Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017

11-12 December 2017, Rome

Editors:

Roberto Basili
Malvina Nissim
Giorgio Satta

aAccademia
university
press

© 2017 by AILC - Associazione Italiana di Linguistica Computazionale
sede legale: c/o Bernardo Magnini, Via delle Cave 61, 38122 Trento
codice fiscale 96101430229
email: info@ai-lc.it

Pubblicazione resa disponibile
nei termini della licenza Creative Commons
Attribuzione – Non commerciale – Non opere derivate 4.0

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it

isbn 0978-88-99982-76-8
www.aAccademia.it/CLIC_2017

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

Table of Contents

I Front Matter

II Contributed Papers

Part I

Front Matter

Preface[*]

On behalf of the Programme Committee, a very warm welcome to the Fourth Italian Conference on Computational Linguistics (CLiC-it). This edition of the conference is held in the wonderful city of Rome! The conference is locally organised by the University of Rome "Tor Vergata", and is hosted at the headquarters of the Italian Research Council (CNR). The CLiC-it conference series is an initiative of the Italian Association for Computational Linguistics (AILC) which, after four years of activity, has clearly established itself as the premier national forum for research and development in the fields of Computational Linguistics and Natural Language Processing, where leading researchers and practitioners from academia and industry meet to share their research results, experiences, and challenges.

This year CLiC-it received 72 submissions, against 64 submissions in 2015 and 69 submissions in 2016. The Programme Committee worked very hard to ensure that every paper received at least two careful and fair reviews. This process finally led to the acceptance of 21 papers for oral presentation and 37 papers for poster presentation, with a global acceptance rate of 80% motivated by the inclusive spirit of the conference and which is in line with the previous editions (81% in 2015 and 80% in 2016). The conference is also receiving considerable attention from the international community, with 21 (29%) submissions this year showing at least one author affiliated to a foreign institution. This amounts to a total of 40 authors over 186 (21%) affiliated to 11 foreign countries: Croatia, Czech Republic, France, Germany, Netherlands, Romania, Spain, Sweden, Switzerland, United Kingdom, and United States. Regardless of the format of presentation, all accepted papers are allocated 6 pages in the proceedings, available as open access publication.

In line with previous editions, the conference is organised around 12 thematic areas. This is a slight reduction with respect to the 13 thematic areas in the 2016 edition of CLiC-it: we have merged the area "information retrieval and question answering" and the area "information extraction, entity linking and (linked) open data". This year we have also implemented a considerable reduction on the number of area chairs, moving from 30 area chairs in 2016, with two or three area chairs per area, to 16 area chairs in 2017, with one or two area chairs per area, on the basis of the expected number of submissions. On a retrospect, the upper bound of two area chairs per area proved to be a reasonable one, given that the most populated thematic area received 13 submissions.

In addition to the technical programme, this year we are honoured to have as invited speakers such internationally recognised researchers as Marco Baroni (Facebook Artificial Intelligence Research), Yoav Goldberg (Bar-Ilan University), and Rada Mihalcea (University of Michigan). We are very grateful to Marco, Rada and Yoav for agreeing to share with the Italian community their knowledge and expertise on key topics in computational linguistics.

The programme also includes two panels. One focuses on teaching NLP both at the bachelor and master levels in Italy and also Europe, and involves panelists who are lecturers at Italian Universities and teach students with very different backgrounds. In the context of this panel we will also discuss the results of a survey launched in the months prior to CLiC-it 2017 and aimed at obtaining a panoramic overview of all Computational Linguistics and Natural Language Processing-related courses taught at Italian institutions. The second panel revolves around the work that is being done within the AI task force launched by the Italian Government in order to better understand and explore the opportunities offered by Artificial Intelligence towards public services. We would like to thank warmly all the panelists who accepted to be involved in the two events.

[*] The CLiC-it 2017 proceedings were originally published online by CEUR (CEUR-WS.org, ISSN 1613-0073).

Traditionally, around one half of the participants at CLiC-it are young postdocs, PhD students, and even undergraduate students. This year we want to pay some special attention to these people by featuring three novel, outreach activities that we would like to highlight here. The first activity is intended to expose to excellent research our young students who might not be able to travel every year to top-notch conferences. We have thus started a special track called "Research Communications", encouraging authors of articles published in 2017 at outstanding international conferences in our field to submit short abstracts of their work. Research communications are not published in the proceedings, but are orally presented within a dedicated session at the conference, in order to enforce dissemination of excellence in research. Out of 7 submissions, we selected 5 excellent works that will be presented at the conference. As a second activity, intended to recognise excellence in student research, this year we are introducing a prize for the best Master Thesis (Laurea Magistrale) in Computational Linguistics, submitted at an Italian University. This special prize is also endorsed by AILC. We have received 10 candidate theses, which have been evaluated by a special jury. The prize will be awarded at the conference, by a member of the jury, accompanied by an oral presentation of the thesis by the student. The third initiative is the introduction of two tutorials, one at the beginning and one at the end of the conference. They are complementary inasmuch one ("Stretching the Meaning of Words: Inputs for Lexical Resources and Lexical Semantic Models", by Elisabetta Ježek) is targeted to those researchers in NLP who might be less accustomed to lexical theories, while the other one ("Implementing dynamic neural networks for language with DyNet", by Yoav Goldberg) is targeted to those who want to catch up with state-of-the-art neural approaches, with an applied flavour. We are extremely grateful to Elisabetta and Yoav who have agreed to teach these tutorials in the context of CLiC-it 2017. And to particularly highlight the importance that such opportunities have for young researchers, we are proud of having made the tutorials' attendance free for all registered students.

Even if CLiC-it is a medium size conference, pulling together this meeting requires major effort on the part of many people. This conference would not have been possible without the dedication, devotion and hard work of the members of the Local Organising Committee, who volunteered their time and energies to contribute to the success of the event. We are also extremely grateful to our Programme Committee members for producing 207 detailed and insightful reviews, as well as to the Area Chairs who assisted the Programme Chairs in their duties. All these people are named in the following pages. We also want to acknowledge the support from endorsing organisations and institutions and from all of our sponsors, who generously provided funds and services that are crucial for the realisation of this event. Special thanks are also due to the University of Rome "Tor Vergata" and to the Italian Research Council for their support in the organisation of the event and for hosting the conference. Finally, we want to acknowledge the EasyChair infrastructure for the management of the review process and the support in the collection of the camera-ready papers for the composition of the conference proceedings.

Please join us at CLiC-it 2017 to interact with experts from academia and industry on topics related to Computational Linguistics and Natural Language Processing and to experience and share new research findings, best practices, state-of-the-art systems and applications. We hope that this year's conference is intellectually stimulating and that you take home many new ideas and techniques that will help extend your own research.

Roberto Basili, Malvina Nissim, Giorgio Satta
CLiC-it 2017 General Chairs

Organising Committee

Conference and Programme Chairs

- Roberto Basili, University of Rome “Tor Vergata”
- Malvina Nissim, University of Groningen
- Giorgio Satta, University of Padua

Area Chairs

- Vito Pirrelli, Istituto di Linguistica Computazionale “A. Zampolli”, CNR, Pisa (Cognitive Modelling of Language Processing and Psycholinguistics)
- Marco De Gemmis, University of Bari, and Fabrizio Sebastiani, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, CNR, Pisa (Information Extraction, Information Retrieval and Question Answering)
- Claudia Soria, Istituto di Linguistica Computazionale “A. Zampolli”, CNR, Pisa, and Tommaso Caselli, University of Groningen (Language Resources)
- Elisabetta Ježek, University of Pavia (Linguistic Issues in Computational Linguistics and in Natural Language Processing)
- Alessandro Moschitti, University of Trento (Machine Learning and Language Processing)
- Marcello Federico, Fondazione Bruno Kessler, Trento (Machine Translation and Multilinguality)
- Felice Dell’Orletta, Istituto di Linguistica Computazionale “A. Zampolli”, CNR, Pisa (Morphology and Syntax Processing)
- Rachele Sprugnoli, Fondazione Bruno Kessler, Trento (Natural Language Processing for Digital Humanities)
- Valerio Basile, University of Rome “La Sapienza”, and Nicole Novielli, University of Bari (Natural Language Processing for Web and Social Media)
- Viviana Patti, University of Turin (Pragmatics and Creativity)
- Giovanni Semeraro, University of Bari, and Alessandro Lenci, University of Pisa (Semantics and Knowledge Acquisition)
- Fabio Tamburini, University of Bologna (Spoken Language Processing and Automatic Speech Understanding)

Local Organisers

- Danilo Croce, University of Rome “Tor Vergata”
- Giuseppe Castellucci, Almawave
- Andrea Vanzo, University of Rome “La Sapienza”
- Simone Filice, University of Rome “Tor Vergata”
- Paola Cibello, University of Rome “Tor Vergata”

Reviewers

Simona Amenta, Fahad Anas Khan, Giuseppe Attardi, Alexandra Balahur, Valentina Bambini, Francesco Barbieri, Gianni Barlacchi, Valerio Basile, Pierpaolo Basile, Nuria Bel, Andrea Bellandi, Luisa Bentivogli, Raffaella Bernardi, Laurent Besacier, Andrea Bolioli, Daniele Bonadiman, Anna Borghi, Federico Boschetti, Cristina Bosco, Dominique Brunato, Cristina Burani, Davide Buscaldi, Elena Cabrio, Cristina Cacciari, Basilio Calderone, Nicoletta Calzolari, Angelo Cangelosi, Annalina Caputo, Tommaso Caselli, Giuseppe Castellucci, Diego Ceccarelli, Emmanuele Chersoni, Cristiano Chesi, Isabella Chiari, Francesca Chiusaroli, Andrea Cimino, Chloé Clavel, Giovanni Colavizza, Bonaventura Coppola, Elisa Corino, Piero Cosi, Davide Crepaldi, Stefano Cresci, Danilo Croce, Francesco Cutugno, Giovanni Da San Martino, Rossana Damiano, Morena Danieli, Marco De Gemmis, Thierry Declerck, Felice Dell'Orletta, Giorgio Maria Di Nunzio, Maud Ehrmann, Andrea Esuli, Benamara Farah, Stefano Faralli, Marcello Federico, Marcello Ferro, Nicola Ferro, Simone Filice, Antske Fokkens, Mikel Forcada, Diego Frassinelli, Francesca Frontini, Aldo Gangemi, Andrea Gesmundo, Emiliano Giovannetti, Aurelie Herbelot, Elisabetta Ježek, Gabriella Lapesa, Alberto Lavelli, Gianluca Lebani, Alessandro Lenci, Eleonora Litta, Vincenzo Lombardo, Pasquale Lops, Simone Magnolini, Alessandro Maisto, Diego Marcheggiani, Marco Marelli, Claudia Marzi, Francesca Masini, Alessandro Mazzei, Massimo Melucci, Stefano Menini, Anne-Lyse Minard, Monica Monachini, Alejandro Moreo Fernández, Véronique Moriceau, Alessandro Moschitti, Claudio Mulatti, Cataldo Musto, Federico Nanni, Fedelucio Narducci, Massimo Nicosia, Nicole Novielli, Debora Nozza, Antonio Origlia, Rainer Osswald, Giulio Paci, Denis Paperno, Lucia Passaro, Marco Passarotti, Viviana Patti, Serena Pelosi, Giovanni Pezzulo, Paola Pietrandrea, Daniele Pighin, Michael Piotrowski, Arianna Pipitone, Vito Pirrelli, Roberto Pirrone, Simone Paolo Ponzetto, Valeria Quochi, Diego Reforgiato, Giuseppe Rizzo, Geoffrey Rockwell, Matteo Romanello, Salvatore Romeo, Francesco Ronzano, Paolo Rosso, Irene Russo, Manuela Sanguinetti, Diana Santos, Enrico Santus, Fabrizio Sebastiani, Giovanni Semeraro, Marco S. G. Senaldi, Aliaksei Severyn, Maria Simi, Claudia Soria, Caroline Sporleder, Rachele Sprugnoli, Carlo Strapparava, Simone Sulpizio, Fabio Tamburini, Francesco Tarasconi, Sara Tonelli, Antonio Toral, Antonio Uva, Andrea Vanzo, Giulia Venturi, Francesco Vespignani, Renata Vieira, Laure Vieu, Fabio Zanzotto.

CLiC-it 2017 is endorsed by

CLiC-it 2017 is sponsored by

Gold Sponsor

Silver Sponsors

Bronze Sponsors

Part II

Contributed Papers

I Verbi Neologici nell'Italiano del Web: Comportamento Sintattico e Selezione dell'Ausiliare

Matteo Amore
Università degli studi di Pavia
matteo.amore@gmail.com

Abstract

Italiano. Si è analizzata la selezione dell'ausiliare da parte dei verbi intransitivi, all'interno di un gruppo di neologismi italiani. I verbi studiati sono stati tratti da elenchi di neologismi disponibili online. Si è quindi verificata la loro presenza e il loro comportamento sul corpus itTenTen. Tali verbi hanno dimostrato una evidente preferenza per l'ausiliare *avere*.

English. *We analyzed the auxiliary verb selection made by Italian intransitive verbs, purposely by neological verbs. The verbs were chosen from online lists of neologisms. We checked the presence and the behaviour of these verbs on the itTenTen corpus. We found that almost all verbs choose the* avere *'have' auxiliary.*

1 Introduzione

L'obiettivo di questo lavoro consiste nell'individuare l'Ausiliare (Aux) selezionato dai verbi neologici intransitivi dell'italiano. Prima di discutere approfonditamente l'ipotesi di ricerca, però, è necessario presentare il fenomeno dell'intransitività scissa, in quanto esso costituisce lo sfondo teorico in cui questo studio si vuole inserire.

1.1 Intrasitività scissa in italiano

In italiano i tempi verbali composti sono formati da una forma flessa dell'Aux *essere* o *avere* e da un participio passato. I Verbi (V) transitivi attivi (p.e. *mangiare*) selezionano *avere*,[1] i riflessivi (p.e. *specchiarsi*) selezionano *essere*, mentre i Verbi Intransitivi (V Intr) mostrano un comportamento più variegato: alcuni selezionano essere (p.e. *andare*),[2] altri *avere* (p.e. *camminare*), altri possono comparire con entrambi (p.e. *correre*). Per dare conto della diversificazione interna al gruppo dei V Intr, Perlmutter (1978) postula l'esistenza di due sub-classi di verbi: i V inaccusativi (con Aux *essere*) e i V inergativi (con Aux *avere*). La differenza tra queste due categorie corrisponderebbe ad una diversa struttura sintattica profonda: nei V inaccusativi il soggetto (Subj) superficiale corrisponderebbe ad un oggetto diretto nella struttura profonda; nei V inergativi il Subj superficiale corrisponderebbe ad un Subj anche nella struttura profonda. All'interno di questa teoria, quindi, la spiegazione del fenomeno risiede nel livello sintattico. Sono in accordo con questa impostazione, tra gli altri, anche gli studi di Burzio (1986). Altri studiosi hanno proposto che la distinzione tra inaccusativi e inergativi sia basata unicamente su criteri semantici (Van Valin, 1990; Dowty, 1991; Bentley e Eythórsson, 2003). A partire dallo studio di Levin e Rappaport Hovav (1995), si è però affermata una tendenza che punta ad integrare i due piani di analisi; un approccio di questo tipo è utilizzato, nell'ambito della Role and Reference Grammar, da Centineo (1996), in cui particolare importanza è data al ruolo del Subj rispetto al verbo.

Seguendo un'altra ipotesi, Sorace (tra gli altri, Sorace, 2000; Bard *et al.*, 2010) identifica con l'aspetto verbale il fattore semantico determinante per l'intransitività e, parallelamente, individua 4 sottoclassi gerarchicamente ordinate all'interno dei V intransitivi. Tale gerarchia è chiamata ASH (Auxiliary Selection Hierarchy). A un polo della gerarchia si trovano i V che selezionano esclusivamente l'Aux *essere* (inaccusativi, massimamente telici), all'altro polo si trovano i V che selezionano esclusivamente l'Aux *avere* (inergativi, non telici). I V appartenenti alle due sottoclassi intermedie mostrano maggiore flessibilità di interpretazione. All'interno degli inaccusativi la distinzione è dovuta alla staticità del predicato, mentre all'interno degli inergativi il parametro distintivo è costituito dall'agentività. Tale sistema gerarchico, oltre ad essere stato dimostrato per più lingue, è stato anche testato attraverso una

[1] Nelle forme passive e in quelle impersonali l'Aux è *essere* (nel passivo si può trovare anche *venire* invece di *essere*).

[2] Si noti che, tra le lingue romanze, l'italiano è la lingua con il più alto numero di V Intr con Aux *essere* (Bentley & Eythórsson, 2003).

serie di esperimenti psicolinguistici (Bard *et al.*, 2010).

Il lavoro qui presentato costituisce la fase iniziale di una ricerca più ampia che conterrà un'analisi più approfondita dei Verbi Neologici (VNeo). Pertanto in questo contributo vengono presentati dati di tipo quantitativo, riservando alle fasi di ricerca successive le ipotesi sulle cause dell'inergatività.

1.2 Obiettivo di ricerca

L'obiettivo di questo contributo consiste in un'analisi del comportamento sintattico dei VNeo dell'italiano. Particolare attenzione viene riservata ai V Intr e alla selezione dell'Aux di questi V nei tempi composti, in quanto l'Aux corrisponde a una delle discriminanti fondamentali per determinare l'inaccusatività o inergatività di un V Intr in italiano (cfr. §1.1). Fine ultimo di questa ricerca è stato, perciò, fornire un'analisi quantitativa in grado di mostrare quanti VNeo Intr selezionano *avere* come Aux. In questo modo si è cercato di individuare ed evidenziare eventuali linee di tendenza nel comportamento sintattico dei VNeo.

1.3 Neologismi

Nell'ambito degli studi sull'intransitività, i VNeo costituiscono un'area ancora non indagata. Un neologismo è una parola o espressione nuova, formata attraverso le regole di formazione proprie del sistema lessicale di una lingua e non ancora registrata nei vocabolari (Adamo e Della Valle, 2017:8). Tuttavia riconoscere quali parole siano effettivamente neologismi non è un compito facile, in quanto la percezione di una parola come nuova può dipendere in larga parte dalle competenze e dai criteri soggettivi propri di ogni utente della lingua (Quemada, 2006:9;11).

Il motivo per cui si è scelto di indagare i neologismi consiste nel loro essere elementi il cui uso non si è ancora stabilizzato e che pretanto possono presentare oscillazioni e/o suscitare incertezza in chi li utilizza. Si pensi, p.e., alla traduzione italiana del V inglese *to scan* nel significato di 'riprodurre digitalmente un'immagine attraverso uno scanner': in questo caso l'uso oscilla tra due V identici nel significato ma differenti nella forma, ossia *scansionare* e *scannerizzare*. Tali oscillazioni appaiono naturali e possono essere viste come sintomi dell'instabilità concettuale, morfologica e pragmatica delle nuove parole (Quemada, 2006:9; Adamo e Della Valle, 2017:23-24).

Nonostante la presenza o l'assenza di una voce nei dizionari sia un fattore indiscutibilmente importante per l'identificazione di un neologismo, in questo lavoro, si è dato maggior risalto agli aspetti di incertezza e instabilità tipici delle neoformazioni, poiché si è voluto verfificare se tale incertezza possa realizzarsi anche nella scelta dell'Aux nei tempi composti. Proprio per dare più spazio a tali possibili oscillazioni si è scelto di utilizzare anche una risorsa online basata sulle segnalazioni degli utenti del sito dell'Accademia della Crusca (cfr. §2.2).

In conclusione, proprio perché lo statuto neologico di una espressione può assumere a volte contorni sfumati (cfr. *supra*), la nozione di neologismo utilizzata in questo lavoro è piuttosto ampia ed è volta ad includere, oltre ai neologismi *stricto sensu*, non solo termini che stanno entrando (o potrebbero entrare) nel lessico comune a partire da linguaggi settoriali o varietà non standard, ma anche parole recentemente registrate in opere lessicografiche. Da questi presupposti derivano pertanto le scelte metodologiche operate (cfr.§2).

2 Metodologia

2.1 Scelta del corpus

La scelta di una ricerca corpus-based è stata effettuata per rispondere a più esigenze contemporaneamente. Dato che il contesto sintattico è considerato decisivo nella scelta di un Aux, l'utilizzo di un corpus è sembrata una scelta valida. Inoltre il corpus itTenTen (Jakubíček *et al.*, 2013) è il più grande corpus disponibile per l'italiano (4,9 miliardi di parole), ed è anche un corpus web-based e ciò ha sicuramente favorito la presenza, nelle attestazioni, di molti verbi usati in ambito informatico, settore particolarmente esposto all'influenza dell'inglese e continuo portatore di nuovi referenti.

2.2 Creazione della lista

Per ottenere una lista di neologismi quanto più aggiornata possibile, si è scelto di basare l'indagine su due elenchi disponibili online. Il primo è costituito dalla pagina riservata ai neologismi su Treccani.it (http://www.treccani.it/lingua_italiana/neologismi/searchNeologismi.jsp).[3] Tale lista è stata scelta per la sua notevole ampiezza (più di 12000 voci) e perché, nonostante sia un lavoro in continua fase di sviluppo, ogni voce presente è corredata da un contesto d'uso reale tratto da un quotidiano o rivista a diffusione nazionale. Ciò dimostra come dietro a questo elenco ci sia un processo di revisione e controllo delle voci.

Il secondo riferimento corrisponde, invece, all'elenco dei termini nuovi che sono stati segnalati più frequentemente dagli utenti del sito dell'Accademia della Crusca (http://www.accademiadellacrusca.it/it/lingua-

[3] Ultima consultazione 15/01/2017.

italiana/parole-nuove/parole-piu-segnalate).[4] È necessario sottolineare che su tale elenco non viene esercitato nessun controllo editoriale,[5] né vi è un confronto con risorse lessicografiche esistenti, perciò la lista può contenere voci già incluse nei dizionari oppure elementi legati a varietà regionali. Nonostante questi limiti, si è scelto di utilizzare la risorsa considerandola come un riflesso della percezione dei parlanti riguardo alle parole avvertite come neoformazioni. Un ulteriore e più importante vantaggio offerto da questa lista è la presenza di V meno diffusi e afferenti ad ambiti più eterogenei[6] rispetto a quelli presenti nell'elenco di Treccani.it.[7]

Un aspetto postivo di entrambe le risorse è costituito dal loro costante aggiornamento che, offrendo una rappresentazione dinamica del lessico, ha permesso di analizzare termini la cui diffusione fosse la più recente possibile.[8] Quindi, sebbene le risorse presentino dei limiti, si è scelto di utilizzare entrambi gli elenchi senza apportare modifiche o filtri.

Non esistendo una risorsa che ne permettesse la consultazione offline, l'elenco completo dei neologismi presenti su Treccani.it è stato estratto automaticamente.[9] La lista di tutte le forme provenienti dalle due fonti comprende circa 12500 parole ed espressioni complesse.[10] All'interno della lista, sono state esaminate tutte le parole terminanti in *-are*, *-ere* o *-ire*, con lo scopo di selezionare esclusivamente i VNeo. Lo spoglio manuale dell'elenco ha condotto a un totale di 368 lemmi.

2.3 Ricerca su itTenTen

Partendo da tale lista, si è condotta una ricerca sul corpus ItTenTen, attraverso l'interfaccia di Sketch Engine (Kilgarriff *et al.*, 2014; http://www.sketchengine.co.uk), la quale permette, oltre ad altre tipologie, sia una ricerca per lemma, sia una ricerca per forme singole. Per ogni verbo si sono cercate sia tutte le occorrenze del relativo lemma (per ottenere tutte le forme flesse), sia le occorrenze del solo participio passato. La ricerca per participio passato è stata necessaria per vari motivi: a) per trovare anche occorrenze che avessero presentato una lemmatizzazione errata (e, trattandosi di neologismi, si è ipotizzato che non tutti fossero lemmatizzati correttamente); b) per ottenere in maniera più immediata risultati contenenti forme verbali composte da Aux e participio; c) per includere anche i casi in cui il participio passato fosse stato etichettato come aggettivo (per moltissimi verbi, infatti, parte delle occorrenze è costituita da participi passati in funzione aggettivale). Considerata la quantità degli elementi da ricercare per la ricerca sul corpus si è utilizzato il comando in linguaggio Python webbrowser.open( ) che ha permesso l'apertura di più pagine web (e quindi più ricerche) contemporaneamente.[11]

2.4 Analisi morfosintattica

Per ogni elemento della lista sono stati individuati gli eventuali suffissi derivativi (p.e. *-izzare*, *-eggiare*) e, quando possibile, la base lessicale da cui il neologismo è derivato (p.e. *slalom* è base di *slalomeggiare*). Per quanto riguarda la base lessicale, si è scelto di registrare anche le informazioni relative alla lingua di origine (italiano o inglese). La lingua d'origine, però, non è stata annotata nelle seguenti situazioni: casi in cui la base fosse un nome proprio di persona o di luogo (p.e. *Berlusconi* in *berlusconeggiare*, *Lisbona* in *lisbonizzare*), casi in cui la base fosse il nome di un marchio (p.e. *Facebook* in *facebookare)* e casi in cui la base corrispondesse ad una sigla (p.e. LOL in *lollare*). In quest'ultima situazione si è scelto di non segnalare la lingua d'origine in quanto si è ritenuto che l'uso (e la conoscenza del significato) di alcune sigle o acronimi può essere indipendente dalla conoscenza delle singole parole che hanno dato vita alla sigla stessa (Adamo e Della Valle, 2017:103).

L'analisi più pertinente all'obiettivo di questo lavoro è quella relativa al comportamento sintattico dei V. Nei casi in cui nel corpus fosse presente almeno una occorrenza per lemma, le informazioni sintattiche sono state dedotte dal corpus; per i V non rappresentati nel corpus (esclusi dalle analisi successive) l'attribuzione del tipo sintattico è stata basata sul giudizio di chi scrive. Le categorie utilizzate per descrivere il tipo sintattico del verbo sono le seguenti: Transitivo (Tr); Intransitivo (Intr); Alternante Transitivo/Intransitivo (Tr\Intr);[12] inoltre sono

[4] Ultima consultazione 15/01/2017.

[5] L'unica revisione che viene effettuata è volta ad eliminare dall'elenco volgarità, bestemmie *et simlia*.

[6] Si danno qui alcuni esempi: *screenshottare* 'salvare come immagine ciò che viene riprodotto su uno schermo, p.e. di uno smartphone'; *camperare* 'pratica attuabile in alcuni videogiochi che consiste nel rimanere a lungo nascosti per evitare di essere colpiti'. Entrambi questi V sono stati riscontrati nel corpus itTenTen.

[7] L'elenco di neologismi costruito dall' ONLI (Osservatorio Neologico della Lingua Italiana) non è stato incluso nella ricerca in quanto quest'ultimo presentava un gruppo più ristretto di V, parzialmente rappresentato anche nell'elenco di Treccani.it.

[8] Si noti che i più recenti repertori di neologismi italiani stampati sono stati pubblicati nel 2008 (Adamo e Della Valle, 2008; Sanguineti, 2008).

[9] Tale operazione è stata effettuata attraverso la funzione "carica dati esterni da web" di Microsoft Excel 2010®.

[10] Oltre a parole semplici (p.e. *admin*), figurano anche composti (p.e. *aereo bomba*, *baby-hacker*); sintagmi di vario genere (p.e. *adozione mite*, *a colpi di maggioranza*); sigle (p.e. *ADSL*).

[11] Per ulteriori informazioni sul comando si rimanda a https://docs.python.org/2/library/webbrowser.html .

[12] Oltre ai V esclusivamente trasitivi e V esclusivamente intransitivi, in italiano esiste anche una classe piuttosto numerosa di

state riconosciute altre categorie presenti in misura minore rispetto alle precedenti.[13] Infine, per i V Intr (o alternanti Tr\Intr) è stato individuato, quando possibile, l'Aux selezionato dal V (distinguendo dagli altri i V che presentavano doppio Aux). Al fine di ottenere dati più solidi da un punto di vista empirico, si è scelto di prendere in considerazione solo i verbi rappresentati nel corpus. Trattandosi, in alcuni casi, di lemmi molto rari e poco utilizzati si è scelto di esaminare solo i VNeo che fossero rappresentati da più di un'occorrenza nel corpus.[14] I VNeo presenti nel corpus sono stati separati dagli altri e costituiscono una lista di 206 lemmi.

Si noti, infine, che qui viene proposta un'analisi di tipo quantitativo e che tale indagine si configura come un primo stadio di uno studio più ampio che comprenderà anche analisi di natura qualitativa.

3 Risultati

3.1 Comportamento sintattico

All'interno del gruppo di VNeo riscontrati nel corpus, la maggioranza dei V risultano essere transitivi (1), mentre i V Intr costituiscono un insieme molto più piccolo (2). In misura ancora inferiore vi sono i V che mostrano un'alternanza del tipo Transitivo(3)a)\Intransitivo(3)b). La (Figura 1) riassume questi dati.

(1) Gli investigatori *hanno attenzionato* entrambe le abitazioni

(2) Anni fa girava a *comiziare* su una camionetta

(3)

a. Molta gente che *ha opinionato* questo film

b. Il tuttologo può *opinionare* su tutto e tutti

Per quanto riguarda la relazione tra comportamento sintattico e lingua di origine della base del neologismo, nei dati analizzati è stata riscontrata un'asimmetria: nei neologismi derivati da termini italiani i V con alternanza Tr\Intr sono un gruppo più ristretto rispetto a quelli non alternanti, mentre per i neologismi derivati da termini inglesi vi è una sostanziale parità tra V Intr e V Tr\Intr (cfr. Figura 2).

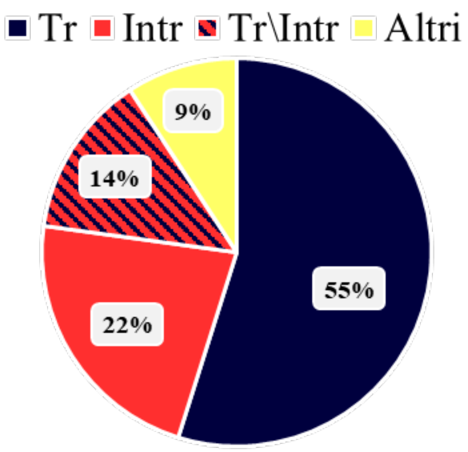


Figura 1. Comportamento sintattico dei neologismi esaminati

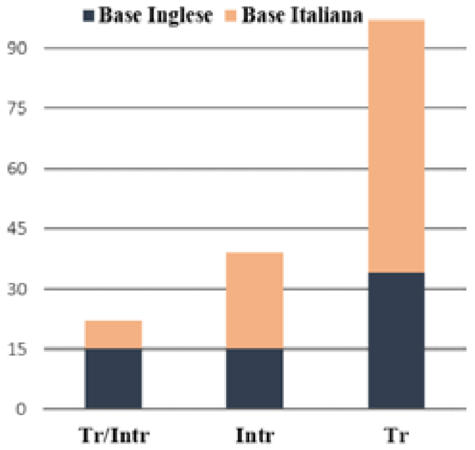


Figura 2. Comportamento sintattico e lingua di origine della base del neologismo

3.2 Selezione dell'ausiliare

Per quanto riguarda la selezione dell'Aux da parte dei V Intr, i dati raccolti mostrano una tendenza netta. Se si considerano solo i V Intr o i V che presentano alternanza con il tipo Intr (per un totale di 80 V), solamente in 3 casi l'Aux è *essere* (ma cfr. *infra*); in altri 3 casi l'Aux può alternare tra *essere* e *avere*; in 24 casi non è stato rintracciato nessun Aux. In tutti gli altri 50 casi l'Aux selezionato è *avere* e, se si escludono i casi per cui l'Aux non è stato rinvenuto, tale cifra corrisponde all'89% del totale.

I verbi con Aux *essere* sono: *pacsare* (usato anche nella forma *pacsarsi con qcn.*) 'unirsi in un contratto coniugale denominato PACS' (4); *imbufalire/imbufalirsi* 'arrabbiarsi'(5); *loggare/loggarsi* ma solo nel significato 'effettuare l'accesso ad un sistema protetto tramite delle credenziali' (6).[15] Si noti

verbi che possono essere, a seconda dei casi, sia transitivi sia intransitivi. Es.: *suonare* è Tr in *Giulia sta suonando una nuova canzone*, mentre è Intr in *Giulia sta suonando* (Ježek, 2003:94).

[13] Tali categorie sono: Riflessivo/pronominale; Passivo (usato nel caso in cui di un V siano state trovate solo occorenze in forma passiva, p.e. *alluminizzare*); Alternante Intransitivo/riflessivo; Alternante Transitivo/intransitivo/riflessivo; Alternante Transitivo/passivo; Alternante Intransitivo/passivo; Non identificabile su base intuitiva.

[14] Si noti che itTenTen viene sottoposto a revisione periodica, per cui la presenza di alcuni dati potrebbe variare nel tempo.

[15] Questo verbo presenta anche un altro significato di tipo transitivo che corrisponde a 'registrare le operazioni effettuate'.

che *pacsare/-rsi* può essere considerato come un troponimo di *sposare/-rsi*, mentre *loggare/-rsi* è un sinonimo (tecnico) di 'entrare'. Entrambi gli equivalenti non neologici presentano l'Aux *essere*. Si noti che in alcune occorrenze di questi V, il participio passato sembra svolgere una funzione aggettivale.

(4) Il mio compagno, con cui *sono pacsato* da più di 6 anni

(5) Davide *era* letteralmente *imbufalito* contro la situazione dei parcheggi

(6) Se *sei loggato*, verrai identificato con il tuo nome utente

Per quanto riguarda il V *imbufalire/imbufalirsi*, la forma *imbufalire* in casi come (5) presenta l'Aux *essere* in modo coerente con altri verbi parasintetici dell'italiano (p.e. *ingiallire, sbiancare, arrossire*) il cui significato equivale a 'diventare X' (e in cui il soggetto non ha controllo sull'azione). Anche nella forma riflessiva, ovviamente, il V mostra l'Aux *essere*.

I tre verbi che presentano l'alternanza *avere/essere* sono invece: *crashare* 'smettere di funzionare per motivi legati ad un software'(7); *sifonare* 'rubare; fare sesso' (8); *colazionare* 'fare colazione' (9).

(7)

a. Il programma *ha crashato* varie volte

b. Infatti *è crashato* solo una volta

(8)

a. Ormai si *è sifonato* Gabriela

b. I due traditori *avrebbero* sicuramente *sifonato* selvaggiamente

(9)

a. Stamane *avevo* già *colazionato* con caffèlatte

b. Prometto *sarò* già *colazionato*

Questi ultimi due V però non andrebbero considerati come realmente alternanti: infatti *sifonare* presenta l'Aux *essere* solo nella forma pronominale *sifonarsi qcn.* (modellato sui vari verbi che indicano l'attività sessuale come *scopare/-rsi*, etc.), mentre quando il participio passato *colazionato* appare con il V *essere*, esso si comporta come un aggettivo indicante uno stato (interpretabile come 'sazio a causa della colazione').

La classe dei V Intr con Aux *avere* è piuttosto eterogenea e per completezza se ne riportano solo alcuni esempi: *outperformare* 'produrre prestazioni superiori alla media'(10); *saltapicchiare* 'saltellare, passare da un posto ad un altro'(11).

(10) Le banche oggi *hanno outperformato* l'indice generale

(11) *Ho saltapicchiato* qua e là.

4 Conclusioni

La netta maggioranza dei neologismi esaminati seleziona *avere* come Aux nei tempi composti, mentre in presenza dell'Aux *essere* il participio tende a indicare uno stato e sembra assumere un valore aggettivale.

Tale comportamento potrebbe essere indice di una propensione, in italiano contemporaneo, per una separazione netta tra le funzioni svolte dai due Aux. In alternativa si può ipotizare una preferenza per la creazione di V inergativi, rispetto agli inaccusativi.

Come accennato in precedenza (§1.1 e §2.4), tale analisi ha voluto indagare aspetti principalmente quantitativi, perciò sarà necessariamente ampliata per stabilire quali sono, se esistono, le motivazioni per i dati riscontrati. In particolare, si indagheranno ipotesi riguardanti l'aspetto semantico dei VNeo. Un'analisi incentrata sulla semantica dei neologismi studiati potrebbe essere utile, infatti, per mostrare eventuali tendenze nella creazione di nuovi verbi (in termini di preferenze semantiche e/o aspettuali).

Un ulteriore campo di indagine meritevole di approfondimento potrebbe essere quello relativo ai valori e alle funzioni svolte dal participio passato in cooccorrenza con l'Aux *essere*.

Nonostante i limiti costituiti da una nozione di neologismo ampia ma tenue e dall'utilizzo di un corpus vasto ma dalla rappresentatività relativa, il lavoro offre dei dati quantitativamente chiari e che confermano l'importanza di risorse come i corpora per l'avanzamento degli studi lessicografici (e non solo).

Bibliografia

Adamo, Giovanni e Valeria Della Valle. 2017. *Che cos'è un neologismo.* Carocci Editore, Roma.

Adamo, Giovanni e Valeria Della Valle (edd.). 2008. *Il Vocabolario Treccani. Neologismi. Parole nuove dai giornali.* Istituto della Enciclopedia Italiana, Roma.

Bard, Ellen Gurman, Cheryl Frenck-Mestre e Antonella Sorace. 2010. Processing auxiliary selection with Italian intransitive verbs. *Linguistics* 48(2): 325-361.

Bentley, Delia e Thórhallur Eythórsson. 2003. Auxiliary selection and the semantics of unaccusativity. *Lingua* 114(4): 447–471.

Burzio, Luigi. 1986. *Italian Syntax: A Government-Binding Approach*. Springer Science and Business Media, Dordrecht.

Centineo, Giulia. 1996. A lexical theory of auxiliary selection in Italian. *Probus* 8(3): 223-272.

Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. *Language* 67(3): 547–619.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ e Vít Suchomel. 2013. The tenten corpus family. *7th International Corpus Linguistics Conference CL*: 125–127.

Ježek Elisabetta. 2003. *Classi di verbi tra semantica e sintassi*. Edizioni ETS, Pisa.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ e Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1): 7–36.

Levin, Beth e Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-lexical Semantics Interface*. MIT Press, Cambridge, MA.

Perlmutter, David M. 1978. Impersonal Passives and the Unaccusative Hypothesis. *Annual Meeting of the Berkeley Linguistics Society* 38: 157–190.

Quemada, Bernard. 2006. Problématiques de la néologie. In Giovanni Adamo e Valeria Della Valle (edd.) *Che fine fanno i neologismi?* (Lessico Intellettuale Europeo): 1–21. Olschki Editore, Firenze.

Sanguineti, Edoardo (ed.). 2008. *Grande Dizionario della lingua italiana: Supplemento 2009*. UTET, Torino.

Sorace, Antonella. 2000. Gradients in Auxiliary Selection with Intransitive Verbs. *Language* 76(4): 859-890.

Valin, Robert D. van. 1990. Semantic Parameters of Split Intransitivity. *Language* 66(2): 221–260.

Sitografia

http://www.accademiadellacrusca.it/it/lingua-italiana/parole-nuove/parole-piu-segnalate
(Ultima consultazione 15/01/2017)

http://www.sketchengine.co.uk

http://www.treccani.it/lingua_italiana/neologismi/searchNeologismi.jsp
(Ultima consultazione 15/01/2017)

Predicting Land Use of Italian Cities using Structural Semantic Models

Gianni Barlacchi[1,2], **Bruno Lepri**[3], **Alessandro Moschitti**[1,4]
[1]Department of Information Engineering and Computer Science, University of Trento
[2] TIM Semantics and Knowledge Innovation Lab, Trento
[3] Fondazione Bruno Kessler, Trento
[4]Qatar Computing Research Institute, HBKU
{gianni.barlacchi,amoschitti}@gmail.com
lepri@fbk.eu

Abstract

English. We propose a hierarchical semantic representation of urban areas extracted from a social network to classify the most predominant land use, which is a very common task in urban computing. We encode geo-social data from Location-Based Social Networks with standard feature vectors and a conceptual tree structure that we call Geo-Tree. We use the latter in kernel machines, which can thus perform accurate classification, exploiting hierarchical substructure of concepts as features. Our comparative study on three datasets extracted from Milan, Rome and Naples shows that Tree Kernels applied to Geo-Trees are very effective improving the state of the art.

Italiano. *In questo lavoro, proponiamo un nuovo modello semantico per la rappresentazione di aree urbane utilizzando dati da social media. In particolare, modelliamo tale informazione con una struttura ad albero che abbiamo chiamato Geo-Tree. Questa viene utilizzata, in combinazione con un vettore di feature classico, nelle kernel machine per fare classificazione della destinazione di uso delle aree urbane. Abbiamo valutato il nostro approccio su tre grandi metropoli italiane quali Milano, Roma e Napoli. I risultati mostrano come i Geo-Tree, applicati ai Tree Kernel, riescono a raggiungere risultati di molto superiori ad altri modelli attualmente stato dell'arte.*

1 Introduction

The growing availability of data from cities (Barlacchi et al., 2015a) (e.g., traffic flow, human mobility and geographical data) opens new opportunities for predicting and thus optimizing human activities. For example, the automatic analysis of land use enables the possibility of better administrating a city in terms of resources and provided services. However, such analysis requires specific information, which is often not available for privacy concerns. In this paper we follow the approach proposed in (Barlacchi et al., 2017) and we use public textual descriptions of urban areas to design a novel machine learning representation. We represent urban areas as: (i) a bag-of-concepts (BOC), e.g., the terms *Arts and Entertainment*, *College and University*, *Event*, *Food* extracted from the Foursquare description of the area; and (ii) the same concepts above organized in a tree, which reflects the hierarchical organization of Foursquare activities. We combine BOC vectors with Tree Kernels (TKs) (Collins and Duffy, 2002; Moschitti, 2006) applied to concept trees (Geo-Tree) and use them in Support Vector Machines (SVMs). The Geo-Tree allows the model to learn complex structural and semantic patterns from the hierarchical conceptualization of an area. We show that TKs not only can capture semantic information from natural language text, e.g., as shown for semantic role labeling (Moschitti et al., 2008) and question answering (Severyn and Moschitti, 2013; Barlacchi et al., 2015b), but they can also learn from the hierarchy above to perform semantic inference, such as deciding which is the major activity of a land.

We carried out a study on land use prediction of three Italian cities: Milan, Rome and Naples as follows: (i) we divided each city in squares of 200x200 meters; (ii) then, we classify the most predominant land use class (e.g., *High Density Urban Fabric* or *Open Space and Outdoor*), assigned by the city administration. The results show that GeoTKs achieve an impressive improvement over state-of-the-art classification approaches based on BOC., i.e., 21.2%, 13.6% and 54.3% of relative improvement in Macro-F1 over Milan, Rome and

Naples datasets, respectively.

2 Related Work

Previous work has modeled land use classification by means of different sources of information. For example, Yuan et al. (2012) built a framework that, using human mobility patterns derived from taxicab trajectories and Point Of Interests (POIs), classifies the functionality of an area for the city of Beijing. Assem et al. (2016) proposed a spatio-temporal approach based on three different clustering algorithms to model the change of functionality of a city's region over time. They extracted features from Foursquare's POIs and check-in activities of Manhattan. Yao et al. (2017) built sequences of POI concepts reflecting their spatial distance. Then, they applied Word2Vec (Mikolov et al., 2013) to these sequences to derive vectors representing each area, which was used to train a land use classifier. In general, most previous work applies extensive feature engineering, which is typically costly as it requires to fully understand the target domain. Our approach alleviates this problem with automatic feature engineering applied to an abstract land representation.

3 Land Description Data

Geospatial city areas are described with the popular shape file format, where each shape is a collection of points geo-localized using their coordinates. The latter are provided with the well-known Coordinate Reference System (CRS) WGS84, adopted for the common latitude/longitude geolocation. We use (i) shape files provided by Urban Atlas[1], a website providing data for large urban areas (more than $100,000$ inhabitants) and (ii) POIs from Foursquare[2].

3.1 Land Use

Cities are divided in small areas associated with a main land use. In total, there are 17 different land use classes defined from the open dataset Urban Atlas [3]. We focused on those related to city centers, discarding those less interesting from a social viewpoint, i.e., associated with rural areas such as forests, agricultural, semi-natural and wetland areas and mineral extraction and dump sites. Thus, we selected the following categories: (i) *High Density Urban Fabric*, (ii) *Medium Density Urban Fabric*, (iii) *Low Density Urban Fabric*, (iv) *Industrial, commercial, public, military and private units*, (v) *Open Space & Recreation*, (vi) *Transportation*. We collapsed *Medium* and *Low Density Urban Fabric* into one single category, *ML-Density Urban Fabric* as they only have few samples. Land use distribution is very fine-grained, making its classification based on POI information very difficult. A trade-off between classification accuracy and the desired area granularity consists in segmenting the regions in squared cells. As each cell can contain more than one land use label, we consider the predominant label as its primary use.

3.2 Point-Of-Interest

A POI is usually characterized by a location (i.e., latitude and longitude), textual information (e.g., a description of the activity in that place) and a hierarchical categorization that provides different levels of detail about the activity of the place (e.g., *Food*, *Asian Restaurant*, *Chinese Restaurant*). We used POIs extracted from Foursquare, a geolocation-based social network supported with web search facilities for places and a recommendation system. In particular, we extracted 46,731, 43,389 and 7,219 POIs from Milan, Rome and Naples[4], respectively. We focused on the ten macro-categories of such POIs[5], each one specialized in maximum four levels of detail.

4 Structural Models

In most machine learning algorithms data examples are transformed in feature vectors, which in turn are used in dot products to carry out both learning and classification. Kernel Machines (KMs) allow for replacing the dot product with kernel functions, which directly compute it on the examples, i.e., they avoid the transformation of examples in vectors. The main advantage of KMs is a much lower computational complexity as it does not directly depend on the feature space size.

4.1 Point-of-interests Features

The most straightforward way to represent an area by means of Foursquare data is the use its POIs. Every venue is hierarchically categorized (e.g., *Professional and Other Places* → *Medical Center* → *Doctor's office*) and the categories are used to produce an aggregated representation of the area.

[1] https://www.eea.europa.eu/data-and-maps/data/urban-atlas

[2] https://foursquare.com/

[3] https://www.eea.europa.eu/data-and-maps/data/urban-atlas#tab-additional-information

[4] For some reasons Foursquare is less popular in Naples

[5] https://developer.foursquare.com/categorytree

We define a feature vector for a grid cell by counting the macro-level category (e.g., *Food*) in all the POIs that we found in that cell.

4.2 Geographical Tree Kernel

Foursquare has its own hierarchy of categories, which is used to characterize each location and activity (e.g., restaurants or shops) in the database. Thus, each Foursquare POI is associated with a hierarchical path, which semantically describes the type of location/activity (e.g., for *Chinese Restaurant*, we have the path *Food* → *Asian Restaurant* → *Chinese Restaurant*). The path is much more informative than just the target POI name, as it provides feature combinations following the structure and the node proximity information, e.g., *Food* & *Asian Restaurant* or *Asian Restaurant* & *Chinese Restaurant* are valid features whereas *Food* & *Chinese Restaurant* is not.

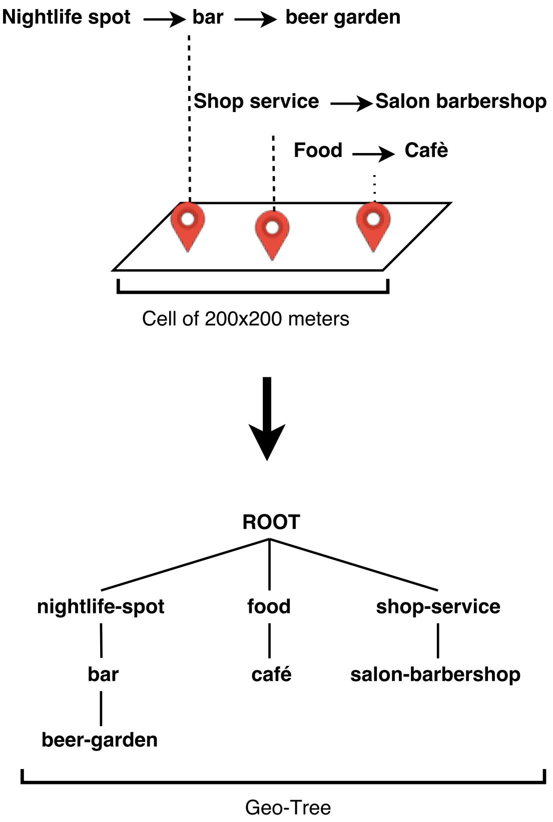


Figure 1: Example of Geo-Tree built from a collection POIs in a cell.

Geo-Tree: we propose a new tree structure, i.e., Geo-Tree, whose nodes and edges among them are subsets of the Foursquare hierarchy (FH). A Geo-Tree of a grid cell is constituted by a new root node connecting the subtrees of FH rooted in concepts present in the cell. In other words, we connect all the paths of FH starting from grid concepts. Figure 1 shows an example of the FH paths of a cell and the resulting Geo-Tree.

This way, the nodes of the first level, i.e., the root children, correspond to the most general FH categories, e.g., *Arts & Entertainment*, *Event*, *Food*, etc., the second level of our tree corresponds to the second level of the hierarchical tree of Foursquare, and so on. The terminal nodes are the finest-grained descriptions in terms of category about the area, e.g., *College Baseball Diamond* or *Southwestern French Restaurant*. For example, Fig. 2 illustrates the semantic structure of a grid cell obtained by combining all the categories' chains of each venue.

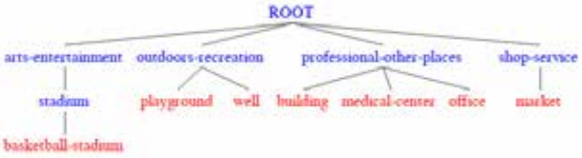


Figure 2: Example of Geo-Tree in Milan for an area labeled as Open Space & Recreation.

GeoTK: given a Geo-Tree, we can encode all its substructures in kernel machines using TKs. In particular, we used the Syntactic Tree Kernels (STK$_b$) with Bag-Of-Words and the Partial Tree Kernel (PTK) (Moschitti, 2006). Our TKs by construction do not consider the frequency[6] of the POIs present in a given grid cell.

BOC kernel: to complement GeoTK, we represent a cell also creating a BOC representation, namely we count the macro-level category (e.g., *Food*) in all the POIs that we found in any cell grid. This way, we generate feature vectors by counting the number of each activity under each macro-category. In order to take into consideration the popularity of the area, we included (i) the total sum of unique users that did at least one check-in in the cell, and (ii) the total sum of check-in done in the cell. Note that, given an area, the number of unique users provides an idea on how many people visited it, while the number of check-in can be used to represent its popularity.

Kernel combination: finally, given two geographical areas, x^a and x^b, we define a kernel combining Geo-Tree and BOC as: $K(x^a, x^b) = TK(\mathbf{t}^a, \mathbf{t}^b) + KV(\mathbf{v}^a, \mathbf{v}^b)$, where TK is any

[6] It is possible to add the frequency in the kernel computation but for our study we preferred to have a completely different representation from previous typical frequency-based approaches.

structural kernel function applied to tree representations, $\mathbf{t}^a$ and $\mathbf{t}^b$ of the geographical areas and KV is a kernel applied to the feature vectors, $\mathbf{v}^a$ and $\mathbf{v}^b$, extracted from x^a and x^b using any data source available (e.g., text, social media, mobile phone and census data).

5 Experiments and Results

We performed our experiments on the data from Milan, Rome and Naples. We used a grid of 200x200meters as it is indicated as the best size from other similar previous work on land use classification (Toole et al., 2012; Zhan et al., 2014; Barlacchi et al., 2017). We applied a pre-processing step in order to filter out cells for which land use classification cannot be performed. In particular, for Milan and Rome, we selected the central point of the shape and we included those cells that have their centroid in the radius of 15 and 8 kilometers, respectively. For Naples, we kept all the cells due to the smaller size of the city. Then, for all the three cities, we removed the cells that (i) cover areas without a specified land use (e.g., the cells in the sea) and (ii) do not have POIs (e.g., the countryside cells). After this step, we obtained a grid with 2,581, 5,657 and 1,314 cells for Milan, Rome and Naples, respectively. We created, separately for each city, the training and test set randomly sampling 80% vs. 20% of the cells. We labelled the dataset following the same category aggregation strategy proposed by Zhan et al. (2014), who assigned the predominant land use class to each grid cell.

To train our models, we applied SVM-Light-TK[7], which enables the use of structural kernels (Moschitti, 2006) in SVM-Light[8]. In particular, due to the nature of the task, we used the Python wrapper around SVM-Light-TK to perform multiclass classification[9]. We experimented with linear, polynomial and radial basis function kernels applied to standard feature vectors. We measured the performance of our classifier by averaging Precision, Recall and F1 over all land use categories.

5.1 Results for Land Use Classification

We trained multi-class classifiers using common learning algorithm such XGboost (Chen and Guestrin, 2016), and SVM using linear, polynomial and radial basis function kernels, named SVM-{Lin, Poly, Rbf}, respectively, and our structural semantic models, indicated with STK$_b$ and PTK. We also combined kernels with a simple summation, e.g., PTK+Lin indicates an SVM using such kernel combination.

City	Model	Prec.	Rec.	F1
Milan	baseline	0.200	0.119	0.149
	XGBoost	0.294	0.317	0.297
	STK_b+Rbf	0.368	**0.364**	**0.360**
	PTK+Rbf	0.430	0.350	0.345
	STK_b	**0.448**	0.307	0.320
	PTK	0.364	0.302	0.309
Rome	baseline	0.200	0.089	0.124
	XGBoost	0.291	0.306	0.279
	STK_b+Lin	**0.359**	**0.314**	**0.317**
	STK	0.338	0.300	0.302
	PTK	0.340	0.300	0.299
	PTK+Lin	0.359	0.297	0.291
Naples	baseline	0.200	0.100	0.133
	XGBoost	0.236	0.272	0.219
	STK_b+Rbf	**0.361**	**0.331**	**0.338**
	STK_b+Lin	0.338	0.302	0.300
	STK_b	0.409	0.290	0.299
	PTK	0.318	0.298	0.297

Table 1: Classification results on Rome, Milan and Naples. Prec., Rec. and F1 are averaged over all categories.

Table 1 shows the average of F1, Precision and Recall over the different categories. The model *baseline* is obtained by always classifying an example with the label *High Density Urban Fabric*, which is the most frequent. Due to space constraint, we only reported six models, namely: the baseline, XGBoost and the top four kernel models.

We note that: (i) GeoTK always outperforms XGBoost and the baseline, demonstrating the superiority of our novel approach. This is an interesting finding as XGboost is the current state of the art for land use classification. (ii) STK$_b$ combined with feature vector always produces the best results, improving the F1-score over XGBoost up to 6.3, 3.8 and 11.9 absolute points for Milan, Rome and Naples, respectively. (iii) Kernel combinations always provide the best results.

6 Conclusions

In this paper, we have introduced Geo-Trees, a novel semantic representation based on a hierarchical classification of POIs, to better exploit geo-social data to the classification of the primary land use of an urban area. This is an important task as it gives the urban planners and policy makers the possibility to better administrate and renew a city in terms of infrastructures, resources and services. More in detail, we have built our classi-

[7] http://disi.unitn.it/moschitti/Tree-Kernel.htm

[8] http://svmlight.joachims.org/

[9] https://github.com/aseveryn/SVMTK-Multiclass-Classifier

fiers with combinations of a kernel over BOC and TKs applied to Geo-Trees, thus exploiting hierarchical substructure of concepts as features. Our comparative study on three large Italian cities, Milan, Rome and Naples shows that our models can relatively improve the state of the art up to 11.9 absolute points in F1-score.

Acknowledgments

This work has been partially supported by the EC project CogNet, 671625 (H2020-ICT-2014-2, Research and Innovation action).

References

Haytham Assem, Lei Xu, Teodora Sandra Buda, and Declan O'Sullivan. 2016. Spatio-temporal clustering approach for detecting functional regions in cities. In *ICTAI*, pages 370–377. IEEE.

Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015a. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2:150055.

Gianni Barlacchi, Massimo Nicosia, and Alessandro Moschitti. 2015b. Sacry: Syntax-based automatic crossword puzzle resolution system. *ACL-IJCNLP 2015*, page 79.

G Barlacchi, A Rossi, B Lepri, and A Moschitti. 2017. Structural semantic models for automatic analysis of land use.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794, New York, NY, USA. ACM.

Michael Collins and Nigel Duffy. 2002. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In *ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Alessandro Moschitti, Daniele Pighin, and Roberto Basili. 2008. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *ECML*, pages 318–329. Springer.

Aliaksei Severyn and Alessandro Moschitti. 2013. Automatic feature engineering for answer selection and extraction. In *EMNLP*, volume 13, pages 458–467.

Jameson L Toole, Michael Ulm, Marta C González, and Dietmar Bauer. 2012. Inferring land use from mobile phone activity. In *SIGKDD International Workshop on Urban Computing*, pages 1–8. ACM.

Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang, and Ke Mai. 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. *International Journal of Geographical Information Science*, 31(4):825–848.

Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and pois. In *KDD*, pages 186–194. ACM.

Xianyuan Zhan, Satish V Ukkusuri, and Feng Zhu. 2014. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics*, 14(3-4):647–667.

Predicting Controversial News Using Facebook Reactions

Angelo Basile
Faculty of ICT, Univ. of Malta
CLCG, Univ. of Groningen
Groningen, NL
a.basile@student.rug.nl

Tommaso Caselli
CLTL, VU Amsterdam
CLCG, Univ. of Groningen
Amsterdam/Groningen, NL
t.caselli@gmail.com

Malvina Nissim
CLCG, Univ. of Groningen
Groningen, NL
m.nissim@rug.nl

Abstract

English. Different events and their reception in different reader communities may give rise to controversy. We propose a distant supervised entropy-based model that uses Facebook reactions as proxies for predicting news controversy. We prove the validity of this approach by running within- and across-source experiments, where different news sources are conceived to approximately correspond to different reader communities. Contextually, we also present and share an automatically generated corpus for controversy prediction in Italian.

Italiano. *Diversi tipi di eventi e la loro percezione in diverse comunità di utenti/lettori possono dare vita a controversie. In questo lavoro proponiamo un modello basato su entropia e sviluppato secondo il paradigma della "distant supervision" per predire controversie sulle notizie usando le reazioni di Facebook come "proxy". La validità dell'approccio è dimostrata attraverso una serie di esperimenti usando dati provenienti dalla stessa fonte o da fonti diverse. Contestualmente, presentiamo anche un corpus generato automaticamente per la previsione delle controversie in italiano.*

1 Introduction and Background

The explosion of social media (e.g. Facebook, Twitter, Disqus, Reddit, Wikipedia, among others) and the increased interactions with readers-users that traditional newspapers embraced, have transformed the Web in a huge *agora*, where news are shared, opinions are exchanged, and debates arise. On many topics, such as climate change, abortion, vaccination, among others, people strongly disagree. Following the work by Timmermans et al. (2017), we call *controversies* situations where, even after lengthy interactions, opinions of the involved participants tend to remain unchanged and become more and more polarized towards extreme values.

Modeling and understanding controversies may be useful in many situations. Journalists and news agencies may pay additional attention in the framing of a certain news, government officials and policy makers may be more aware of the issues involved in specific laws, social media managers might be more careful, i.e. monitor controversial content, in order to avoid the spreading of hate speech, and the general public may benefit as well thanks to a reduction of the "filter bubble" effect (Pariser, 2011).

Recently, computational approaches on controversy detection have been developed with varying degrees of success (Awadallah et al., 2012; Borra et al., 2015; Dori-Hacohen and Allan, 2015; Lourentzou et al., 2015). Works in the areas of Sentiment Analysis (Zhou et al., 2013; Deng and Wiebe, 2015; Deng et al., 2013; Chambers et al., 2015; Russo et al., 2015), Emotion Detection (Strapparava and Mihalcea, 2007; Strapparava and Mihalcea, 2008; Russo et al., 2011; Pool and Nissim, 2016), and Stance Detection (Mohammad et al., 2016) are, on the other hand, only partially related, as they focus on predicting/classifying the content of a message with respect to specific categories, such as "positive", "negative", "neutral", or "joy", "sadness" (among others), or as "being in favour" or "being against". They may be seen as necessary but not sufficient tools for detecting/predicting controversy (Timmermans et al., 2017).

The main contribution of this work is two-fold: i.) we propose a distant supervised entropy-based

Table 1: Sample rows from the dataset showing how entropy varies in relation to the reactions.

ID	TEXT	LIKE	LOVE	ANGRY	HAHA	WOW	SAD	entropy
1.)	In volo sul Piemonte con biplano anni '30	32	0	0	0	0	0	0.0
2.)	Medico anti vaccini radiato	5700	216	220	36	42	22	0.5
3.)	Piacenza, abbattuto il cinghiale Agostino	125	7	34	33	5	78	1.9

model to predict controversial news; and ii.) we present and share an automatically created corpus to train and test models for controversy detection. At this stage of development, we focused only on Italian, although the methods are completely language independent and can be reproduced for any language for which news are available on Facebook. The remainder of the paper is structured as follows: Section 2 illustrates the methods used to collect the data and develop the entropy-based model. Section 3 reports on the experiments and results both in a within- and across-source setting. Finally, Section 4 draws conclusions and outlines future research. Data and code are made available at `https://anbasile.github.io/predictingcontroversy/`.

2 Data and Methodology

We used the Facebook Graph API[1] to download news headlines (including the `description` and `body` fields) from four major Italian newspapers. Of these, two are slightly politically biased (*Corriere della Sera* and *La Repubblica*, both centre/centre-left), two openly biased ones (*Il Manifesto*, left-wing, and *il Giornale*, right-wing), and one news agency (*ANSA*).

Together with each news, we also downloaded all users' reactions.[2] Facebook reactions can be used as a proxy for annotations (Pool and Nissim, 2016), allowing to train a model for predicting the degree of controversy associated to news. On the basis of the definition of controversy previously introduced, our working hypothesis is that if users' reactions fall in two or more emotion classes (not necessarily opposed in terms of "polarity") with high frequencies, the controversy of a news item is higher. Building on this, we assume that entropy can be explanatory in modelling news' controversy: the higher the entropy, the more controversial the news. To better clarify this aspect, consider the data in Table 1. Each sample is the text of a Facebook post, for which we report the reaction breakdown (including `LIKE`), and its overall entropy based on reaction counts. Users expressing different reactions suggest that a text is likely to be controversial as it is shown by the high values of the entropy, as illustrated in examples 2.) and 3.) vs. example 1.).

For each source, namely the newspaper pages mentioned at the beginning of this section, we downloaded a collection of posts which appeared between mid-April and early July 2017. Posts with less than 30 reactions in total were discarded. For each post, we collected: i.) the link to the full article on the source's website (a large majority of the posts include this); ii.) an excerpt of the article (the variable `text`); iii.) additional texts commenting the article, when available (the variable `descriptor`); iv.) the full list of users' reactions. Finally, for a portion of the posts (1024 out of 3595, i.e. 28,48%; column "# body" in Table 2) we downloaded the entire text of the article (the variable `body`).[3]. Table 2 provides an overview of the data collected, including, for each source, the number of Facebook posts, the number of tokens, the number of posts for which the full article was retrieved, the token-post ratio, i.e. the number of tokens per post, and, finally, the average entropy.

Table 2: Dataset shape and average entropy score (avg H) per source.

SOURCE	# POSTS	# TOKEN	# BODY	RATIO TOKEN-POST	AVG H
AgenziaANSA	883	18,635	528	21.10	1.0216
corrieredellasera	594	23,811	124	40.08	0.9135
ilgiornale	1,022	8,665	124	8.47	1.1266
ilmanifesto	752	36,479	124	48.5	0.6195
repubblica	344	7,763	124	22.56	0.9078
total	3,595	95,353	1,024	26.52	0.9386

To further verify the soundness of using entropy as an indicator of controversy, we inspected the top-10 and bottom-10 news in the full dataset

[1] `https://developers.facebook.com/docs/graph-api`

[2] Since February 2016, Facebook users can react to a post not only with a like but by choosing from a set of 5 different emotions: `ANGRY, LIKE, HAHA, WOW, SAD, LOVE.`

[3] The full text of the article is not always available or accessible. Furthermore, there is a monthly limit to the data that can be downloaded. We made sure that the final dataset we used contained, for each source, the same number of posts for which the full body could be downloaded. This constraint did not apply to *ANSA*

Table 3: Sample of entropy-ranked top-5 and bottom-5 posts.

	TOPIC	TEXT
TOP	Incident	Fuggono dall'aereo in fiamme ma si fermano per scattare un selfie a pochi metri dall'aereo
	25th April	#25Aprile #Anpi: ""Festa di tutti gli italiani"". Roma divisa, due celebrazioni
	Gender/LGBTQ	"Genere: Sconosciuto". E il Canada gli dà' ragione
	Immigration	Emergenza #migranti, nave Rio Segura arrivata a Salerno. A bordo 11 donne incinte, 256 minori e 13 neonati #FOTO
	Animals	#Piacenza, abbattuto il cinghiale #Agostino. Da giorni nel parco urbano di Galleana, avrebbe caricato il personale
BOTTOM	25th April	#25aprile, ecco i musei statali aperti'
	Movies	"La La Land" meritava la statuetta del miglior film, andata poi a "Moonlight"?'
	Sport	Il Presidente della Sampdoria Massimo Ferrero è raggiante per la vittoria nel derby di Genova'
	Arts	Quando Eugenio Corti morì, il 4 febbraio 2014, Sébastien Lapaque, sul quotidiano parigino Le Figaro, lo definì "uno degli immensi scrittori del nostro tempo"'
	Arts	New York New York ricostruisce i legami artistici dal '28 a metà anni '60"

sorted by entropy (high values on top, high controversy) and manually assigned them to a topic. Table 3 illustrates the results for the top 5 and bottom 5 posts, in terms of entropy score. In addition to identifying a different distribution of topics according to degrees of controversy, we also observed that in some cases, the entities and the specific event mentions interact to generate controversy. For instance, in the case of the "*25th April*" topic[4], the controversial news involves a political actor (i.e. *ANPI*, the National Association of Italian Partisans), and divisions on the celebration of this day, while the non-controversial news reports on museums being open on that day. The entropy score appears to capture this distinction.

3 Experiments

We use the *ANSA* dataset to develop our model. The rationale behind this is that, being *ANSA* a news agency, the texts should be more objective and the controversy should depend on the event itself rather than by its framing in a specific, potentially biased, community. We treat this task as a regression problem, and use mean squared error (MSE) to measure the performance of our system. As baseline, we use a dummy regressor which always predicts the mean entropy of the train dataset: considering that the values range between 0 and 2.9, with a standard deviation of 0.4, a system that always predicts the mean entropy is already performing reasonably well. Furthermore, this is in line with the average entropy values of each dataset, ranging from 0.6195 (Table 2, *Il Manifesto*) up to 1.1266 (Table 2, *Il Giornale*).

Settings We use two main settings. Firstly, the data for training and testing the model originates from the same Facebook page, and we use cross-validation. Secondly, we train and test across pages, so as to investigate the model's portability across potentially different communities. This second setting can shed light on the issue of *perspective bias*, as controversy around a specific topic or entity could exist in one domain (or, in this case, in one community as proxied by Facebook pages) and not in another one. In both settings, we run our best model, developed as described below.

Features For predicting the entropy of the reactions to a given text, we built a system using a sparse feature representation and an SVM regressor, with the *scikit-learn* LinearSVR implementation (Buitinck et al., 2013). We used a tf-idf vectorizer to represent the text as both word and character n-grams.

As sentiment might contribute to controversy prediction (Dori-Hacohen and Allan, 2015), we also extended the features with coarse-grained prior polarity information derived from Sentix (Basile and Nissim, 2013), a resource for Italian automatically mapped from the English SentiWordNet (Esuli and Sebastiani, 2006). We represent each token with the absolute values of its polarity (which in Sentix ranges from -1 to +1). This allows us to ignore the specific positive/negative values, and get a more abstract representation on the subjectivity relevance of a token: high values indicate that the text is rich of subjectivity relevant tokens; 0 means that the text is merely objective. For each post, we then compute the average polarity and encoded it into a separate vector. Missing words in the lexicon are simply skipped.

[4] April 25th is a national holiday in Italy to celebrate the end of World War II.

Model development For development, as mentioned, we only used *ANSA*. We experimented with different features and different sizes of texts. In particular, we ran experiments using: i.) only the `text` variable; ii.) a combination of the `text` and the `descriptor` variables; and iii.) a combination of the `text`, the `descriptor`, and the `body` variables. Furthermore, these three basic settings have been extended with the polarity values from Sentix. To fine tune the parameters, a grid-search of the model using a 10-fold cross-validation was conducted. Table 4 reports the results of the different models as well as of the baselines.

Table 4: Results for the cross-validated *ANSA* dataset.

DATA	BASELINE	MODEL	+ SENTIX
text	0.24	0.154	0.155
text+descriptor	0.24	0.146	0.148
text+descriptor+body	0.24	0.146	0.148

The best model shows an improvement of 0.094 MSE with respect to the baseline when extending the variable `text` with `descriptor` and `body`. The use of the variable `text` alone still beats the baseline, but obtains a lower score than the models which include both the `descriptor` and the `body` variables. The extensions with the polarity scores from Sentix decrease the model performances (though still outperforming the baselines). We believe that this behaviour is mainly due to noise in the resource itself and calls for better and more context-oriented sentiment lexicons in Italian. Table 5 summarises the features of the best model, which is based on a combination of the three text variables only: `text`, `descriptor`, and `body` (whenever available), represented as word and character n-grams, ignoring the polarity vectors. This model was used on the reminder of the datasets.

Results on the test set Table 6 illustrates cross-validated results for the newspaper datasets. For comparison and completeness, we report also the results of the cross-validation on the full test set, with and without the extension of the data with *ANSA*.

With the exception of *Il Giornale*, our model always beats the baseline, confirming the validity of the designed approach. Extending the newspaper dataset with the data from *ANSA*, we can observe a reinforcement of the predicting power of the model, with a range between 0.04 to 0.1 points with respect to the corresponding baselines. The positive effect on *Il Giornale* dataset can be due to an extension of the number of tokens, since *Il Giornale* is the dataset with the lowest token-post ration (8,47 tokens per post), which clearly affects our model.

Table 5: Best model's settings and features.

PARAMETER	VALUE
SVR C	10
character ngrams	(2,3)
character binary features	True
character normalization	l2
character sublinear tf	False
word ngrams	(1,3)
word binary features	False
word normalization	l2
word sublinear tf	True

Table 6: Cross-validated results on all datasets.

	BASELINE	STD	MODEL	STD
ilgiornale	**0.21**	0.03	0.22	0.04
ilgiornale+ansa	0.23	0.04	**0.19**	0.03
ilmanifesto	0.15	0.04	**0.11**	0.04
ilmanifesto+ansa	0.24	0.04	**0.14**	0.03
repubblica	0.22	0.07	**0.18**	0.07
repubblica+ansa	0.24	0.04	**0.15**	0.04
corrieredellasera	0.24	0.06	**0.16**	0.06
corrieredellasera+ansa	0.24	0.03	**0.14**	0.04
full_dataset	0.24	0.02	**0.17**	0.03
full_dataset-ansa	0.24	0.03	**0.17**	0.04

Cross-source results in Table 7 are less clear-cut. In these experiments, it clearly emerges that our model works in the large majority of cases, although with no big gains over the baselines. All datasets fail to beat the baseline when predicting controversy on *Il Giornale* and, on the contrary, training on *Il Giornale* only fails to beat the baseline when testing on *La Repubblica*. This suggests that either there must be a difference in the wording used by *Il Giornale* with respect to the other datasets, or that the controversy is affected by perspective bias associated to different communities.

On the other hand, slightly politically oriented newspapers (*La Repubblica* and *Il Corriere della Sera*) and the *ANSA* news agency tend to have a homogeneous behavior, being able to correctly predict controversy in highly politically oriented news (see results for *Il Manifesto* in Table 7). As a matter of fact, the more the post/token ratio is similar between different sources, the better the model works in predicting controversy. For instance, *Il Corriere della Sera* and *Il Manifesto* have a very similar post/token ratio (40,08 and 48,5, respectively) and not surprisingly both cross-source experiments beat the baseline.

Table 7: Cross-source results on all datasets.

TRAIN	TEST	BASELINE	MODEL
ilgiornale	ilmanifesto	0.40	**0.36**
ilgiornale	AgenziaANSA	0.25	**0.24**
ilgiornale	repubblica	**0.26**	0.29
ilgiornale	corrieredellasera	0.28	**0.26**
ilmanifesto	ilgiornale	**0.46**	0.46
ilmanifesto	AgenziaANSA	**0.40**	0.40
ilmanifesto	repubblica	0.30	**0.28**
ilmanifesto	corrieredellasera	0.32	**0.29**
AgenziaANSA	ilgiornale	**0.22**	0.23
AgenziaANSA	ilmanifesto	**0.31**	0.38
AgenziaANSA	repubblica	0.23	**0.21**
AgenziaANSA	corrieredellasera	0.25	**0.23**
repubblica	ilgiornale	**0.25**	0.28
repubblica	ilmanifesto	0.23	**0.23**
repubblica	AgenziaANSA	0.25	**0.23**
repubblica	corrieredellasera	0.23	**0.20**
corrieredellasera	ilgiornale	**0.25**	0.25
corrieredellasera	ilmanifesto	0.23	**0.20**
corrieredellasera	AgenziaANSA	0.25	**0.21**
corrieredellasera	repubblica	0.21	**0.18**

4 Conclusions and Future Work

This paper presents a simple regression model to predict the entropy of a post's reactions based on the Facebook reaction feature. We take this measure as a proxy to predict the *controversy* of news, where the higher the entropy (indicated by highly mixed reactions), the bigger the controversy. We run experiments both within and across communities, exemplified by the Facebook pages of specific newspapers. As a by-product, we have also automatically generated a first reference corpus for controversy prediction in Italian.

The results are promising, given that our model beats the baseline in almost all cases in cross-validation of same source data (see Table 6), and in the large majority of cases when applied cross-sources (see Table 7). At this stage of development, we observed that coarse-grained sentiment values are not useful, although this may depend on the quality of the lexicon employed. Test and training on openly biased datasets (e.g. *Il Giornale*$_{TRAIN}$ - *Il Manifesto*$_{TEST}$, and vice-versa) results in the lowest entropy, suggesting perspective bias in the different community.

The approach we have developed is based on discrete linguistically motivated features. This has an impact in the learned model as it is not able to generalise enough when dealing with low-frequency features and unseen data in the test set. To alleviate this issue, we are planning to model the post representations by using word embeddings.

We are planning to expand the model to account for perspective bias in different communities. News from different sources may be aggregated per event type, for example via the EventRegistry API[5], allowing to explore entropy (and polarisation of reactions) on exactly the same event instance. A first step in this direction would be to detect and match Named Entities to approximately identify similar events. At the reaction-level, the obvious next step is to explore and experiment with *clusters* of reactions (for instance, positive (`LIKE`, `LOVE`, `AHAH`), negative (`ANGRY`, `SAD`), or ambiguous (`WOW`)), instead of treating them all as single and distinct indicators.

Another follow-up is to extend this work to other social media data, such as Twitter. Twitter does not allow for nuances in reactions in the same way that Facebook does, as only one kind of "like" is provided. However, the substantial use of hashtags and emojis might offer alternative proxies to capture a variety of reactions. There is plenty of work on the usefulness of leveraging hashtags as reaction proxies both at a coarse and finer level (Mohammad and Kiritchenko, 2015), but this information, to the best of our knowledge, has not been used to predict likelihood of controversy.

Acknowledgments

One of the authors wants to thank the Spinoza-NWO Project "Understanding Language by Machines" subtrack 3 for making this work possible.

References

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. Opinions network for politically controversial topics. In *Proceedings of the first edition workshop on Politics, elections and data*, pages 15–22. ACM.

[5] `http://eventregistry.org`

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *WASSA@ NAACL-HLT*, pages 100–107.

Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. Societal controversies in wikipedia articles. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 193–196. ACM.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. 2015. Identifying political sentiment between nation states with social media. In *EMNLP*, pages 65–75.

Lingjia Deng and Janyce Wiebe. 2015. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *EMNLP*, pages 179–189.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125, Sofia, Bulgaria, August. Association for Computational Linguistics.

Shiri Dori-Hacohen and James Allan. 2015. Automated controversy detection on the web. In *European Conference on Information Retrieval*, pages 423–434. Springer.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.

Hyunseo Hwang, Youngju Kim, and Catherine U Huh. 2014. Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media*, 58(4):621–633.

Ismini Lourentzou, Graham Dyer, Abhishek Sharma, and ChengXiang Zhai. 2015. Hotspots of news articles: Joint mining of news text & social media to discover controversial points in news. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2948–2950. IEEE.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June. Association for Computational Linguistics.

Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. Emocause: an easy-adaptable approach to emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 153–160. Association for Computational Linguistics.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *SemEval@ NAACL-HLT*, pages 443–450.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM.

Benjamin Timmermans, Lora Aroyo, Evangelos Kanoulas Tobias Kuhn, Kaspar Beelen, and Gerben van Eerten Bob van de Velde. 2017. Controcurator: Understanding controversy using collective intelligence. In *Collective Intelligence 2017*.

Xujuan Zhou, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. In *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*, pages 557–562. IEEE.

Bi-directional LSTM-CNNs-CRF for Italian Sequence Labeling

Pierpaolo Basile, Giovanni Semeraro, Pierluigi Cassotti
Department of Computer Science, University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{firstname.surname}@uniba.it, pierluigicassotti@gmail.com

Abstract

English. In this paper, we propose a Deep Learning architecture for sequence labeling based on a state of the art model that exploits both word- and character-level representations through the combination of bidirectional LSTM, CNN and CRF. We evaluate the proposed method on three Natural Language Processing tasks for Italian: PoS-tagging of tweets, Named Entity Recognition and Super-Sense Tagging. Results show that the system is able to achieve state of the art performance in all the tasks and in some cases overcomes the best systems previously developed for the Italian.

Italiano. *In questo lavoro viene descritta un'architettura di Deep Learning per l'etichettatura di sequenze basata su un modello allo stato dell'arte che utilizza rappresentazioni sia a livello di carattere che di parola attraverso la combinazione di LSTM, CNN e CRF. Il metodo è stato valutato in tre task di elaborazione del linguaggio naturale per la lingua italiana: il PoS-tagging di tweet, il riconoscimento di entità e il Super-Sense Tagging. I risultati ottenuti dimostrano che il sistema è in grado di raggiungere prestazioni allo stato dell'arte in tutti i task e in alcuni casi riesce a superare i sistemi precedentemente sviluppati per la lingua italiana.*

1 Background and Motivation

Deep Learning (DL) gained a lot of attention in last years for its capacity to generalize models without the need of feature engineering and its ability to provide good performance. On the other hand good performance can be achieved by accurately designing the architecture used to perform the learning task. In Natural Language Processing (NLP) several DL architectures have been proposed to solve many tasks, ranging from speech recognition to parsing. Some typical NLP tasks can be solved as sequence labeling problem, such as part-of-speech (PoS) tagging and Named Entity Recognition (NER). Traditional high performance NLP methods for sequence labeling are linear statistical models, including Conditional Random Fields (CRF) and Hidden Markov Models (HMM) (Ratinov and Roth, 2009; Passos et al., 2014; Luo et al., 2015), which rely on hand-crafted features and task/language specific resources. However, developing such task/language specific resources has a cost, moreover it makes difficult to adapt the model to new tasks, new domains or new languages. In (Ma and Hovy, 2016), the authors propose a state of the art sequence labeling method based on a neural network architecture that benefits from both word- and character-level representations through the combination of bidirectional LSTM, CNN and CRF. The method is able to achieve state of the art performance in sequence labeling tasks for the English without the use of hand-crafted features.

In this paper, we exploit the aforementioned architecture for solving three NLP tasks in Italian: PoS-tagging of tweets, NER and Super Sense Tagging (SST). Our research question is to prove the effectiveness of the DL architecture in a different language, in this case Italian, without using language specific features. The results of the evaluation prove that our approach is able to achieve state of the art performance and in some cases it is able to overcome the best systems developed for the Italian without the usage of specific language resources.

The paper is structured as follows: Section 2 provides details about our methodology and summarizes the DL architecture proposed in (Ma and Hovy, 2016), while Section 3 shows the results of the evaluation. Final remarks are reported in Section 4.

2 Methodology

Our approach relies on the DL architecture proposed in (Ma and Hovy, 2016), where the authors combine two aspects previously exploited separately: 1) the use of a character-level representation (Chiu and Nichols, 2015); 2) the addition of an output layer based on CRF (Huang et al., 2015). The architecture is sketched in Figure 1: the input level of the Convolution Neural Network is represented by the character-level representation. A dropout layer (Srivastava et al., 2014) is applied before feeding the CNN with character embeddings. Then the character embeddings are concatenated with the word embeddings to form the input for the Bi-directional LSTM layer. The dropout layer is also applied to output vectors from the LSTM layer. The output layer is based on Conditional Random Fields and it modifies the output vectors of the LSTM in order to find the best output sequence. The CRF layer is useful for learning correlations between labels in neighborhoods, for example generally a noun follows an article in PoS-tagging, or the I-ORG tag[1] cannot follow the I-PER tag in the NER task.

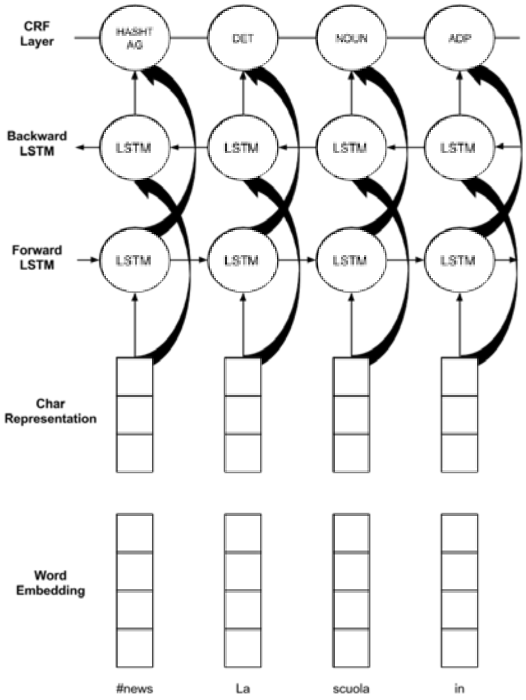


Figure 1: The DL architecture for the sequence labeling.

The aforementioned architecture can be easily adapted to other languages since it does not rely on language dependent features. The only components outside the architecture are the word embeddings that can be built by relying on a corpus of documents of the specific language. In Section 3, we provide details about the setup of the architecture parameters and the building of word embeddings for Italian, in particular we adopt two different word embeddings: ones for PoS-tagging and ones for NER and SST. Moreover, we re-implement[2] the architecture by using the Keras[3] framework and Tensorflow[4] as back-end.

3 Evaluation

We provide an evaluation in the context of three sequence labeling tasks: 1) PoS tagging of Italian tweets; 2) NER of Italian news and 3) Super Sense Tagging. All tasks are performed using Italian datasets, in particular we exploit data coming from the last edition (2016) of EVALITA[5] (Basile et al., 2016) and previous ones (2009 (Magnini and Cappelli, 2009) and 2011[6]). EVALITA[7] is a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language. The usage of a standard benchmark allows to compare our system with the state of the art approaches for the Italian language.

Each task has its specific parameters, but there are some ones that are in common as reported in Table 1. We do not perform any parameters optimization and we use the values proposed in the English evaluation (Ma and Hovy, 2016). We choose this strategy in order to not reduce the training set since validation set is not provided in all the tasks.

3.1 PoS tagging of Tweets

The goal of the task is to perform PoS-tagging of tweets. The task is more challenging with respect to the classical PoS-tagging due to the short and noisy nature of tweets. For the evaluation we adopt the dataset used during the EVALITA 2016 PoSTWITA task (Bosco et al., 2016) in order

[1] Generally, the NER task uses the IOB2 schema for data annotation.

[2] The code is available on line: https://github.com/pippokill/bilstm-cnn-crf-seq-ita

[3] https://keras.io/

[4] https://www.tensorflow.org/

[5] https://github.com/evalita2016/data

[6] http://www.evalita.it/2011/working_notes

[7] http://www.evalita.it/

Parameter	**Value**
Framework	Keras 2.0.1
Back-end	Tensorflow 1.1.0
Char embed. dimension	30
Word embed. dimension	300
Window size	3
LTSM dimension	200 (bi-LTSM 400)
Optimization	Adadelta
Gradient clipping	5.0
Epochs	100 (PoS), 60 (NER and SST)

Table 1: Parameters' values.

System	Accuracy
UNIBA-twita	**.9334**
UNIBA-itwiki	.9199
UNIBA-random300	.8790
ILC-CNR	.9319
UniDuisburg	.9286
UniBologna UnOFF	.9279

Table 2: Results for the PoSTWITA task.

to compare our system with the other EVALITA participants. The dataset contains 6,438 tweets (114,967 tokens) for training and 300 tweets (4,759 tokens) for test. The metric used for the evaluation is the classical tagging accuracy: it is defined as the number of correct PoS tag assignment divided by the total number of tokens in the test set. Participants can predict only one tag for each token.

All the top-performing PoSTWITA systems are based on Deep Neural Networks and, in particular, on LSTM, moreover most systems use word or character embeddings as inputs for their systems. This makes other systems more similar to the one proposed in this paper.

Results of the evaluation are reported in Table 2, our best approach (*UNIBA-twita*) is able to overcome the first three PoSTWITA participants. (*UNIBA-twita*) exploits a corpus of 70M tweets randomly extracted from Twita, a collection of about 800M tweets, for building the word embeddings. It is important to underline that the best system (*ILC-CNR*) (Cimino and Dell'orletta, 2016) in PoSTWITA uses a biLSTM and a RNN by exploiting both word and char embeddings, moreover it use further features based on morpho-syntactic category and spell checker. The good performance of our system probably depends by the CRF layer and the corpus used for building the word embeddings. This hypothesis is supported by the fact that the configuration (*UNIBA-itwiki*) based on word embeddings extracted from Wikipedia obtains the worst result. The configuration *UNIBA-random300* adopts random embeddings, we report this result in order to underline the importance of pre-trained word embeddings. Moreover, the second best system (*UniDuisburg*) (Horsmann and Zesch, 2016) in PoSTWITA exploits a CRF classifier using several features without a DL architecture, while the system *UniBologna UnOFF* (Tamburini, 2016) uses a BiLSTM with a CRF layer by exploiting word embeddings and additional morphological features.

3.2 NER Task

Three tasks about named entities have been organized during the EVALITA evaluation campaigns, respectively in 2007 (Speranza, 2007), 2009 (Speranza, 2009), and 2011 (Lenzi et al., 2013). In this paper we take into account the 2009 edition since the I-CAB dataset [8]used in the evaluation is the same adopted in 2009. In 2007 a different version of I-CAB was used, while in 2011 the task was focused on data transcribed by an ASR system. The I-CAB dataset consists of a set of news manually annotated with four kinds of entities: GPE (geo-political), LOC (location), ORG (organization) and PER (person). The dataset contains 525 news for training and 180 for testing for a total number of 11,410 annotated entities for training and 4,966 ones for testing. The dataset is provided in the IOB2 format.

We build word embeddings by exploiting the Italian version of Wikipedia. Word2vec is used for creating embeddings with a dimension of 300, we remove all words that have less than 40 occurrences in Wikipedia, for the other parameters we adopt the standard values provided by word2vec.

Results of the evaluation are reported in Table 3 and Table 4. Table 3 reports precision (P), recall (R) and F1-measure (F1) for different configurations of the system. In particular: *no-case-sensitive* does not perform lowercase of words for both word embeddings and the lookup table, while *case-sensitive* does it. The *random* configuration randomly initializes embeddings without using pre-trained embeddings, while *no char* does not adopt char embeddings. The results show that

[8] http://ontotext.fbk.eu/icab.html

Configuration	ALL			GPE	LOC	ORG	PER
	P	R	F1	F1	F1	F1	F1
no-case-sensitive	.8286	.8182	**.8234**	.8561	.6220	.6587	.9239
case-sensitive	.8220	.8084	.8151	.8444	.6305	.6421	.9178
random	.7153	.6885	.7017	.7564	.4809	.5209	.8037
no char	.8305	.7426	.7841	.8492	.6200	.5945	.8714

Table 3: Results for the Italian NER task using different configurations.

System	ALL			GPE	LOC	ORG	PER
	P	R	F1	F1	F1	F1	F1
UNIBA	.8286	.8182	**.8234**	.8561	.6220	.6587	.9239
FBK_ZanoliPianta	.8407	.8002	.8200	.8513	.5124	.7056	.8831
UniGen_Gesmundo_r2	.8606	.7733	.8146	.8336	.5081	.7108	.8741
UniTN-FBK-RGB_r2	.8320	.7908	.8109	.8525	.5224	.6961	.8689

Table 4: Results for the Italian NER task compared with other EVALITA 2009 participants.

the best performance is obtained by applying lowercase, moreover the contribution of char embeddings is significant.

Table 4 reports the result of our best configuration (*no-case-sensitive*) with respect to the other EVALITA 2009 participants. The system is able to outperform the first three EVALITA participants thanks to the best performance in recall. All the first three participants adopt classical classification methods: the first system (Zanoli et al., 2009) combines two classifiers (HMM and CRF), the second participant (Gesmundo, 2009) uses a Perceptron algorithm, while the third participant (Mehdad et al., 2009) adopts Support Vector Machine and feature selection. We can conclude that the DL architecture is more effective in the model generalization and in tackling the data sparsity problem. This behavior is supported by the good performance in recognizing LOC entities, in fact the LOC class represents about the 3% of annotated entities in both training and test. Other two systems (Nguyen and Moschitti, 2012; Bonadiman et al., 2015) able to overcome the EVALITA 2009 participants have been proposed in the literature. The former (Nguyen and Moschitti, 2012) achieves the 84.33% of F1 by using re-ranking techniques and the combination of two state-of-the-art NER learning algorithms: conditional random fields and support vector machines. The latter (Bonadiman et al., 2015) exploits a Deep Neural Network with a log-likelihood cost function and a recurrent feedback mechanism to ensure the dependencies between the output tags. This system is able to achieves the 82.81% of F1, a performance comparable with our DL architecture.

3.3 Super Sense Tagging

The Super-Sense Tagging (SST) task (Dei Rossi et al., 2011) consists in annotating each significant entity in a text, like nouns, verbs, adjectives and adverbs, within a general semantic taxonomy defined by the WordNet lexicographer classes (called super-senses, for a total of 45 senses). SST can be considered as a task half-way between NER and Word Sense Disambiguation (WSD): it is an extension of NER, since it uses a larger set of semantic categories, and it is an easier and more practical task with respect to WSD. The dataset has been tagged using the IOB2 format as for the NER task and contains about 276,000 tokens for training and about 50,000 for testing. The metric adopted for the evaluation is the F1, results of the evaluation are reported in Table 5. As word embeddings we use the same ones adopted for the NER task and built upon Wikipedia with lowercase.

System	F1
UNIBA-pos-Adagrad	**.7871**
UNIBA-pos	.7787
UNIBA	.7453
UNIBA-SVMcat	.7866
UNIPI-run3	.7827

Table 5: Results for the Super-Sense Tagging task.

The best performance (*UNIBA-pos-Adagrad*) is obtained using Adagrad instead of Adadelta (*UNIBA-pos*) as optimization method. Moreover, we exploits PoS-tags as additional features, while *UNIBA* uses only tokens and word/char embeddings. The difference in performance between *UNIBA-pos* and *UNIBA* proves the effectiveness of the PoS-tag in this task. The best system in EVALITA 2011 SST task, *UNIBA-SVMcat* (Basile, 2013, 2011), is very close to our best configuration. This system combines lexical and distributional features through an SVM classifier, while the second system (*UNIPI-run3*) (Attardi et al., 2011) exploits lexical features and a Maximum Entropy classifier.

4 Conclusions and Future Work

We propose an evaluation of a state of the art DL architecture for sequence labeling in the context of the Italian language. In particular, we consider three tasks: PoS-tagging of tweets, Named Entity Recognition and Super-Sense tagging. All tasks exploit data coming from EVALITA a standard benchmark for the evaluation of Italian NLP systems. Our system is able to achieve good performance in all the tasks without using hand-crafted features. Analyzing the results, we observe the importance of building word embeddings on appropriate corpora and we note that the system in the SST task is not able to generalize a good model without the pos-tag feature, this underline the importance of this kind of feature in the SST task. As future work, we plan to perform a parameters optimization by reducing the training set and using a portion as validation set. Using less data for training could affect the final performance and it could be interesting to have insights on the trade-off between training on more examples versus the parameters optimization.

Acknowledgments

This work is partially supported by the project "Multilingual Entity Liking" funded by the Apulia Region under the program FutureInResearch.

References

Giuseppe Attardi, Luca Baronti, Stefano Dei Rossi, and Maria Simi. 2011. SuperSense Tagging with a Maximum Entropy Classifier and Dynamic Programming. In *Working Notes of EVALITA 2011*.

P. Basile. 2013. Super-sense tagging using support vector machines and distributional features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7689 LNAI:176–185.

P. Basile, F. Cutugno, M. Nissim, V. Patti, and R. Sprugnoli. 2016. EVALITA 2016: Overview of the 5th evaluation campaign of natural language processing and speech tools for Italian. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR Workshop Proceedings, volume 1749.

Pierpaolo Basile. 2011. UNIBA: Super-sense Tagging at EVALITA 2011. In *Working Notes of EVALITA 2011*.

Daniele Bonadiman, Aliaksei Severyn, and Alessandro Moschitti. 2015. Deep neural networks for named entity recognition in italian. In *CLiC-it 2015 Proceedings of the second Italian Conference on Computational Linguistics*. page 51.

C. Bosco, F. Tamburini, A. Bolioli, and A. Mazzei. 2016. Overview of the EVALITA 2016 Part of speech on twitter for Italian task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR Workshop Proceedings, volume 1749.

Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional LSTM-CNNs. *arXiv preprint arXiv:1511.08308* .

A. Cimino and F. Dell'orletta. 2016. Building the state-of-the-art in POS tagging of Italian Tweets. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR Workshop Proceedings, volume 1749.

Stefano Dei Rossi, Giulia Di Pietro, and Maria Simi. 2011. EVALITA 2011: Description and Results of the SuperSense Tagging Task. In *Working Notes of EVALITA 2011*.

Andrea Gesmundo. 2009. Bidirectional sequence classification for named entities recognition. In *Proceedings of the Workshop Evalita 2009*.

T. Horsmann and T. Zesch. 2016. Building a social media adapted PoS tagger using flexTag - A case study on Italian tweets. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR Workshop Proceedings, volume 1749.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* .

Valentina Bartalesi Lenzi, Manuela Speranza, and Rachele Sprugnoli. 2013. Named entity recognition on transcribed broadcast news at EVALITA 2011. In *Revised Papers from EVALITA11: International Workshop on the Evaluation of Natural Language and Speech Tools for Italian*. Springer, volume 7689, pages 86–97.

G. Luo, X. Huang, C.-Y. Lin, and Z. Nie. 2015. Joint named entity recognition and disambiguation. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. pages 879–888.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* .

Bernardo Magnini and Amedeo Cappelli. 2009. Introduction to Evalita 2009. In *Proceedings of the Workshop Evalita 2009*.

Yashar Mehdad, Vitalie Scurtu, and Evgeny Stepanov. 2009. Italian named entity recognizer participation in NER task @ Evalita 09. In *Proceedings of the Workshop Evalita 2009*.

Truc-Vien T Nguyen and Alessandro Moschitti. 2012. Structural reranking models for named entity recognition. *Intelligenza Artificiale* 6(2):177–190.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv:1404.5367* .

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 147–155.

Manuela Speranza. 2007. Evalita 2007: the named entity recognition task. In *Proceedings of the Workshop Evalita 2007*.

Manuela Speranza. 2009. The named entity recognition task at evalita 2009. In *Proceedings of the Workshop Evalita 2009*.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

F. Tamburini. 2016. A BiLSTM-CRF PoS-tagger for Italian tweets using morphological information. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR Workshop Proceedings, volume 1749.

Roberto Zanoli, Emanuele Pianta, and Claudio Giuliano. 2009. Named entity recognition through redundancy driven classifiers. In *Proceedings of the Workshop Evalita 2009*.

Monitoring Adolescents' Distress using Social Web data as a Source: the InsideOut Project

Basili Roberto[†‡], Bellomaria Valentina[‡], Bugge Niels J.[⋆], Croce Danilo[†‡],
De Michele Francesco[•], Fiori Nastro Federico[•], Fiori Nastro Paolo[•],
Michel Chantal[⋆∗], Schmidt Stefanie J.[⋆], Schultze-Lutter Frauke[⋆◦]

[†] University of Roma, Tor Vergata [‡] Reveal srl [⋆] University of Bern [∗]University of Geneva
[•] Sapienza University of Rome [◦] Heinrich-Heine University, Düsseldorf

{basili|croce}@info.uniroma2.it, bellomaria@revealsrl.it, niels.bugge@gmail.com
{chantal.michel|stefanie.schmidt|frauke.schultze-lutter}@kjp.unibe.ch
paolo.fiorinastro@uniroma1.it, {francescodemichele1981|federico.fiori.nastro }@gmail.com

Abstract

English. The role of Social Media in the psychological and social development of adolescents and young adults is increasingly important as it impacts on the quality of their interpersonal communication dynamics. The InsideOut project explores the possibility to use Social Web mining methodologies and technologies to collect information about adolescents' distress from their micro-blogging activities. The project is promoting a complex language processing workflow to approach the collection, enrichment and summarization of user generated contents over Twitter. This paper presents the general architecture of the InsideOut Web Platform and the resources produced by an integrated effort among computer science and mental health professionals.

Italiano. *Il ruolo dei Social Media nella crescita psicologica e sociale risulta essere sempre più importante poiché influisce sulla qualità e sulle dinamiche di comunicazioni interpersonali, specialmente riguardo le ultime generazioni. Il progetto InsideOut esplora la applicabilità di metodologie e tecnologie che consentono l'individuazione nel Web di evidenze riferibili a sorgenti di stress negli adolescenti. Il progetto propone un workflow di elaborazione linguistica in grado di gestire la raccolta, l'arricchimento e la sintesi dei contenuti generati dagli utenti su Twitter. Nel paper verrà presentata l'architettura generale della piattaforma Web InsideOut e le risorse che derivano dal lavoro congiunto di ricercatori provenienti dall'ambito informatico e medico.*

1 Introduction

Among adolescents, the use of Social Media, such as Twitter, Facebook or Instagram, has grown exponentially in the past years. This makes them a valuable source of information on the well-being of adolescents, but also concerning on their mental health. Mental disorders are the main cause of disability in adolescents and young adults (Gore et al., 2011), affecting an average of 10 to 20% of youth worldwide (Kieling et al., 2011). Thus, for the emerging complex relationship between the use of Social Media, mental health and well-being (Best et al., 2014), Social Media are a valuable source of information on the mental health and well-being of adolescents.

Social Media thus play an increasingly important role in the psychological and social development of adolescents as it impacts on the quality of their social interactions and networks. Any attempt to study and govern mental health in young communities (adolescents, students, interest groups) must take into account an effective and large scale methodology to monitor all the behaviors on the Web that exhibit and impact on mental habits, trends and social practices. The possibility of predicting writers demographics from their writings is an important research topic in the Computational Linguistic Community. In fact, the idea that a writer's style may reveal age, gender or other sociodemographic information has been also targeted in the "Plagiarism analysis, Authorship identification, and Near-duplicate detection" (PAN) (e.g., (Rangel et al., 2014; Rangel et al., 2015; Rangel et al., 2016)) or other experiences (Sulis et al., 2016) whose aim was to infer a user's gender, age, native language or personality traits, by analyzing the respective texts.

In this paper, the InsideOut project is presented. It explores the possibility to use Social Web mining methodologies and technologies to collect information about adolescents' distress from their micro-blogging activities. The project is promoting a complex language processing workflow to approach the collection, enrichment of user generated contents on Twitter: messages written by a set of targeted community of users (e.g. from a school) are enriched with semantic metadata reflecting the expressed topics (e.g. social vs intimate relationships) and the attitude of the writers. The goal is to use this large scale evidence to support a comprehensive psychological characterization of adolescent communities and to pave the way towards effective applications of preventive and intervention efforts. The general architecture of the InsideOut Web Platform and the resources produced by an integrated effort of computer science specialists and mental health professionals will be presented. These data supported the exploratory evaluation where inter-annotation agreement scores and the performance over real data in the task of psychologically enriching user writings have been obtained.

In the rest of the paper, Section 2 describes the overall workflow underlying the InsideOut Platform. Section 3 describes the semantic models at the base of the semantic annotation process whose first result is the annotated corpus and the exploratory evaluation presented in Sections 4 and 5, respectively. Section 6 derives the conclusion.

2 The InsideOut Web Platfrom

The InsideOut Web Platform aims at supporting mental health studies concerning the causes of distress in adolescents. To this aim, a comprehensive service-oriented architecture has been designed and implemented to collect messages from Social Networks (such as Twitter) written by targeted communities of adolescents and enrich them with semantic information reflecting discussed topics and corresponding attitudes of the writers.

This enables specific kinds of queries and data aggregations, such as the pie chart shown in Figure 1, which summarizes the topics discussed by a community of users, e.g. concerning SCHOOL, FAMILY, or ALCHOOL AND DRUGS. By selecting a specific topic, such as SCHOOL, the system shows only those messages where the writer expresses a specific attitude, such as a DISTRESS. In the same Figure, the distressful messages concerning school are shown, such as "*Questa scuola fa schifo...*" ("*This school sucks...*") or "*Devo studiare.*" ("*I have to study.*").

In order to enable such queries the following services have been implemented:

Data collection services: services dealing with the extraction of data (messages/user information) from targeted social networks. These services are designed both to collect messages referring to a specific topic or hashtag, such as "*#maturità*" or messages exchanged between users belonging to specific *communities*, such as a members of a targeted school class. Among such services, we also implemented *Author Profiling* services that automatically determine the age of the writers (e.g. to filter adolescent's messages) but these specific services are out of the scope of this work.

Semantic annotation services: services dealing with the semantic annotation of gathered messages; once downloaded, they are automatically annotated with the semantic metadata described in the next section.

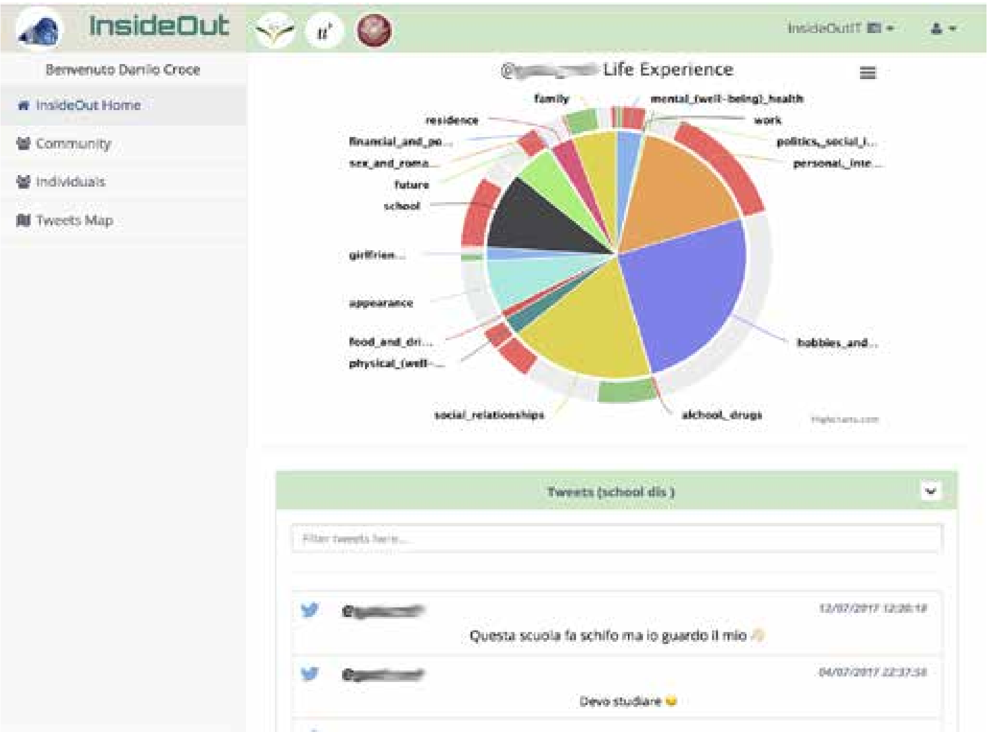
Storage services: services to store (possibly large-scale) collections of messages, communities and semantic metadata in NoSQL databases, implemented in MongoDB.

Reporting services GUI: services that aggregate messages, metadata and users to enable advanced report, such as shown in Figure 1.

3 Distress Characterization: The semantic modeling

In order to synthesize the amount of information made available on Social Media, we need to look at different semantic dimensions that can be associated with the writer's emotion, sentiment and mental status. Given that no direct diagnosis about mental health of an individual can be traced from or over one single message (but it is rather inspired by the observation of behaviors across temporal and social dimensions) we need to frame the mental state related information observable in Social Media within a comprehensive description of a subject.

So we decided to focus on the *experiential dimension* and start from the so-called **Life Event** dimension that expresses topics of interest and daily events in a young person's life. At the moment of writing, these have been discretized in eighteen different classes, as listed in Table 1. Each message can be assigned to one or more classes characterizing the possibly multiple topics



[redacted] 12/07/2017 12:26:18

Questa scuola fa schifo ma io guardo il mio

[redacted] 04/07/2017 22:37:58

Devo studiare

Figure 1: The InsideOut Interface

that can be mentioned in a message. For example, in the message "*Odio la scuola ma adoro i miei compagni*" ("*I hate school but I love my classmates*") the writer refers to the SCHOOL and SOCIAL RELATIONSHIP life events.

Moreover, a **Subjective** emotional dimension is targeted to capture the way the subject relates to the event in the micro-blog he writes, i.e., whether it is related to as a clearly positive or negative event, as a rather neutral statement, or in an ironic way. We referred to the traditional modeling for subjectivity analysis (Rosenthal et al., 2017; Barbieri et al., 2016), adopting POSITIVE, NEGATIVE and NEUTRAL classes; as an example "*Odio la scuola*" ("*I hate school*") is NEGATIVE, while "*Domani la scuola è chiusa*" ("*Tomorrow my school is closed.*") is NEUTRAL.

Finally, a further dimension called **Experience** tried to capture the writer's personal affect towards an event, e.g., whether it (*i*) is causing distress or other negative feelings such as anger or sadness, (*ii*) is regarded as helpful or causing positive feelings such as happiness or affection or (*iii*) is not associated with any perceivable emotional reaction (neutral). As an example, a school performance can be a positive experience if satisfactory for the teacher or the parents, thus being experienced as helpful by the writer, while it might be experienced as a negative event and as distressing when teacher's or parent's judgment is negative. It is worth noting that the Subjective and Experience dimension are nevertheless correlated, but they target different kinds of perception: the following message "*Mi sono rotto una gamba.*" ("*I broke my leg.*") can be considered DISTRESSFUL for the writer even if no agreement or rejection is made w.r.t. the event.

The information observable in a tweet is thus mapped into a set of three independent dimensions: (i) the type of Life Events le the message relates to (ii) the sentiment s of the event (POSITIVE, NEGATIVE, NEUTRAL) and (iii) experience-level e related to the event (among HELPFUL, DISTRESSFUL or NEUTRAL). For example, the tweet "*Quanto odio la mia classe... per fortuna mia sorella mi aiuta!*" ("*I hate my class so much... thankfully, my sister helps me!*") is assigned to the (le,s,e) triples: (SCHOOL, NEGATIVE, DISTRESSFUL) and (FAMILY, POSITIVE, HELPFUL).

4 The InsideOut Annotated Corpus

In the annotation process, annotators selected tweets written by adolescents (that have been previously manually validated) both in English and Italian and enriched them with triples (le, s, e), as discussed in the previous section. In the annota-

Table 1: Life Events description

Life Events	Definition
Alchool, Drugs	All actions and ideas involving the misuse of medications or the use of illegal drugs or alcohol.
Appearance	All messages related to the physical appearance of the writer or of other people.
Crime, abuse and mobbing	Thoughts, references and considerations directly connected to the world of crime or that express an attitude or an opinion of the adolescent towards that sphere.
Family	Events involving or statements related to the family members, such as parental habits, relationships, generational clashes.
Financial and Possession	Event related to the financial status of the young person or his own family; needs of money for important needs or expectations; strongly perceived needs that strictly depend on the economic status and capability of the subject or his family.
Food and drink	All actions and ideas involving food and drink (not alcohol).
Future	Events or thoughts related to the perception the adolescent has about his own future.
Girlfriend/boyfriend personal relationship	(Usually strongly emotional) relationships based on sentimental and sexual attraction, involving gender aspects.
Hobbies and interests	All events, expectations or preferences evoked by entertainment related activities or personal interests (e.g. hobbies, fun, VIP) usually producing fun or connected with time-consuming helpful activities (e.g. games, TV, Social media, Celebrities).
Mental (Well-being) Health	Expressions related to mental well-being and to the health dimension but not related to physical aspects; this class includes sleep problems.
Personal, Internal stressors, beliefs	General opinions, convictions or beliefs of the subject related to his own feelings and his personal sphere; general considerations regarding emotions, spirituality, stressors but not politics or social issues.
Physical (Well-being) Health	Thoughts, complains, considerations related to the physical health dimension, including conditions, nutrition, diseases, remedies and treatments.
Politics, social issues, etc	All thoughts, considerations, reports regarding social, political and anthropological aspects of the close or general environment, as perceived by the young person.
Residence	Every perception about the locations where the subject lives or spends most of his time, including environmental aspects or weather.
School	All events dominated by the school experience (comprehends social interactions IF only limited to school environment).
Sex and Romance	Events or experience specifically grounded at the sexual level, not including boyfriend-hood.
Social relationships	All thoughts and events related to the relational dimension of the young people, but not involving the family, the criminal, the working/school and the boyfriend-hood dimension.
Work	All events related to the relational dimension of the young people, caused or maintained alive by activities or dependencies based on the working condition of the subject of a member of his family.

Table 2: Corpus Statistics

Language	Italian	English
Number of tweets	2,037	1,072
- at least two annotators	1,074	1,072
- only one annotator	963	-
# annotators	4	4
# of (le, s, e) triples	2,517	2,811
Avg (le, s, e) for tweet	1.2	2.6
Avg token per tweet	16	15

Table 3: Inter-annotation agreement

Annotators Agreement - IT			
	Precision	Recall	F1
Life Event	85.76%	60.24%	70.76%
Sentiment	72.43%	50.88%	59.77%
Experience	74.28%	52.17%	61.29%
Annotators Agreement - EN			
	Precision	Recall	F1
Life Event	80.69%	56.17%	66.09%
Sentiment	63.77%	44.05%	51.99%
Experience	64.16%	44.35%	52.35%

tion process, each annotator starts by associating one or more le to a message[1] and, for each of them, the corresponding s and e must be provided. Each message was initially annotated by two annotators. After this first stage, the annotators in disagreement were asked to converge, in order to acquire a gold standard dataset. Only for Italian, we extended the dataset with a set of 963 messages that were annotated by only one annotator, without further refinements. The overall statistics of the dataset are shown in table 2.

In order to measure the complexity of the annotation process, we measured the inter-annotation agreement[2]. Given the possibility to associate more than one le to a message, we decided to measure the agreement in terms of Precision, Recall and F1, by considering the annotations confirmed after the agreement step as gold-standard and the initial annotations as measured annotations. Results are shown in Table 3. For the Sentiment and Experience dimension, we only focused on those messages sharing the same le. These agreement scores are quite low, confirming the difficulty of these kinds of analyses in Social Networks. The lowest score is Recall: it means that annotators generally assign a reasonable class, but it is very difficult to be exhaustive: as an example in the tweet "*Odio la gente che mastica rumorosamente. Mi innervosisce troppo!!!*" (*"I hate the people who chew loudly. It makes me very upset!!!"*) has been assigned to Social Relationships by an annotator while to Personal, internal stressors, beliefs by the other one. At the end, both were accepted and added to the gold standard. Agreement measured over the Italian messages is higher if compared with the English counterpart: one of the main reasons for this is due to

[1] Each annotator can associate zero, one or more les to a message.

[2] The inter-annotation agreement considered only messages annotated by at least two annotators.

Table 4: Results concerning the quantitative analysis of messages.

		Life Event			Sentiment	Experience
Lang.	**Tweets**	**Prec.**	**Rec.**	**F1**	**Accuracy**	**Accuracy**
En	1062	76.0%	31.3%	44.0%	62.8%	62.0%
It	1992	72.2%	47.2%	57.1%	67.8%	67.6%

the fact that Italian messages were annotated by native speakers, while English messages were annotated by German native speakers.

5 Exploratory Evaluation

In order to assess the applicability of the annotation process, we measured the quality of the system in the automatic recognition of Life Event (LE), Sentiment and Experience classes. We modeled this problem as a classification task and adopted the Support Vector Machine learning algorithm (Vapnik, 1995) in a One-VS-ALL schema, implemented within the Kernel-based Learning Platform (KeLP), presented in (Filice et al., 2015)[3]. We evaluated the three targeted dimensions of LE, Subjectivity and Experience separately[4] in a 10-Fold cross-validation schema: at each time a fold is selected as test set, while another set is the validation set used to estimated the SVM parameters. Each tweet is modeled by using the following feature representations: a *Bag-of-words* representation, *Bag-of-n-grams* (with $n = 2$ and $n = 3$) and a distributional representation based on Word Embedding (Mikolov et al., 2013) so that a message is the linear combination of its nouns, verbs, adjective and adverbs. For the LE classifier, we built a similar distributional representation of the eighteen LE definitions shown in Table 1: we introduced additional features in terms of the 18-dimensional vector containing the cosine similarity between the distributional representation of a tweet and the LE definitions. For Subjectivity and Experience, we added some specific features, modeling the presence of emoticons, punctuation marks (such as exclamation points), upper case words and elongated words. Moreover, we added features such as the length of the message (in terms of words and characters).

Regarding the LE dimension, we adopted a conservative strategy so that the system assigns a new LE to a message whereas the SVM classifier provides a positive confidence for the corresponding class while no LE is assigned, otherwise. Performance is thus measured in terms of Precision (the percentage of le correctly introduced by the system), Recall (the percentage of le from the oracle that have been correctly recovered) and F1 (the harmonic mean between Precision and Recall)[5]. Regarding the Subjective and Experience dimensions, once a le is known, the classifier is always requested to associate a message to the s and e labels, in order to generate consistent triples in the form (le, s, e). Being a multi-classification schema were the classifier always outputs a class, Precision is always equal to Recall[6], as well to the F1. In order to avoid redundancy, only one measure is reported and it is referred as Accuracy as it also corresponds to the percentage of messages correctly associated to the gold-standard label.

Preliminary results are shown in Table 4, both for English and Italian. Regarding the LE dimension, the adopted strategy results in a Precision higher than 70%, but at in a lower Recall. We believe this is mainly due to the reduced size of the dataset: it is even more relevant for English where only a 31% of Recall was detected. This number is consistently higher for the Italian dataset, where almost the double of examples is in fact provided and almost half of the tweets were only annotated by one person, thus reducing the odds for differences in annotations. Anyway, these results are consistently higher with respect to a baseline: the correct LE classification given the random selection from 18 classes would achieve a F1 no higher than 3%; if we require two correct classifications, in line with the average le per tweet shown in Table 2, this baseline drops to 0.3%. Moreover, it is worth noting that the adopted conservative strategy has been adopted to have a higher precision: since we are able to collect a huge amount of messages from social network, we can afford to lose

[3] Available at `www.kelp-ml.org`.

[4] When considering Subjectivity and Experience, a gold standard Life event is assumed.

[5] Since a message could be associated to multiple le the evaluation is not message-based but annotation-based.

[6] It may be the case that the LE classifier produces a number of les different from the number of the ones provided in the gold-standard. As a consequence, when evaluating this specific classifier, each message potentially introduces a different number of false positives and false negatives, so Precision and Recall will diverge.

some messages (often characterize by too little information in very short messages) instead of introducing too many noisy meta-data in the overall workflow. Results concerning sentiment are generally consistent with respect to international benchmark in English (Rosenthal et al., 2017) or in Italian (Barbieri et al., 2016) where almost all systems achieved an Accuracy between 60% and 65% (even using larger datasets). Overall, this result seems to be significant, as in line with the first outcome of the inter-annotation agreement. However, a further analysis is required to adopt more complex models for classification of such short messages, such as more complex kernels (Agarwal et al., 2011) or deep methods (Kim, 2014).

6 Conclusions

This paper summarizes the InsideOut project where the possibility to use Social Web mining methodologies and technologies to gather evidence about the adolescents' mental distress.The semantic model defined here and the annotated resource pave the way to a long-term joint research between computer science specialists and mental health professionals. The outcomes suggest the applicability of the devised methodology to larger communities and different languages. Since the system is currently active over Twitter, the final version of the paper will discuss about 5 months of continuous monitoring outcomes towards Italian and English speaking communities, with interesting evidences about the future of our project as a novel and ambitious Social Computational Science application.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Paul Best, Roger Manktelow, and Brian Taylor. 2014. Online communication, social media and adolescent wellbeing: A systematic narrative review. *Children and Youth Services Review*, 41:27 – 36.

Simone Filice, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2015. Kelp: a kernel-based learning platform for natural language processing. In *Proceedings of ACL: System Demonstrations*, Beijing, China, July.

Fiona M Gore, Paul JN Bloem, George C Patton, Jane Ferguson, Vronique Joseph, Carolyn Coffey, Susan M Sawyer, and Colin D Mathers. 2011. Global burden of disease in young people aged 1024 years: a systematic analysis. *The Lancet*, 377(9783):2093 – 2102.

Christian Kieling, Helen Baker-Henningham, Myron Belfer, Gabriella Conti, Ilgi Ertem, Olayinka Omigbodun, Luis Augusto Rohde, Shoba Srinath, Nurper Ulkuer, and Atif Rahman. 2011. Child and adolescent mental health worldwide: evidence for action. *The Lancet*, 378(9801):1515–1525.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*, pages 1746–1751.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd author profiling task at pan 2014. In *CLEF evaluation labs and workshop*, pages 898–927.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, pages 1–8.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. *Working Notes Papers of the CLEF*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August. Association for Computational Linguistics.

Emilio Sulis, Cristina Bosco, Viviana Patti, Mirko Lai, Delia Irazú Hernández Farías, Letizia Mencarini, Michele Mozzachiodi, and Daniele Vignoli. 2016. Subjective well-being and social media. A semantically annotated twitter corpus on fertility and parenthood. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016), Napoli, Italy, December 5-7, 2016.*

V Vapnik. 1995. The nature of statistical learning theory.

"Gimme the Usual" - How Handling of Pragmatics Improves Chatbots

Alessia Bianchini, Francesco Tarasconi,
Raffaella Ventaglio, Mariafrancesca Guadalupi
CELI Language Technology
{bianchini,tarasconi,ventaglio,guadalupi}@celi.it

Abstract

English. We provide our view on the components needed for both the development and further improvement of robust and effective chatbots. We focus on why Pragmatics is important in developing next generation chatbots by bringing a few generalizable examples. We report our current experience on the design and implementation of a task-oriented textual chatbot for a closed-domain Question Answering system, which tackles problems in Pragmatics.

Italiano. *Forniamo la nostra visione su quali sono i componenti necessari per realizzare e migliorare chatbot robusti ed efficaci. Ci concentriamo sul perché la pragmatica sia importante nello sviluppo di chatbot di nuova generazione portando esempi generalizzabili. Riportiamo la nostra esperienza nella progettazione e implementazione di un chatbot testuale task-oriented per un sistema di Question Answering a dominio chiuso che affronta problemi di pragmatica.*

1 Why Pragmatics Matters in Chatbots

Chatbot, chatterbot, natural-language interface, dialogue system are some of the terms used to refer to softwares that aim to carry on conversations with humans (Mauldin, 1994; Lester, Branting and Mott, 2004; Boualem, Casati and Toumani, 2004). We will not go into further details about the classification and definition of such softwares. We will use *chatbot* as if it was a hypernym of the above mentioned softwares instead.

Chatbots and Intent Understanding. The goal of an intelligent chatbot is to understand the user's intent (Yue, 2017) and behave accordingly. Such goal is quite complex to achieve, and beyond the capability of current state of the art chatbots. However, the hype around chatbots has raised awareness of what elements are needed for a chatbot to manage human-like interactions. It is generally agreed that to build effective and solid chatbots the following is needed:

Natural Language Processing (NLP): much of the intelligence needed to understand human intent lies in the processing of human language. Hence, the development and improvement of NLP algorithms is a necessary prerequisite for the creation of intelligent chatbots.

Machine Learning (ML): chatbot design should rely on ML for learning and automatically consolidating NLP rules by means of observation of past experience - i.e., past conversations and their outcomes (Perez-Marin, 2011). Current chatbot development, given enough annotated data, should consider adopting recently developed algorithms that are task-oriented (Bordes and Weston, 2016) or topic aware (Xing, Chen et al., 2017). Developments in reinforcement learning applications seem promising for task-oriented dialogue systems (Rieser and Lemon, 2011).

Context and State Awareness: depending on the purpose of the chatbot, the component responsible for the managing of the conversation (Dialogue Manager System - DMS) should take into account both context and states variables (Allen, Byron, Dzikovska, Ferguson, Galescu and Stent, 2001). From the DMS point of view, chatbots are usually classified as: stateless chatbots; semi-stateful chatbots; stateful chatbots (next generation chatbots). During the conversation, state transitions depend on the information acquired before. As for the follow up action, it depends on the recognized context.

Natural Language Generation (NLG): NLG concerns what information and in what form it should be delivered (Breen, 2014). Dealing with "real" conversation requires being both proactive (e.g. suggest the best option; drive along the compilation of a form; remind planned activities; ...) (Owen et al., 2001) and adaptive (e.g. change style - both in written and spoken scenarios - according to domain, mood of the user, or sociolinguistic variables).

We argue here that, in addition to the above mentioned, moving from Semantics to Pragmatics plays a crucial role in building chatbots. This is because a lot of the knowledge human beings share during a conversation gets constructed along the conversation itself (Robyn, 2002; Pask, 1975). For instance, let's consider the following mock dialogue between Human (H) and Chatbot (C):

H lands on a money transfer service page.
H: Hello, I would like to make a transfer
C: Hello. Sure. Would you like to know more about: [FAQ menu about transfer service is shown]?
H navigates the FAQ menu
H signs up online to proceed with the transfer.
H: I would like to make a transfer
C: Sure. It only takes a couple of minutes
C starts the procedure to execute the transfer.

In this interaction, the sentence “I would like to make a transfer” instantiates two different intents: informative intent, at first (H is looking for information about the transfer service); follow up intent, then (once H is satisfied with transfer service conditions, he/she wants to proceed with the transfer). Such *pragmatic disambiguation* involves taking knowledge from the conversational context into account, which is one of the most difficult tasks for a chatbot. We report below how we deal with this task in our task-oriented closed-domain chatbot.

2 Intent Understanding in Practice

Understanding intents implies handling both semantic meaning and pragmatic meaning. Roughly speaking, while semantics concerns the meaning of a sentence from the linguistic point of view, Pragmatics concerns the interpretation of the same sentence depending on extralinguistic knowledge (Grice, 1975). As mentioned before, a sentence can be ambiguous from the intent point of view. As for classifying intents, there seems to be no comprehensive literature about it, yet - not from the chatbot perspective, at least. However, based on our business experience, we would arrange intents as follows:

Informative Intent: the user is looking for information; e.g. *Question and Answer* (QA), FAQ browsing typically instantiate this intent.

Follow Up Intent: as in regular conversations, the user wants to "do things with words" (Austin, 1962), perform actions; e.g. "Call the call center", "Order pizza", "Turn on washing machine".

Dialogic Intent: the user uses *discourse markers* to connect, organize, manage the conversation; e.g.: greetings, farewells, turns markers, ...

Regular expressions, pattern matching and keyword recognition typically are not enough to achieve *real* intent understanding. This is because the more the interaction is *human-like*, the more complicated it becomes to figure out what the human really wants. Among business intent, real life cases we faced are: Onboarding, Question Answering and Education. In our applications, we break down the understanding process into subtasks. Namely: *intent classification* (e.g. “booking a flight”); *slot filling*, i.e. enriching the intent with more detailed information (such as “destination” and “departure time”); *context modeling*, i.e. keeping track of context to get to the correct meaning (“time” might refer to “flight departure time”, “flight arrival time”, “dinner time”, etc...).

3 System Design and Architecture

Our task-oriented closed-domain financial textual chatbot, *Financial QA Chatbot*, aims to provide users with answers concerning banks and insurances, through a conversation in Italian. The type of answers that a user can obtain are similar to the ones found on a financial platform website[1]: this portal provides a search engine and FAQ section to satisfy the information need. Therefore, it is mainly a QA chatbot, although some additional follow up actions are available on top of providing an answer to questions, such as redirecting to specific websites or services. Financial QA is provided with a proprietary scoring algorithm to match the current user's questions to answers in a database A. In line with previous work (Quarteroni and Manandhar, 2007), we will review key design and architecture aspects, with emphasis on possible solutions to Pragmatics problems discussed in sections 1 and 2. In this sense, the most significant components are the Dialogue Manager and the Context Manager, which provide the scoring algorithm enriched information.
NLP functions such as normalization, tokenization, lemmatization, POS tagging, disambiguation and dependency parsing are made available through the CELI linguistic pipeline[2] (Tarasconi and Di Tomaso, 2015).

Dialogue Scenario: a QA session consists of *actions* that can be performed by the user or by the automated system, according to *Dialogue Management* logic.
User actions: greet, quit, ask a question q, acknowledge the previous utterance, ask for help/suggestions, browse the navigation menu.
System actions: greet, quit, present answer a, acknowledge the previous utterance, ask for clarifications, propose a follow up (question/action), reprimand for using swearwords, suggest questions, present or hide the navigation menu.

User's action classification: each user's utterance is classified into one of five action classes: greet, quit, ask a question, acknowledge, ask for help. This is accomplished using predefined dictionaries and automatic classifiers, which also consider discourse markers and *disfluencies*.
Although there is promising work done on dialog act detection with multi-level information (Rosset et al., 2008), in this step with adopt a simpler approach, leaving further refinements to subsequent components.

Dialogue Management: the conversation proceeds along these logics.

1. An initial greeting (*greet* action), a request for help (*ask for help*) or a direct question q (*ask a question*) from the user.

2. The system, if asked for help, presents the user with a navigation menu, based on current context and on the given hierarchical classification of contexts or topics (see *Context Management* below). This menu can be browsed until a terminal node in the classification is reached, and, at that point, a predefined set of questions related to that topic is suggested. The user can select a question q from that list.

3. q is analyzed to detect *wh-type* (Huang et al., 2008) and whether it is elliptic or anaphoric. This information is passed along with q and the current context to the subsequent QA component.

4. The QA component searches for matches of the query according to the *QA Algorithm*. Each matching answer a_k is accompanied by a $relevance$ score r_k, $r_k \in (0, 1]$. If at least one match has relevance more than a fixed threshold T, only the best match (highest relevance) is returned. Otherwise, up to the top N_r highest results are returned by the QA component. In Financial QA's basic settings, $T = 0.75$ and $N_r = 5$.

5. The QA component results are processed: they can be a single answer or, because of low relevance scores, a list of answers. If a single answer is provided by the QA component, it is returned to the user (*answer* action). In the case of a list, the user is asked for clarifications, and a single answer is selected based on her additional input (*ask for clarifications* action, then *answer* action). After an answer is provided to the user, context is updated accordingly.

6. The system inquires whether the user is interested in a follow up session; if this is the case,

[1] gooruf.com
[2] www.celi.it

the user can enter a question again. Else, the system acknowledges.

7. Whenever the user wants to terminate the interaction, a final greeting is exchanged (*quit* action).

Context Management: intuitively, all the answers a in the knowledge base are grouped in disjoint topics of maximum granularity, which are then organized in a hierarchical structure, used to model context in this QA task.
Managing topic hierarchies can improve performance in a query matching system (Domingues et al., 2014). Formally, context elements are topics of conversation belonging to the finite set $C = \{C_1, \dots, C_N\}$. Topics are arranged in a hierarchical classification structure, which can be represented as a tree $T = (C, E)$, where C is the set of nodes. Edges E express the "C_i has subclass C_j" relation. A context X is, in general, an arbitrarily ordered sequence of topics.
In our current implementation of Financial QA, we support only contexts of length 1, therefore the context X_s at step s of the conversation is the position C_s in T. We assume all interactions start at the root node C_0. X_s is meant to represent the current topic of conversation at step s, according to the last answer provided or the latest click on the navigation menu. By supporting contexts of length > 1, it is also possible to keep track of previous topics of conversation.
Each node C_i has a corresponding nonempty set of topic-related *keywords* W_i.
An important distinction is drawn between *terminal nodes* C^Ω and *nonterminal nodes* $C \setminus C^\Omega$. Each terminal node C^Ω_j has a (potentially empty) set of answers A_j corresponding to it. All the A_j sets are disjointed. Let A be the set of all answers: $A = \cup_{j \in 1,\dots,\omega} A_j$.
In our current implementation, there are 35 classification nodes arranged on 3 levels, 25 of them are terminal ones; the number of answers in the knowledge base is 440, and growing
In the Financial QA chatbot, two types of moves between contexts in C are allowed:

1. To children nodes or root node: using the interactive navigation menu. Context is updated automatically according to the user's selection.
   Example: *You are in the "People" section. Ask me a question or choose one of the following topics:*
   (a) members
   (b) influencers
   (c) contact us
   (d) return to main menu

2. To any terminal node: after answer a_k is provided to the user by the system, new context becomes C_j, where A_j contains a_k.
   Example: after providing the answer *COST_OF_GOORUF = "Gooruf is free, only Premium Providers are required to pay"*, context is changed to *ROOT* → *services* → *info_about_gooruf*, *info_about_gooruf* being the terminal topic containing answer *COST_OF_GOORUF*.

QA Algorithm Design: question q, its wh-type h, and its current context C_s are passed to the algorithm. Keywords $\mathbf{w}_q$ are extracted from q. If q is anaphoric or elliptic, the algorithm evaluates whether to expand W_q to $\dot{\mathbf{w}}_q$ by using keywords W_s corresponding to C_s. The final representation of q is:

$$R(q) = (C_s, h, \dot{\mathbf{w}}_q).$$

Answers $a \in A$ are described by the following feature vector:

$$F(a) = (C^\Omega_a, H_a, W_a)$$

where C^Ω_a is the classification terminal node corresponding to a, W_a the corresponding keywords and H_a a set of related wh-type (for example a user might inquire about Gooruf by referring to it as a *what* or a *who*).
Relevance r_k for each answer a_k is computed, by considering the classification structure T as well, therefore:

$$r_k(q) = \rho(R(q), F(a_k), T).$$

To compute ρ, scores are calculated separately by comparing contexts (using proximity in T between C_s and C^Ω_a in T), wh-types (h and H_a) and a dense semantic representation of keywords ($\dot{\mathbf{w}}_q$ and W_a) obtained using a Word2Vec model for Italian language (Mikolov et al., 2013); before these partial scores are weighted and summed.

Example: we provide below an example of Financial QA interaction which shows how managing hierarchical context helps in accomplishing

the question answering task.
A subtree *taxes* of T models Italian taxes-related topics:

- *taxes* $\rightarrow$ *city_taxes*
  - *TARI*
  - *TASI*
- *taxes* $\rightarrow$ *income_taxes*
  - *IRPEF*
  - *IRAP*

Individual taxes are represented as terminal nodes in T. Let $C^{\Omega}_{\text{taxes}} = \{$*IRPEF, IRAP, TARI, TASI*$\}$. Each $t \in C^{\Omega}_{\text{taxes}}$ has associated answers: "*HOW_TO_PAY* t", "*WHERE_TO_PAY* t", "*AMOUNT_TO_PAY* t".
The interaction could go as follows:

H: How can I pay city taxes?
QA Algorithm detects wh-type *how*, keywords matching the *city_taxes* node, finds two relevant answers in children nodes and Chatbot asks for clarification.
C: Did you mean TARI or TASI?
H: the second one
Chatbot presents answer *"HOW_TO_PAY TASI"*
Context is now *TASI*.
H: and where can I pay it?
QA Algorithm detects wh-type *where* and completes the question with context knowledge.
Chatbot finds a single relevant answer and presents answer *"WHERE_TO_PAY TASI"*.
H: how much IRPEF should I pay?
Chatbot presents *"AMOUNT_TO_PAY IRPEF"*.
Context is now *IRPEF*.
H: where can I pay it?
Chatbot presents *"WHERE_TO_PAY IRPEF"*.

4 Conclusions and Further Work

We are currently in the process of evaluating Financial QA according to a framework based on PARADISE (Walker et al., 1997; Rieser and Lemon, 2011), which considers, among the others, the following indicators: Task Ease, NLU Performance, Expected Behavior, Presentation, Verbal Presentation, Future Use. We plan to finalize our evaluation in the next months.
NLP is crucial for the development of robust chatbots; since extra-linguistic elements are potentially very important in intent understanding, moving from semantics to pragmatics is a necessary step to develop next-generation chatbots. We have shown how Dialogue Management can support a more robust handling of context, at least in closed-domain QA tasks.
Further work is required to handle more business cases and a broader definition of context, such as history of activities conducted by the same user, which can be especially useful in chatbots with recommender functions (Lombardi et al., 2009).

We would like to thank Andrea Bolioli for his help in the review phase and all of our CELI colleagues for their invaluable work and support.

References

John L. Austin. 1962. *How to Do Things With Words*. Oxford University Press.

James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu and Amanda Stent. 2001. *Toward conversational human-computer interaction. AI magazine*, 22(4):27–39.

Antoine Bordes and Jason Weston. *Learning End-to-End Goal-Oriented Dialog*. 2016. arXiv:1605.07683 [cs.CL].

Benatallah Boualem, Fabio Casati, and Farouk Toumani. 2004. *Web service conversation modeling: A cornerstone for e-business automation. IEEE Internet computing*, 8(1):46–54.

Andrew Breen 2014. *Creating Expressive TTS Voices for Conversation Agent Applications. International Conference on Speech and Computer*. Springer.

Marcos Aurélio Domingues, Marcelo Garcia Manzato, Ricardo Marcondes Marcacini, Camila Vaccari Sundermann and Solange Oliveira Rezende. 2014. *Using contextual information from topic hierarchies to improve context-aware recommender systems. Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 3606–3611.

Paul Grice. 1975. *Logic and conversation*. New York Academic Press, 41–58.

Zhiheng Huang, Marcus Thint and Zengchang Qin. 2008. *Question Classification Using Head Words and Their Hypernyms. Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

James Lester, Karl Branting and Bradford Mott. 2004. *Conversational agents. The Practical Handbook of Internet Computing*, 220–240.

Sabrina Lombardi, Sarabjot Singh Anand and Michele Gorgoglione. 2009. *Context and customer behaviour in recommendation.*

Michael Mauldin. 1994. *ChatterBots, TinyMuds, and the Turing Test: Entering the Loebner Prize Competition. AAAI Press - Proceedings of the Eleventh National Conference on Artificial Intelligence*.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space.* arXiv:1301.3781 [cs.CL].

Rambow Owen, Srinivas Bangalore and Marilyn Walker. 2001. *Natural language generation in dialog systems. Proceedings of the first international conference on Human language technology research.*

Carston Pask. 1975. *Conversation, cognition and learning*. New York: Elsevier.

Diana Perez-Marin. 2011. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices.* IGI Global.

Silvia Quarteroni and Suresh Manandhar. 2007. *A Chatbot-based Interactive Question Answering System. Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue.* Edited by Ron Artstein and Laure Vieu.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media.

Carston Robyn. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford Blackwell.

Sophie Rosset, Delphine Tribout and Lori Lamel. 2008. *Multi-level information and automatic dialog act detection in human–human spoken dialogs. Speech Communication*, 50(1):1-3.

Francesco Tarasconi and Vittorio Di Tomaso. 2015. *Geometric and statistical analysis of emotions and topics in corpora. IJCoL - Italian Journal of Computational Linguistics*, Vol.1 n. 1. Accademia University Press.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. *Topic Aware Neural Response Generation. AAAI*, 3351–3357.

Gu Yue. 2017. *Speech Intention Classification with Multimodal Deep Learning. 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Proceeding*. Springer.

Marilyn A. Waler, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. *PARADISE: A framework for evaluating spoken dialogue agents. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Deep-learning the Ropes: Modeling Idiomaticity with Neural Networks

Yuri Bizzoni[1], **Marco S. G. Senaldi**[2], **Alessandro Lenci**[3]
University of Gothenburg - Sweden[1], Scuola Normale Superiore - Italy[2], University of Pisa - Italy[3]
yuri.bizzoni@gu.se[1], marco.senaldi@sns.it[2], alessandro.lenci@unipi.it[3]

Abstract

English. In this work we explore the possibility of training a neural network to classify and rank idiomatic expressions under constraints of data scarcity. We discuss our results comparing them both to other unsupervised models designed to perform idiom detection and to similar supervised classifiers trained to detect metaphoric bigrams.

Italiano. *In questo lavoro esploriamo la possibilità di addestrare una rete neurale per classificare ed ordinare espressioni idiomatiche in condizioni di scarsità di dati. I nostri risultati sono discussi in comparazione sia con altri algoritmi non supervisionati ideati per l'identificazione di espressioni idiomatiche sia con classificatori supervisionati dello stesso tipo addestrati per identificare bigrammi metaforici.*

1 Introduction

Figurative expressions like idioms (e.g. *to learn the ropes* 'to learn how to do a job', *to cut the mustard* 'to perform up to expectations', etc.) and metaphors (e.g. *clean performance*, *that lawyer is a shark*, etc.) are pervasive in language use. Important differences have been stressed between the two types of expressions from a theoretical (Gibbs, 1993; Torre, 2014), neurocognitive (Bohrn et al., 2012) and corpus linguistic (Liu, 2003) prespective. On the one hand, as stated by Lakoff and Johnson (2008), linguistic metaphors reflect an instantiation of conceptual metaphors, whereby abstract concepts in a *target domain* (e.g. the ruthlessness of a lawyer) are described by a rather transparent mapping to concrete examples taken from a *source domain* (e.g. the aggressiveness of a shark). On the other hand, although most idioms originate as metaphors (Cruse, 1986), they have undergone a crystallization process in diachrony, whereby they now appear as fixed and non-compositional word combinations that belong to the wider class of Multiword Expressions (MWEs) (Sag et al., 2002) and always exhibit lexical and morphosyntactic rigidity to some extent (Cacciari and Glucksberg, 1991; Nunberg et al., 1994). It is anyway crucial to underline that idiomaticity itself is a multidimensional and gradient phenomenon (Nunberg et al., 1994; Wulff, 2010) with different idioms showing varying degrees of semantic transparency, formal versatility, proverbiality and affective valence.

The aim of this work is to explore the fuzzy boundary between idiomatic and metaphorical expression, by applying a method designed to discriminate figurative vs. literal usages to the task of distinguishing idiomatic from compositional expressions. Our starting point is the work of Bizzoni et al. (2017). The authors managed to classify adjective-noun pairs where the same adjectives were used both in a metaphorical and a literal sense (e.g. *clean performance* vs. *clean floor*) using a neural classifier trained on a composition of the words' embeddings (Mikolov et al., 2013). Actually, the neural network was able to detect the abstract/concrete semantic shift of nouns when used with the same adjective in figurative and literal compositions respectively, basically treating the noun as the "context" to discriminate the metaphoricity of the adjective. In our attempt, we will use a relatively similar approach to classify idiomatic expressions by training a three-layered neural network on a set of idiomatic and non-idiomatic expressions and we'll compare the performance of the network when trained on different syntactic patterns (Adjective-Noun and Verb-Noun expressions, AN and VN henceforth).

Importantly, the abstract/concrete polarity the network was able to learn in Bizzoni et al. (2017) will not be available this time, since none of the idiom constituents will ever appear in its literal sense inside the expressions, whatever their concreteness may be. What we want to find out is whether the sole information captured by the distributional vector of a single expression is sufficient to learn its potential idiomaticity. Differently from Bizzoni et al. (2017), for each idiom we collect a count-based vector (Turney and Pantel, 2010) of the expression as a whole, taken as a single token. We compare this approach with a model trained on the composition of the individual words of an expression, showing that the latter is less effective for idioms than for metaphors. In both cases we will be operating on scarce training sets (26 AN and 90 VN constructions). Traditional ways to deal with data scarcity in computational linguistics resort to a wide number of different features to annotate the training set (see for example Tanguy et al. (2012)) or rely on artificial bootstrapping of the training set (He and Liu, 2017). In our case we test the performance of our classifier on scarce data without bootstrapping the dataset and relying only on the information provided by the distributional semantic space, showing that the distribution of an expression in large corpora can provide enough information to learn idiomaticity from few examples with a satisfactory degree of accuracy.

2 Related Work

Previous computational research has exploited different methods to perform *idiom type detection* (i.e., automatically telling apart potential idioms like *to get the sack* from only literal combinations like *to kill a man*). For example Lin (1999) and Fazly et al. (2009) label a given word combination as idiomatic if the Pointwise Mutual Information (PMI) (Church and Hanks, 1991) between its constituents is higher than the PMIs between the components of a set of lexical variants of this combination obtained by replacing the component words of the original expressions with semantically related words. Other studies have resorted to Distributional Semantics (Lenci, 2008; Turney and Pantel, 2010) by measuring the cosine between the vector of a given phrase and the single vectors of its components (Fazly and Stevenson, 2008) or between the phrase vector and the sum or product vector of its components (Mitchell and Lapata, 2010; Krčmář et al., 2013). Senaldi et al. (2016b) and Senaldi et al. (2016a) have combined insights from both these approaches by observing that the vectors of VN and AN idioms are less similar to the vectors of lexical variants of these expressions with respect to the vectors of compositional constructions. To the best of our knowledge, neural networks have been previously adopted to perform MWE detection in general (Legrand and Collobert, 2016; Klyueva et al., 2017), but not idiom identification specifically. In Bizzoni et al. (2017), pre-trained noun and adjective vector embeddings are fed to a single-layered neural network to disambiguate metaphorical and literal AN combinations. Several combination algorithms are experimented with to concatenate adjective and noun embeddings. All in all, the method is shown to outperform the state of the art, presumably leveraging the abstractness degree of the noun as a clue to metaphoricity.

3 Dataset

3.1 Target expressions extraction

The two idiom datasets we employ in the current study come from Senaldi et al. (2016b) and Senaldi et al. (2016a). The first one is composed of 45 idiomatic and 45 non-idiomatic Italian V-NP and V-PP constructions (e.g. *tagliare la corda* ‘to flee’ lit. ‘to cut the rope’ and *leggere un libro* ‘to read a book’) that were selected from an Italian idiom dictionary (Quartu, 1993) and extracted from the itWaC corpus (Baroni et al., 2009), composed of about 1,909M tokens. Their frequency spanned from 364 (*ingannare il tempo* ‘to while away the time’) to 8294 (*andare in giro* ‘to get about’). The latter comprises 13 idiomatic and 13 non-idiomatic AN constructions (e.g. *punto debole* ‘weak point’ and *nuova legge* ‘new law’) that were still extracted from itWaC and whose frequency varied from 21 (*alte sfere* ‘high places’, lit. ‘high spheres’) to 194 (*punto debole*).

3.2 Building target vectors

Count-based Distributional Semantic Models (DSMs) (Turney and Pantel, 2010) allow for representing words and expressions as high-dimensionality vectors, where the vector dimensions register the co-occurrence of the target words or expressions with some contextual features, e.g. the content words that linearly precede and follow the target element within a fixed contextual window. We built two DSMs on itWaC, where our target AN and VN idioms and non-idioms were represented as target vectors and co-occurrence statistics counted how many times each target construction occurred in the same sentence with each of the 30,000 top content words in the corpus. Differently from Bizzoni et al. (2017), we did not opt for prediction-based vector representations (Mikolov et al., 2013). Although some studies have brought out that context-predicting models fare better than count-based ones on a variety of semantic tasks (Baroni et al., 2014), including compositionality modeling (Rimell et al., 2016), others (Blacoe and Lapata, 2012; Cordeiro et al., 2016) have shown them to perform comparably. Moreover, Levy et al. (2015) highlight that much of the superiority in performance exhibited by word embeddings is actually due to hyperparameter optimizations, which, if applied to traditional models as well, can bring to equivalent outcomes. Therefore, we felt confident in resorting to count-based vectors as an equally reliable representation for the task at hand.

3.3 Gold standard idiomaticity judgments

In Senaldi et al. (2016b) and Senaldi et al. (2016a), we collected gold standard idiomaticity judgments for our target AN and VN constructions. 9 Linguistics students were presented with a list of our 26 AN constructions and were asked to evaluate how idiomatic each expression was from 1 to 7, with 1 standing for ‘totally compositional’ and 7 standing for ‘totally idiomatic’. Inter-coder agreement, measured with Krippendorff’s α (Krippendorff, 2012), was equal to 0.76. The same procedure was repeated for our 90 VN constructions, but in this case the inital list was split into 3 sublists of 30 expressions, each one to be rated by 3 subjects. Krippendorff’s α was 0.83 for the first sublist and 0.75 for the other two.

4 Classifier

We built a neural network composed of three “dense” or fully connected layers[1] of dimensionality 12, 8 and 1 respectively. Our network takes in input a single vector at a time, which can be a word embedding, a count-based distributional vector or a composition of several word vectors. For the core part of our experiment we used as input single distributional vectors of two-word expressions. Due to our input’s magnitude, the most important reduction of data dimensionality is carried out by the first layer of our model. The last layer applies a sigmoid activation function on the output in order to produce a binary judgment. While binary scores are necessary to compute the model classification accuracy and will be evaluated in terms of F1, our model’s continuous scores can be retrieved and will be used to perform an ordering task on the test set, that we will evaluate in terms of Interpolated Average Precision (IAP) [2] and against the human idiomaticity judgments with Spearman’s ρ.

5 Evaluation

We trained our model on the 30,000 dimensional distributional vectors of VN and AN expressions as well as on the composition of their individual words’ vectors. We tried with different semantic spaces as well. When trained on PPMI- (Church and Hanks, 1991) and SVD-transformed (Deerwester et al., 1990) vectors of 150, 200, 250 and 300 dimensions, our models performed comparably or even worse; so, results for these cases won’t be presented here. Details of both classification and ordering task are shown in Table 1.

5.1 Verb-Noun

We ran our model on the VN dataset, composed of 90 elements, 45 idioms and 45 non-idiomatic expressions. This is the larger of the two datasets. We trained our model both on 30 and 40 elements for 20 epochs and tested on the remaining 60 and 50 elements respectively, reaching a maximum IAP of 0.87 and Spearman’s ρ of 0.76. In general we found the model’s performance, both in accuracy and in correlation, comparable to the results reported in Senaldi et al. (2016b), who reached a maximum IAP of 0.91 and a maximum Spearman’s ρ of -0.67.

5.2 Adjective-Noun

We ran our model on the AN dataset, composed of 26 elements, 13 idioms and 13 non-idiomatic expressions. We empirically found that our model was able to perform some generalization on the data when the training set contained at least 14 elements, evenly balanced between positive and negative examples. We trained our model on 16 elements for 30 epochs and tested on the remaining 10 elements. While accuracy’s exact value can

[1] We used Keras, a library running on TensorFlow (Abadi et al., 2016).

[2] Following Fazly et al. (2009), IAP was computed at recall levels of 20%, 50% and 80%.

Vector	Training	Test	IAP	rho	F1
VN	15+15	30+30	0.82	0.50***	0.8
VN	20+20	15+15	0.82	0.76***	0.87
Concat (VN)	15+15	14+14	0.7	0.47*	0.69
AN	8+8	6+4	1?	0.93***	0.9
VN+AN	23+23	14+14(VN)	0.9	0.76***	0.82
VN+AN	23+23	18+20(joint)	0.8	0.64***	0.76
VN+AN	23+23	5+5(AN)	0.57	-0.31	0.58

Table 1: Interpolated Average Precision, Spearman's correlation with the speaker judgments and F-measure for Vector-Noun training (VN), Adjective-Noun training (AN), joint training and training through vector concatenation (** = $p < .01$, *** = $p < .001$). Training and test set are expressed as the sum of positive and negative examples.

undergo some fluctuations when a model is trained on very small sets, we always registered accuracies higher than 80%, with 4 out of 5 idioms correctly labeled in every trial. We reached an IAP of 1.0 and a ρ of 0.93, although it is important to keep in mind that such scores are computed on a very restricted test set. Senaldi et al. (2016b) reached a maximum IAP of 0.85 and a maximum ρ of -0.68. When the training size was under the critical threshold, accuracy dropped significantly. With training sets of 10 or 12 elements, our model naturally went in overfitting, quickly reaching 100% accuracy on the training set and failing to correctly classify unforeseen expressions. In these cases a partial learning was still visible in the ordering task, where most idioms, even if labeled incorrectly, received higher scores than non-idioms.

5.3 Joint training

We also tried to train our model on both datasets together, to check to what extent it would be able to recognize the same underlying semantic phenomenon through different syntactic constructions. We used two different approaches for this experiment. Training our model first on one dataset, e.g. the AN pairs, and then on the other required more epochs overall (more than 100) to stabilize and resulted in a poorer performance (66% F-measure on both test sets). Training our model on a mixed dataset containing the elements of both training sets, our model employed only 12 epochs to reach an F-measure of 76% on the mixed training set. Anyway, we also noticed that VN expressions were learned better than AN expressions. In short, our model was able to generalize over the two datasets, but this involved a loss in accuracy.

5.4 Vector composition

In addition to using the vector of an expression as a whole, we tried to feed our model with the concatenation of the vectors of the single words in an expression, as in Bizzoni et al. (2017). For example, instead of using the 30,000 dimensional vector of the expression *cambiare musica*, we used the 60,000 dimensional vector resulting from the concatenation of *cambiare* and *musica*. We ran this experiment only on the VN dataset, being the largest and the one that yielded the best results in the previous settings. We used 30 elements in training and 26 in testing and trained our model for 80 epochs overall. Predictably enough, vector composition resulted in the worst performance, differently from what happened with metaphors (Bizzoni et al., 2017); nonetheless, the results are not completely random: with an F1 of 69%, the model seems able to learn idiomaticity to a lower, but not null, degree; these findings would be in line with the claim that the meaning of the subparts of several idioms, while less important than in metaphors, is not completely obliterated (McGlone et al., 1994).

6 Error Analysis

Two frequent false positives are *tagliare il traguardo* and *abbassare la guardia*. While we labeled them as non-idioms in our dataset, since they're rather compositional, nonetheless they can be very often used figuratively and that's probably why our algorithms identified them as idioms. A frequent false negative was *vedere la luce*, which probably occurs more often in its literal sense in the corpus we used.

7 Discussion and Conclusions

It seems that the distribution of idiomatic and compositional expressions in large corpora can suffice for a supervised classifier to learn the difference between the two linguistic elements from small training sets and with a good level of accuracy. Unlike with metaphors (Bizzoni et al., 2017), feeding the classifier with a composition of the individual words' vectors of such expressions performs quite scarcely and can be used to detect only some idioms. This takes us back to the core difference that while metaphors are more compositional and preserve a transparent source domain to target domain mapping, idioms are by and large non-compositional. Since our classifiers rely only on contextual features, their ability in classification must stem from a difference in distribution between idioms and non-idioms. A possible explanation is that while the literal expressions we selected, like *vedere un film* or *ascoltare un discorso*, tend to be used with animated subjects and thus to appear in more concrete contexts, most of our idioms (e.g. *cadere dal cielo* or *lasciare il segno*) allow for varying degrees of animacy or concreteness of the subject, and thus their context can easily get more diverse. At the same time, the drop in performance we observe in the joint models seems to indicate that the different parts of speech composing our elements entail a significant contextual difference between the two groups, which introduces a considerable amount of uncertainty in our model. It is also possible that other contextual elements we did not consider have played a role in the learning process of our models. We intend to deepen this aspect in future works.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247.

Bizzoni, Y., Chatzikyriakidis, S., and Ghanimifard, M. (2017). "Deep" learning: Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52.

Blacoe, W. and Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 546–556. Association for Computational Linguistics.

Bohrn, I. C., Altmann, U., and Jacobs, A. M. (2012). Looking at the brains behind figurative language: A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. *Neuropsychologia*, 50(11):2669–2683.

Cacciari, C. and Glucksberg, S. (1991). Understanding idiomatic expressions: The contribution of word meanings. *Advances in Psychology*, 77:217–240.

Church, K. W. and Hanks, P. (1991). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cordeiro, S., Ramisch, C., Idiart, M., and Villavicencio, A. (2016). Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1986–1997.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 1(35):61–103.

Fazly, A. and Stevenson, S. (2008). A distributional account of the semantics of multiword expressions. *Italian Journal of Linguistics*, 1(20):157–179.

Gibbs, R. W. (1993). Why idioms are not dead metaphors. *Idioms: Processing, structure, and interpretation*, pages 57–77.

He, X. and Liu, Y. (2017). Not enough data?: Joint inferring multiple diffusion networks via network generation priors. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 465–474.

Klyueva, N., Doucet, A., and Straka, M. (2017). Neural networks for multi-word expression detection. *Proceedings of the 13th Workshop on Multiword Expressions*, pages 60–65.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

Krčmář, L., Ježek, K., and Pecina, P. (2013). Determining Compositionality of Expresssions Using Various Word Space Models and Measures. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 64–73.

Lakoff, G. and Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Legrand, J. and Collobert, R. (2016). Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 67–71.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324.

Liu, D. (2003). The most frequently used spoken american english idioms: A corpus analysis and its implications. *Tesol Quarterly*, 37(4):671–700.

McGlone, M. S., Glucksberg, S., and Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse Processes*, 17(2):167–190.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26tth International Conference on Neural Information Processing System*, pages 3111–3119.

Mitchell, J. and Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Nunberg, G., Sag, I., and Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

Quartu, M. B. (1993). *Dizionario dei modi di dire della lingua italiana*. RCS Libri.

Rimell, L., Maillard, J., Polajnar, T., and Clark, S. (2016). RELPRON: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics*, 42(4):661–701.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.

Senaldi, M. S. G., Lebani, G. E., and Lenci, A. (2016a). Determining the compositionality of noun-adjective pairs with lexical variants and distributional semantics. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 268–273.

Senaldi, M. S. G., Lebani, G. E., and Lenci, A. (2016b). Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In *Proceedings of the 12th Workshop on Multiword Expression*, pages 21–31.

Tanguy, L., Sajous, F., Calderone, B., and Hathout, N. (2012). Authorship attribution: Using rich linguistic features when training data is scarce. In *PAN Lab at CLEF*.

Torre, E. (2014). *The emergent patterns of Italian idioms: A dynamic-systems approach*. PhD thesis, Lancaster University.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Wulff, S. (2010). *Rethinking Idiomaticity: A Usage-based Approach*. A&C Black.

Neural Sentiment Analysis for a Real-World Application

Daniele Bonadiman[‡], **Giuseppe Castellucci**[†],
Andrea Favalli[†], **Raniero Romagnoli**[†], **Alessandro Moschitti**[‡◇]
[‡]Department of Computer Science and Information Engineering, University of Trento, Italy
[◇]Qatar Computing Research Institute, HBKU, Qatar
[†]Almawave Srl., Italy
d.bonadiman@unitn.it, amoschitti@gmail.com
{g.castellucci,a.favalli,r.romagnoli}@almawave.it

Abstract

English. In this paper, we describe our neural network models for a commercial application on sentiment analysis. Different from academic work, which is oriented towards complex networks for achieving a marginal improvement, real scenarios require flexible and efficient neural models. The possibility to use the same models on different domains and languages plays an important role in the selection of the most appropriate architecture. We found that a small modification of the state-of-the-art network according to academic benchmarks led to a flexible neural model that also preserves high accuracy.

Italiano. *In questo lavoro, descriviamo i nostri modelli di reti neurali per un'applicazione commerciale basata sul* sentiment analysis. *A differenza del mondo accademico, dove la ricerca è orientata verso reti anche complesse per il raggiungimento di un miglioramento marginale, gli scenari di utilizzo reali richiedono modelli neurali flessibili, efficienti e semplici. La possibilitá di utilizzare gli stessi modelli per domini e linguaggi variegati svolge un ruolo importante nella scelta dell'architettura. Abbiamo scoperto che una piccola modifica della rete allo stato dell'arte rispetto ai benchmarks accademici produce un modello neurale flessibile che preserva anche un'elevata precisione.*

1 Introduction

In recent years, Sentiment Analysis (SA) in Twitter has been widely studied. Its popularity has been fed by the remarkable interest of the industrial world on this topic as well as the relatively easy access to data, which, among other, allowed the academic world to promote evaluation campaigns, e.g., (Nakov et al., 2016), for different languages. Many models have been developed and tested on these benchmarks, e.g., (Li et al., 2010; Kiritchenko et al., 2014; Severyn and Moschitti, 2015; Castellucci et al., 2016). They all appear very appealing from an industrial perspective, as SA is strongly connected to many types of business through specific KPIs[1]. However, previous academic work has not provided clear indications on how to select the most appropriate learning architecture for industrial applications.

In this paper, we report on our experience on adopting academic models of SA to a commercial application. This is a social media and microblogging monitoring platform to analyze brand reputation, competition, the voice of the customer and customer experience. More in detail, sentiment analysis algorithms register customers' opinions and feedbacks on services and products, both direct and indirect.

An important aspect is that such clients push for easily adaptable and reliable solutions. Indeed, multi-tenant applications and sentiment analysis requirements cause a high variability of the approaches to the tasks within the same platform. This should be capable of managing multi-domain and multi-channel content in different languages as it provides services for several clients in different market segments. Moreover, scalability and lightweight use of computational resources preserving accuracy is also an important aspect. Finally, dealing with different client domains and data potentially requires constantly training new models with limited time availability.

To meet the above requirements we started from

[1] Key Performance Indicators are strategic factors enabling the performance measurement of a process or activity.

the state-of-the-art model proposed in (Severyn and Moschitti, 2015), which is a Convolutional Neural Network (CNN) with few layers mainly devoted to encoding a sentence representation. We modified it by adopting a recurrent pooling layer, which allows the network to learn longer dependencies in the input sentence. An additional benefit is that such simple architecture makes the network more robust to biases from the dataset, generalizing better on the less represented classes. Our experiments on the SemEval data in English as well as on a commercial dataset in Italian show a constant improvement of our networks over the state of the art.

In the following, Section 2 places the current work in the literature. Section 3 introduces the application scenario. Sections 4 and 5 presents respectively our proposal for a flexible architecture and the experimental results. Finally, Section 6 reports the conclusions.

2 Related Work

Although sentiment analysis has been around for one decade, a clear and exact comparison of models has been achieved thanks to the organization of international evaluation campaigns. The main campaign for SA in Twitter in English is SemEval, which has been organized since 2013. A similar campaign in the Italian language (SENTIPOLC) (Barbieri et al., 2016) is promoted within Evalita since 2014.

Among other approaches, Neural Networks (NNs), and in particular CNNs, outperformed the previous state of the art techniques (Severyn and Moschitti, 2015; Castellucci et al., 2016; Attardi et al., 2016; Deriu et al., 2016). Those systems share some architectural choices: (i) use of Convolutional Sentence encoders (Kim, 2014), (ii) leveraging pre-trained word2vec embeddings (Mikolov et al., 2013) and (iii) use of distant supervision to pre-train the network (Go et al., 2009). Despite this network is simple and provides state of the art results, it does not model long-term dependencies in the tweet by construction.

3 Application Scenario

Our commercial application is a social media and micro-blogging monitoring platform, which is used to analyze brand reputation, competitors, the voice of the customer and customer experience. It is capable of managing multi-domain and multi-channel content in different languages and it is provided as a service for several clients on different market segments.

The application uses an SA algorithm to analyze the customers' opinions and feedbacks on services and products, both direct and indirect. The sentiment metric is used by the application clients to point out customer experience, expectations, and perception. The final aim is to promptly react and identify improvement opportunities and, afterward, measure the impact of the adopted initiatives.

3.1 Focused Problem Description

Industrial applications, used by demanding clients, and dealing with real data tend to prefer easily adaptable and reliable solutions. Major problems are related to multi-tenant applications with several client requirements on the sentiment analysis problem, often requiring variations on task approaches within the same platform. Moreover, high attention is put on scalability and lightweight use of computational resources, preserving accurate performance. Finally, dealing with different client domains and data potentially requires constantly training new models with limited time availability.

3.2 Data Description

The commercial social media and micro-blogging monitoring platform continuously acquires data coming from several sources; among these, we selected Twitter data as the main source for our purposes.

First, the public Twitter stream was collected for several months without specific domain restriction to build the dataset used for the word embedding training. The total amount of tweets used accounts for 100 million Italian tweets and 50 million English tweets.

Then, a dataset has been constructed from a specific market sector in Italian. The data collection was performed on the public Twitter stream with specific word restriction performed in order to filter the tweets of interest on the automotive domain. Afterward, the commercial platform applies different techniques in order to exclude from these collections the tweets that are not relevant for the specific insight analysis.

The messages were then used to construct the dataset for our experiments. A manual annotation phase has been performed together with the demanding client in order to best suit the insight objective requirement. Even though structured guidelines were agreed upon before creating the dataset and continuously checked against, this approach tended to generate dataset characteristics: in particular, unbalanced distribution of the examples over the different classes has been measured. It makes necessary a flexible model capable of handling such phenomena without the need of costly tuning phases and/or network re-engineering.

4 Our Neural Network Approach

The task of SA in Twitter aims at classifying a tweet $t \in T$ into one of the three sentiment classes $c \in C$, where $C = \{positive, neutral, negative\}$. This can be achieved by learning function $f : T \to C$ through a neural network. The architecture here proposed is based on (Severyn and Moschitti, 2015) and it is structured in three steps: (i) a tweet is encoded into an embedding matrix, (ii) an encoder maps the tweet matrix into a fixed size vector and (iii) a single output layer (a logistic regression layer) classifies this vector over the three classes.

In contrast to Severyn and Moschitti (2015), we adopted a Recurrent Pooling layer that allows the network to learn longer dependencies in the input sentence (i.e. sentiment shifts). This architectural change makes the network less sensible to learn biases from the dataset and therefore generalize better on poorly represented classes.

Embedding: a tweet t is represented as a sequence of words $\{w_1, .., w_j, .., w_N\}$. Tweets are encoded into a sentence matrix $\mathbf{t} \in \mathbb{R}^{d \times |\mathbf{t}|}$, obtained by concatenating its word vectors $\mathbf{w}_j$, where d is the size of the word embeddings.

Sentence Encoder: it is a function that maps the sentence matrix $\mathbf{t}$ into a fixed size vector x representing the whole sentence. Severyn and Moschitti (2015) used a convolutional layer followed by a global max-pooling layer to encode tweets. The convolution operation applies a sliding window operation (with window of size m) over the input sentence matrix. More specifically, it applies a non-linear transformation generating an output matrix $\tilde{\mathbf{x}} \in \mathbb{R}^{N \times d_{conv}}$ where d_{conv} is the number of convolutional filters and N is the length of the sentence. The max-pooling operation applies an element-wise max operation to the transformed sentence matrix $\tilde{\mathbf{x}}$, resulting in a fixed size vector representing the whole sentence.

In this work, we propose to substitute the max-pooling operation with a Bidirectional Gated Recurrent Unit (BiGRU) (Chung et al., 2014; Schuster and Paliwal, 1997). The GRU is a Gated Recurrent Neural Network capturing long term dependencies over the input. A GRU processes the input in a direction (e.g., from left to right), updating a hidden state that keeps the memory of what the network has processed so far. In this way, a whole sentence can be represented by taking the hidden state at the last step. In order to capture dependencies in both directions, i.e., a stronger representation of the sentence, we apply a BiGRU, which performs a GRU operation in both the directions $BiGRU(\tilde{\mathbf{x}}) = [\overrightarrow{GRU}(\tilde{\mathbf{x}}); \overleftarrow{GRU}(\tilde{\mathbf{x}})]$.

Classification: the final module of the network is the output layer (a logistic regression) that performs a linear transformation over the sentence vector by mapping it in a d_{class} dimensional vector followed by a softmax activation, where d_{class} is the number of classes.

5 Experiments

5.1 Setup

Similarly to Severyn and Moschitti (2015), for the CNN, we use a convolutional operation of size 5 and $d_{conv} = 128$ with rectified linear unit activation, ReLU. For the BiGRU, we use 150 hidden units for both $\overleftarrow{GRU}$ and $\overrightarrow{GRU}$ obtaining a fixed size vector of size 300.

Word embeddings: for all the proposed models, we pre-initialize the word embedding matrices with the standard skip-gram embedding of dimensionality 50 trained on tweets retrieved from the Twitter Stream.

Training: the network is trained using SGD with shuffled mini-batches using the Adam update rule (Kingma and Ba, 2014) and an early stopping (Prechelt, 1998) strategy with patience $p = 10$. Early stopping allows avoiding overfitting and to improve the generalization capabilities of the network. Then, we opted for adding dropout (Srivastava et al., 2014) with rates of 0.2 to improve generalization and avoid co-adaptation of features (Srivastava et al., 2014).

Datasets: we trained and evaluated our architecture on two datasets: the English dataset of Semeval 2015 (Rosenthal et al., 2015) described by

Table 1: Splits of the Semeval dataset

	pos.	neu.	neg.	total
train	5,895	471	3,131	9,497
valid	648	57	430	1,135
test 2013	2,734	160	1,541	4,435
test 2015	1,899	190	1,008	3,097

Table 2: Splits of the Italian dataset

	pos	neu	neg	total
train	4,234	6,434	2,170	12,838
valid	386	580	461	1,427
test	185	232	83	500

Table 1 in terms of the size of the data splits and positive, negative and neutral instances. We used the validation set for parameter tuning and to apply early stopping whereas the systems are evaluated on the two test sets of 2013 and 2015, respectively.

The Italian dataset was built in-house for the automotive domain: we collected from the Twitter stream as explained in Section 3.2 and divided it into three different splits for training, validation and testing, respectively. Table 2 shows the size of the splits. Due to the nature of the domain, many tweets in the dataset are neutral or objective, this makes the label distribution much different from the usual benchmarks. For example, the neutral class is the least represented in the English dataset (see Table 1) and the most represented in the Italian data. The imbalance can potentially bias neural networks towards the most represented class. One of the features our approach is to diminish such effect.

Evaluation metrics: we used the following evaluation metrics, Macro-F1 (the average of the F1 over the three sentiment categories). Additionally, we report the $F1_{p,n}$, which is the average F1 of the positive and negative class. This metric is the official evaluation score of the SemEval competition.

5.2 Results on English Data

Table 3 presents the results on the English dataset of SemEval 2015. The first row shows the outcome reported by Severyn and Moschitti (2015) (S&M). CNN+Max is a reimplementation of the above system with Convolution and Max-Pooling but trained just on the official training data without distant supervision. This system is used as a strong baseline in all our experiments. Lastly, we report the results obtained with the BiGRU pooling strategy described in Section 4. The proposed architecture presents a slight improvement over the strong baseline ($\sim$ 1 point of both $F1$ and $F1_{p,n}$ score on the test).

Table 3: English results on the SemEval dataset

	2013 test		2015 test	
	$F1$	$F1_{p,n}$	$F1$	$F1_{p,n}$
S&M (2015)	—	72.79	—	64.59
CNN+Max	**72.04**	67.71	67.14	62.63
CNN+BiGRU	71.67	**68.10**	**68.03**	**63.82**

Table 4: Italian results on the automotive dataset

	Valid		Test	
	$F1$	$F1_{p,n}$	$F1$	$F1_{p,n}$
CNN+Max	**65.34**	62.35	**69.35**	62.88
CNN+BiGRU	64.85	**67.71**	68.32	**67.55**

5.3 Results on Italian Data

Table 4 presents the result on the Italian dataset. Despite that on this dataset the proposed CNN+BiGRU model obtains lower F1 scores, it shows improved performance in terms of $F1_{p,n}$ (5 points on both validation and test sets). This suggests that the proposed model tends to generalize better on the less represented classes, which, in the case of the Italian training dataset, are the positive and negative classes (as pointed out in Table 2).

5.4 Discussion of the Results

We analyzed the classification scores of some words to show that our approach is less affected by the skewed distribution of the dataset. The sentiment trends, as captured by the neural network in terms of scores, are shown in Table 5.4). For example, the word *Mexico* classified by CNN+Max produces the scores, 0.06, 0.35, 0.57, while CNN+BiGRU outcome, 0.18, 0.52, 0.30, for the negative, neutral and positive classes, respectively. This shows that CNN+BiGRU is less biased by the data distribution of the sampled word in the dataset, which is, 0, 1, 5, i.e., *Mexico* appears 5 times more in positive than in neutral messages and never in negative messages.

This skewed distribution biased more CNN+Max as the positive class gets 0.57 while the negative one only 0.06. CNN+BiGRU is able, instead, to recover the correct neutral class. We believe that CNN+Max is more influenced by

	Cnn+Max	Cnn+BiGRU
Mexico	(.06, .35, .57)	(.18, .51, .30)
Italy	(.06, .54, .38)	(.18, .54, .26)
nice	(.007, .009, .98)	(.05, .07, .87)

Table 5: Word classification scores obtained with the two neural architectures on English language. The scores refer to the negative, neutral and positive classes, respectively.

the distribution bias as the max pooling operation seems to capture very local phenomena. In contrast, BiGRU exploits the entire word sequence and thus can better capture larger informative context.

A similar analysis in Italian shows the same trends. For example, the word *panda* is classified as, 0.05, 0.28, 0.66, by CNN+Max and 0.07, 0.56, 0.35 by CNN+BiGRU, for negative, neutral and positive classes, respectively. Again, the distribution in the Italian training set of this word is very skewed towards the positive class: it confirms that CNN+Max is more influenced by the distribution bias, while our architecture can better deal with it.

6 Conclusions

In this paper, we have studied state-of-the-art neural networks for the Sentiment Analysis of Twitter text associated with a real application scenario. We modified the network architecture by applying a recurrent pooling layer enabling the learning of longer dependencies between words in tweets. The recurrent pooling layer makes the network more robust to unbalanced data distribution. We have tested our models on the academic benchmark and most importantly on our data derived from a real-world commercial application. The results show that our approach works well for both English and Italian languages. Finally, we observed that our network suffers less from the dataset distribution bias.

References

Giuseppe Attardi, Daniele Sartiano, Chiara Alzetta, and Federica Semplici. 2016. Convolutional neural networks for sentiment analysis on italian tweets. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.* CEUR-WS.org.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016).*

Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2016. Context-aware convolutional neural networks for twitter sentiment analysis in italian. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.* CEUR-WS.org.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555.*

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *SemEval@ NAACL-HLT*, pages 1124–1128.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50(1):723–762, May.

Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu-Ren Huang, and Guodong Zhou. 2010. Sentiment classification and polarity shifting. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 635–643. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *SemEval@ NAACL-HLT*, pages 1–18.

Lutz Prechelt. 1998. Early stopping-but when? In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, pages 55–69, London, UK, UK. Springer-Verlag.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado, June. Association for Computational Linguistics.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962. ACM.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Emo2Val: Inferring Valence Scores from fine-grained Emotion Values

Alessandro Bondielli, Lucia C. Passaro and **Alessandro Lenci**
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica
University of Pisa (Italy)
alessandro.bondielli@gmail.com
lucia.passaro@for.unipi.it
alessandro.lenci@unipi.it

Abstract

English. This paper studies the relationship between the *valence*, one of the psycholinguistic variables in the Italian version of ANEW (Montefinese et al., 2014), and emotive scores calculated by exploiting distributional methods (Passaro et al., 2015). We show two methods to infer valence from fine grained emotions and discuss their evaluation.

Italiano. *Questo lavoro studia la relazione tra la valenza, una delle variabili psicolinguistiche presenti nella versione italiana di ANEW (Montefinese et al., 2014) e degli score emotivi calcolati distribuzionalmente (Passaro et al., 2015). Mostriamo due metodi per inferire la valenza a partire da tali valori e ne discutiamo la valutazione.*

1 Introduction

Recent years have seen a surge in studies concerning emotional ratings, both in psycholinguistics and in affective computing. Traditionally, the three main behavioral dimensions to measure the emotional value of a word are *valence*, *arousal* and *dominance*. Warriner et al. (2013) define valence as the "pleasantness of the stimulus", usually ranging from 1 (very unpleasant) to 9 (very pleasant). The word *dead* has a low valence rating, whereas *holiday* has a higher one. Arousal is the intensity of the feeling evoked on a scale from "stimulated" to "unaroused". A highly stimulating word is *passion*. On the contrary, *sleep* is not arousing. Finally, dominance is identified with the degree to which the stimulus makes the reader feel "in control" (Louwerse and Recchia, 2014). *Victory* is a word with high dominance.

In the domain of Affective Computing, the goal moves from the identification of such variables to the annotation of the texts with the emotions they express and - for Sentiment Analysis - with their degree of positivity and/or negativity.

The aim of this work is to study the relationship between the most important psycholinguistic variables and emotive scores calculated by exploiting distributional methods. In particular, we will focus on valence ratings, assuming that, within these three dimensions, valence is the most highly related with a positive, negative or neutral emotional content. In fact, it can be defined as the "the polarity of emotional activation" (Lang et al., 1999).

A possible approach to infer the valence of the words from co-occurrence statistics is the one adopted by Louwerse and Recchia (2014), who followed a bootstrapping method to extend the ANEW lexicon (Bradley and Lang, 1999). Another approach would be to exploit a resource such as SenticNet (Cambria et al., 2016) to infer valence based on values of polarity for words or conceptual primitives. An alternative strategy is to infer the valence from an emotive lexicon such as ItEM (Passaro et al., 2015; Passaro and Lenci, 2016), a distributional lexicon for Italian, in which words are associated with an emotive score for 8 different emotions. In our opinion, this solution has several advantages: first of all, ItEM has been proven to be quite robust, and guarantees high coverage over Italian words; secondly, it is not only a static resource, but it can be easily expanded with new words, allowing for a quick adaptation to different contexts. Finally, associating words with fine-grained emotional values allows for a wide range of analyses, such as for instance hate and violence detection in texts.

Experimental results showed, in an indirect way, that distributional emotive ratings can be very useful in the implementation of systems for polarity classification (Passaro and Lenci, 2016;

Bondielli, 2016). However, what is the real relation between emotive scores and valence? Our hypothesis is that emotions can be seen as a representation of valence on a more granular scale. The Plutchik's emotion taxonomy (Plutchik, 1994; Plutchik, 2001) is partitioned into positive or negative emotions. However, borderline emotions such as SURPRISE are harder to be included into a positive or negative class, and therefore to be attributed with a direct valence rating. Words like *party* and *gun* will have widely differing valence ratings, but both strongly elicit the emotion of SURPRISE. Hence it is interesting to ask the following question: given ItEM, are we able to predict the valence (i.e., positivity and/or negativity) of its words? In order to address this latter point, we performed a simple regression model to predict the valence ratings of words in ANEW (Montefinese et al., 2014) given the respective emotive values in ItEM (Passaro et al., 2015; Passaro and Lenci, 2016).

This paper is organized as follow: in Section 2 we describe the resources used for the creation of the model. Section 3 shows our method and the results obtained. Finally, in Section 4 we evaluate the results and discuss our findings.

2 Resources

The main resources we used for our experiments are the Italian version of the Affective Norms for English Words (Montefinese et al., 2014) and the Italian EMotive lexicon (Passaro et al., 2015).

2.1 Italian ANEW

ANEW (Affective Norms for English Words) (Bradley and Lang, 1999) is a database created from a rating of 1034 English words with values for *valence*, *arousal* and *dominance*. Montefinese et al. (2014) provided an Italian version of ANEW, developed by translating the English ANEW words, and by adding the words taken from the Italian semantic norms (Montefinese et al., 2012), for a total of 1121 words. Ratings have been obtained via an experiment where participants had to rate words for the target variables. The reported ratings are the average of the ratings for all participants.

2.2 ItEM

ItEM (Passaro et al., 2015; Passaro and Lenci, 2016) is an emotive lexicon for Italian, in which each target term is associated with a score quantifying its association with each emotion in the Plutchik's taxonomy (Plutchik, 1994): JOY, SADNESS, ANGER, FEAR, TRUST, DISGUST, SURPRISE and ANTICIPATION. The resource has been created as follows: in a first phase, feature elicitation was used to create a small set of seed lemmas highly associated to one or more of the emotions in the taxonomy. Then, these lemmas have been distributionally expanded with the most frequent words in two Italian corpora (Baroni et al., 2004; Baroni et al., 2009). Finally, the emotive scores for each word were calculated by measuring the cosine similarity between the lemma and eight emotive centroids built from the collected seeds.

3 From fine-grained Emotion Values to Polarity

We used 2 main regression models to predict the valence from the distributional emotive scores. The first experiment, described in section 3.1 shows a polynomial regression model, and the second one (section 3.2) shows a logistic model in which the valence scores in ANEW have been discretized into two classes representing the positiveness and negativeness of the word.

A simple preprocessing phase has been applied to align the two resources. ANEW has 1121 words, but 65 of them have multiple POS (e.g. *aereo* (plane) can be both a noun and an adjective). We duplicated each word, extending the dataset to 1189 elements, and extracted distinct emotive scores for each <lemma,PoS> pair. In addition, we replaced word forms like "scorie" (waste), with their most frequent word type (scoria) in ItaWaC (Baroni et al., 2004) and La Repubblica (Baroni et al., 2004). Eventually, 57 ANEW words were left out of the analysis because they were not in ItEM. Overall, the resulting size of the aligned dataset is 1129 elements. Finally, to cope with the different distribution of data among the various emotions in ItEM, we normalized the scores with their z-score.

3.1 Polynomial regression

Due to the bimodal distribution of the data in ANEW, we decided to use a polynomial regression model to predict the valence of the words in ANEW by exploiting their emotive normalized scores in ItEM. Preliminary tests had in fact shown that a simple multiple linear regression model was not able to properly fit the data. The histogram

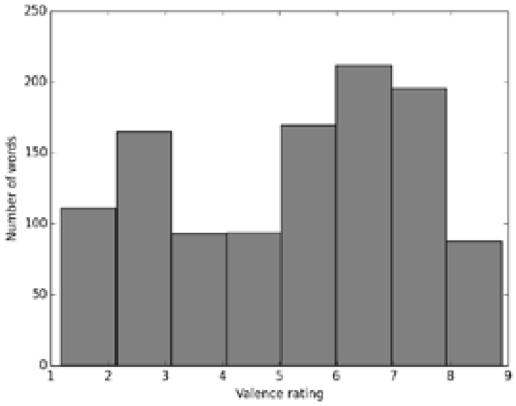


Figure 1: Valence ratings distribution

in Figure 1 shows such data distribution, in which most of the ANEW words have a valence score in the ranges 2-3 and 6-8, with a slight bias towards higher values.

To define the most performing degree (Deg) of the polynomial function, we performed 10-fold cross validation for degrees in the range $\{1...5\}$. The results, presented in Table 1, clearly show overfitting for degrees equal or higher than 3. This is due to the fact that, given the number of parameters (#P), the estimated minimum number of observations (Min. Obs.), computed as $\#P \times 15$, must be at most around the total number of observations. This is true only for polynomial of degree 1 and 2. This finding is in line with Schmidt (1971) and Harrell (2001) who demonstrated that to guarantee the reliability of the prediction, each parameter in the regression model should have a minimum number of observations between 10 and 20.

Deg	#P	Min. Obs.	R^2	MSE
1	9	~ 135	0.46	2.24
2	45	~ 675	0.53	1.82
3	165	~ 2475	0.31	1.50
4	495	~ 7425	-81.29	0.96
5	1287	~ 19305	-11 B	0.00

Table 1: Experiments performed to define the most performing Deg for the polynomial

Given this result, we performed a polynomial interpolation over our parameters with a polynomial of degree 2. Then, we applied a simple multiple linear regression over the new data for predicting the valence. Figure 2 shows the result of the regression fitting. For this model, we obtained a R-Squared (R^2) of 0.58, a mean absolute error (MeanAE) of 1.08, a mean squared error (MSE) of 1.81, and a Median absolute error (MedianAE) of 0.95.

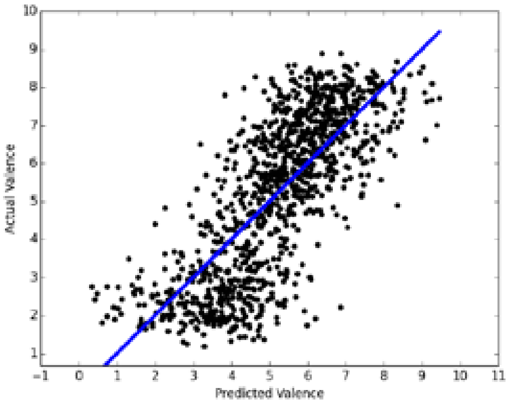


Figure 2: Fitting of predictions

For this experiment, we also provide two additional evaluations (the corresponding results are shown in Table 2):

A) the results of prediction by means of a 10-fold cross validation;

B) the results of prediction by means of split of the data between training (66%) and test (33%).

Method	R^2	MeanAE	MSE	MedianAE
A	0.53	1.13	1.99	0.98
B	0.54	1.13	2.00	0.93

Table 2: Results of the evaluations

We would like to notice that our prediction performs better for words with a very high arousal. In fact, emotionally arousing words were more likely to be produced as an emotive *prototypical* word in the elicitation phase of ItEM. As a consequence, since ItEM's emotive centroids have been constructed using the vectors of these words (namely the seeds), also their nearest neighbors (i.e., the most emotive words) are assumed to have a high level of arousal. Moreover, the distribution of the data in Figure 3, clearly shows how, in ANEW, high arousal corresponds to very high (or very low) valence ratings, suggesting that highly arousing words tend to be very positive or very negative (i.e. polarized). Building on this evidence, we performed an additional experiment in which we used the portion of the data (573 words) with an arousal

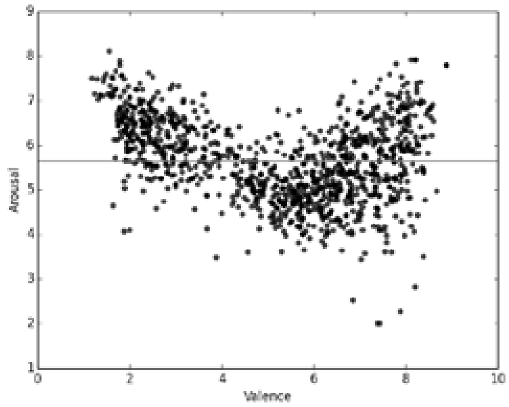


Figure 3: Valence-Arousal distribution

rating higher than its median (5.64) for prediction. In such model, in fact, R^2 is attested to ~ 0.64.

Given the distribution of the data showed in Figure 2, it is clear that a polynomial regression might not be a perfect fit for valence ratings. Nevertheless, it is very important to focus on MeanAE and MSE values. These errors are relatively low with respect to the scale of the human-rated valences.

This means that, on average, the difference between human-rated valence and predicted valence is between 1 and 2. To prove this point, we also compared the obtained scores with the original human annotations, by exploiting the standard deviation for each valence rating. We found that $73,5\%$ of our predictions fall into the correct range around the average valence. If we consider a word having (in ANEW) a valence score of around 8 (e.g. *pace* (peace)) the system will predict a score between 6 and 9, leaving the word around the same (positive) area of the distribution. The same (and opposite) goes for low-valenced words, such as *drogato* (drug addicted) and *feccia* (scum). Problems arise in the case of the words with a medium valence. Examples can be *corridoio* (corridor) and *insipido* (bland). In this case, the word will have the same chance to be attributed with a high valence score (5-6) or with a low one (3-4). Supposing to discretize valence ratings in two classes, a positive and a negative one, with a cut on the median, predictions will fall in the right class for most of the high (or low) valenced words, and (possibly) in the wrong one for the words of medium valence. In fact, by constructing a shallow mapping of the valence into positive (with $valence >= 5.5$) and negative class, we found a correlation of 0.73 between predicted and actual data.

3.2 Logistic regression

Building on the last experiment, and supposing a discretization of the valence into the positive and negative class, we also used a logistic regression model to predict this *binary valence*. The results of this experiment are very promising. We performed 10-fold cross validation to evaluate the effectiveness of the logistic regression over the transformed valence ratings, and obtained an average mean accuracy of 0.80. Detailed results for this evaluation are shown in Table 3.

	Precision	**Recall**	**F1**
MicroAVG	0.806	0.803	0.802
MacroAVG	0.803	0.803	0.803

Table 3: Logistic regression (Cross Validation)

4 Results and discussion

The results provided in previous experiments showed both pros and cons of this approach.

The main advantage of exploiting distributional emotive scores to predict the word's valence is that such scores can be easily obtained in an unsupervised way by means of co-occurrence statistics.

Moreover, predicted data showed a rather good accuracy with respect to the actual distribution, especially considering the logistic regression experiment. In fact, our models reach peak performances by focusing the analysis on the sign of the valence with logistic regression instead of working with continuous values.

On the other hand, the main drawback of our approach derives from the dimension of the ANEW dataset, and in particular from the lack of examples around the medium valence score ratings. It is clear that the ratings distribution in this resource prevented us from obtaining reliable results for continuous values. This might also provide an explanation for the errors concerning the logistic regression experiment. We are confident that having access to a new resource covering the full spectrum of the valence more evenly would have a positive impact on our model.

5 Conclusions and ongoing research

In this work we studied the relationship between *valence* and distributional emotive scores. We modeled our data with regression in order to predict both a continuous score for valence and its corresponding binarized version (i.e., polarity).

Despite the difficulties of modeling an accurate representation of a continuous valence rating from a small and unbalanced dataset like the Italian ANEW, we can identify a clear relationship between distributional emotional scores and a discrete valence obtained by categorizing the ratings into a positive and a negative class.

In the near future, we plan to improve our regression models, with the aim of reducing the impact of the distribution of the data in ANEW, possibly implementing new strategies able to cope with non linear data. ANEW is a highly renown psycholinguistic dataset, but we plan to extend the present work to predict sentiment polarity scores taken from SentiWordNet (Esuli and Sebastiani, 2006a; Esuli and Sebastiani, 2006b), thereby exploiting the larger coverage of this resource.

Moreover, we plan to follow the approach employed in ItEM to create a polarity lexicon for Italian, using ANEW words as seed to build positive and negative polarity centroids. This would also be beneficial for evaluating performances on a emotion-based approach and a polarity-based one.

Finally, we aim at testing the effectiveness of our system for Sentiment Polarity Classification.

References

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the la repubblica corpus: A large, annotated, tei (xml)-compliant corpus of newspaper italian. *issues*, 2:5–163.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Alessandro Bondielli. 2016. Da facebook a twitter: Creazione e utilizzo di una risorsa lessicale emotiva per la sentiment analysis di tweet. Master's thesis, University of Pisa, Italy.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn W Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING*, pages 2666–2677.

A. Esuli and F. Sebastiani. 2006a. Determining term subjectivity and term orientation for opinion mining. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL06)*, Trento (Italy). Association for Computational Linguistics.

A. Esuli and F. Sebastiani. 2006b. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 417–422, Genoa (Italy). European Language Resource Association (ELRA).

F.E. Harrell. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics. Springer.

Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. 1999. International affective picture system (iaps): Technical manual and affective ratings. *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida*, 2.

MM Louwerse and G Recchia. 2014. Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, 68(12):1–15.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2012. Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods*, pages 1–22, oct.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46(3):887–903.

Lucia C. Passaro and Alessandro Lenci. 2016. Evaluating context selection strategies to build emotive vector space models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoro (Slovenia).

Lucia C Passaro, Laura Pollacci, and Alessandro Lenci. 2015. Item: A vector space model to bootstrap an italian emotive lexicon. *CLiC it*, 60(15):215.

Robert Plutchik. 1994. *The psychology and biology of emotion.* HarperCollins College Publishers.

R. Plutchik. 2001. The nature of emotions. *American Scientist*, 89:344–350.

Frank L Schmidt. 1971. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 31(3):699–714.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus

Peter Bourgonje, Yulia Grishina, Manfred Stede
Applied Computational Linguistics
University of Potsdam / Germany
{bourgonje,grishina,stede}@uni-potsdam.de

Abstract

English. We report on experiments to validate and extend two language-specific connective databases (German and Italian) using a word-aligned corpus. This is a first step toward constructing a bilingual lexicon on connectives that are connected via their discourse senses.

Italiano. *Presentiamo una serie di esperimenti per validare ed estendere due database dei connettivi, che sonospecifici per la lingua italiana e per quella tedesca. Abbiamo utilizzato un corpus parallelo allineato a livello della parola. Si tratta di un primo passo verso la costruzione di un lessico bilingue dei connettivi che sono collegati attraverso i loro sensi del discorso.*

1 Introduction

An important part of discourse processing deals with uncovering coherence relations that hold between individual, "elementary" units of a text. The lexical items that can signal such a relation are referred to as *discourse connectives*, and examples of these relations, also called the connectives' senses, are *contrast* (e.g., 'but'), *elaboration* (e.g., 'in particular'), or *cause* (e.g., 'therefore'). Notice, however, that relations need not always be signalled in text, if the context or world knowledge is sufficient for the reader to infer it, as (1)-(4) demonstrate:

(1) We should hurry, because it's late.

(2) We should hurry. It's late.

(3) The red pen costs \$2, while the blue one is \$2.50.

(4) The red pen costs \$2; the blue one is \$2.50.

On the other hand, example (6) is a perfectly grammatical sentence but the meaning is different from (5), so for this case of a Concession relation, the connective is in fact indispensable.

(5) Although it is late, we don't need to hurry.

(6) It is late; we don't need to hurry.

Recognizing these relations, which can hold within a sentence, between two sentences, or between larger spans of text, is a central task for uncovering the structure of a text, as it has been studied in theories like Rhetorical Structure Theory (Mann and Thompson, 1988) or Segmented Discourse Representation Theory (Asher and Lascarides, 2003). While the usage of connectives can sometimes be optional, the *set* of connectives that a language offers is generally taken as important (if not exhaustive) evidence for the set of coherence relations that should be assumed.

1.1 Background: Connectives

From a syntactic viewpoint, 'connective' is not a homogeneous class, as it contains conjunctions, different kinds of adverbials, as well as certain prepositions. Our underlying definition of discourse connectives is based on (Pasch et al., 2003, p. 331):

(7) **Def.:** A *discourse connective* is a lexical item x that exhibits each of the following five properties:
(M1) x cannot be inflected.
(M2) x does not assign case features to its syntactic environment.
(M3) The meaning of x is a two-place relation.
(M4) The arguments of the relation (the meaning of x) are propositional structures.
(M5) The expressions of the arguments of the relation can be sentential structures.

Following (Stede, 2002), we drop M2 because our lexicon deliberately includes several prepositions that can be used as connectives (in the sense of M1, M3-M5), e.g., *trotz* (‘despite’) or *wegen* (‘due to’).

1.2 Motivation and contribution

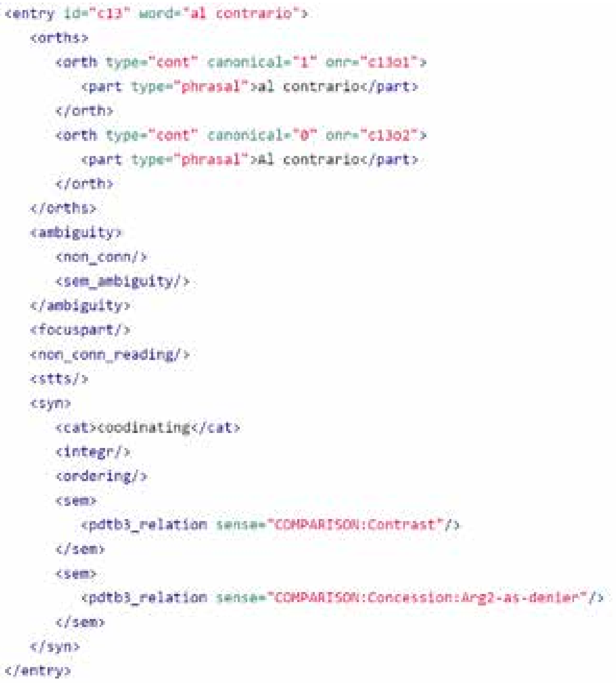
Connectives can pose interesting challenges to translation and for language learners, as the differences in meaning between similar connectives can be quite subtle. For these reasons, we are interested here specifically in a *bilingual* Italian–German lexical resource, to be built on top of two existing single-language lexicons. As a case study, we focus on the subgroup of contrastive/concessive connectives, which we determined to comprise (in the existing lexicons) 31 German connectives and 12 Italian connectives; see Tables 3.2.2 and 3.2.2.

The main contributions of this paper are (1) suggestions for improving the existing language-specific resources used in this study through the technique of *cross-lingual projection* in a parallel corpus, which reveals correspondences between connectives and can point to gaps in either of the resources; and (2) an overview of the distribution of connectives and their senses, to be used in a bilingual database. Section 2 explains the two monolingual lexicons we work with, and Section 3 describes the corpus. Section 4 reviews related work in this area. Section 5 elaborates the idea of bilingual connective databases, and Section 6 summarises our findings.

2 Lexicons: DiMLex and LICo

We extracted the German contrastive connectives from DiMLex (Scheffler and Stede, 2016), a connective lexicon with several different fields describing orthographical variants, syntactic type, discourse sense, and usage examples. It contains 275 entries. The sense annotations are based on the Penn Discourse Treebank (PDTB) senses (Miltsakaki et al., 2008) in its latest version 3. The lexicon is publicly available[1] and aims to exhaustively describe the set of connectives for German, thus providing a basis for our case study.

The set of Italian contrastive connectives comes from LICo (Feltracco et al., 2016), a similar lexicon for Italian containing 170 entries.[2] LICo was inspired by DiMLex and contains annotations on the same attributes and uses essentially the same structure (i.e., the same PDTB senses, orthographic variants, usage examples, etc.). An example entry of LICo is shown in Figure 1. We refer the reader to Feltracco et al. (2016) for details.



```
<entry id="c13" word="al contrario">
   <orths>
      <orth type="cont" canonical="1" onr="c13o1">
         <part type="phrasal">al contrario</part>
      </orth>
      <orth type="cont" canonical="0" onr="c13o2">
         <part type="phrasal">Al contrario</part>
      </orth>
   </orths>
   <ambiguity>
      <non_conn/>
      <sem_ambiguity/>
   </ambiguity>
   <focuspart/>
   <non_conn_reading/>
   <stts/>
   <syn>
      <cat>coodinating</cat>
      <integr/>
      <ordering/>
      <sem>
         <pdtb3_relation sense="COMPARISON:Contrast"/>
      </sem>
      <sem>
         <pdtb3_relation sense="COMPARISON:Concession:Arg2-as-denier"/>
      </sem>
   </syn>
</entry>
```

Figure 1: *al contrario* entry in LICo

3 Exploiting a parallel corpus

For the parallel German/Italian corpus we used Europarl (Koehn, 2005), as it still appears to be the biggest resource of this kind, and it is, conveniently, already sentence-aligned. From the 1,832,053 sentences in the German-Italian part of the corpus we extracted the word alignments using MGIZA++ (Gao and Vogel, 2008). In the following, we sketch our method for obtaining the correspondence information on connectives based on these word alignments, and then present the results.

3.1 Method: Iterative lookup

We approach the problem from two sides: First we look up every German connective (31 in total) to get Italian alignments. 30 of them appeared in our Europarl corpus (with *dementgegen* missing). Then we look up every Italian connective to get German alignments (all 12 connectives present in

[1] https://github.com/discourse-lab/dimlex

[2] https://hlt-nlp.fbk.eu/technologies/lico

the corpus). We end up with a list of target language words or phrases (or empty elements, since a source language connective can also be covert in the target language) that are candidate contrastive connectives. Note that the lookup procedure does not differ structurally between words and phrases. In both cases, single words (stand-alone or in a phrase) can correspond to zero, one or more target words. The target representation is collected in a key-value structure, where the key is the position in the sentence and the value the word. This list is then sorted by position to return the target word or phrase (which is potentially discontinuous). Because the word alignment is not guaranteed to be correct, to filter for unlikely translations we focus on only the 3 most frequent alignments for every connective. We expect to find at least a subset of the already known (contrastive) connectives (from DiMLex or LICo), potentially complemented by a set of words or phrases that can help filling gaps in either of the lexicons.

This procedure produces at least some incorrect results for the following two reasons: 1) discourse connectives often can appear in a text with a connective reading or with a non-connective reading; and 2) connectives can have multiple senses, so that a connective may not have the contrastive reading in the particular sentence. The candidates produced hence have to be evaluated manually. Resulting candidates that have a connective reading are added to the seed list, in order to repeat the step back from the target language to the source language[3].

3.2 Results

3.2.1 German–Italian

The results of the first step of the iteration using the 31 German seed connectives are displayed in Table 3.2.2, where an underscore indicates an empty string (meaning that the connective was not aligned to a particular word or phrase in the target language) and the number after the underscore represents the (normalised) frequency of the alignment.

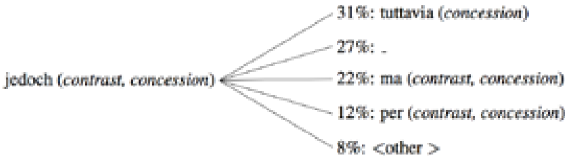

jedoch (*contrast, concession*)
31%: tuttavia (*concession*)
27%: _
22%: ma (*contrast, concession*)
12%: per (*contrast, concession*)
8%: <other >

Figure 2: Most frequent alignments of *jedoch*

For the evaluation, we asked a native speaker of Italian with expert knowledge in linguistics to validate the resulting top 3 bilingual mappings. Firstly, we identified several possible connective candidates that were aligned to German contrastive connectives, but were not present in LICo, such as *al contempo, solo che, doppo tutto*. Secondly, we observed several possible orthographic variants of the already existing Italian connectives: *contro* or *contrario* (as possible variants of *al contrario*), and *d'altro canto* (as a variant of a discontinious connective *da un canto...dall'altro*). Finally, we found that several Italian connectives only had the *concession* sense, while the corresponding German connectives also had the Contrast sense, such as *comunque*, for which we found the German alignments *aber*, *allerdings* and *doch*, for example.

As an example of a visualisation (for a single connective) the above analysis is based on, consider Figure 2, showing the most frequent alignments of *jedoch*, which always has a connective reading, thus nullifying the first problem mentioned in 3.1.

3.2.2 Italian–German

The results of the first step of the iteration using the 12 Italian seed connectives are displayed in Table 3.2.2. For 11 of the 12 contrastive connectives from LICo, the top 3 alignments yielded an existing DiMLex entry. The only connective without a DiMLex entry in the top 3 was *al contrario*, for which a possible new German connective candidate *im Gegenteil* was found through alignment.

Upon further investigation of the lower-ranked alignments (not included in Table 3.2.2), we were able to identify several other gaps in the German lexicon. Firstly, we observed that the Italian connective *invece* is frequently aligned to the German word *anstelle*, which is not in DiMLex (but *anstelle dessen* is). After examining the corresponding examples, we conclude that *anstelle* should be added to DimLex as a separate entry (similarly to the already existing *aufgrund* vs. *aufgrund dessen*). Also, we found that DiMLex lacks

[3] Ideally going back and forth until a stable and exhaustive set of candidates is found. For this study, we only did the first step, and then projected the found Italian connectives back to German.

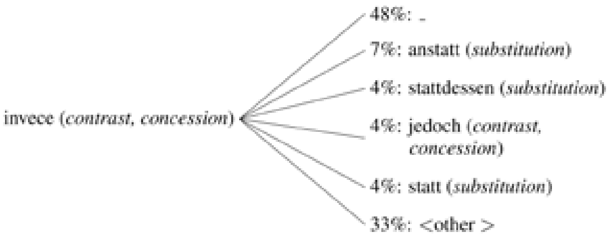

invece (contrast, concession)
48%: _
7%: anstatt (substitution)
4%: stattdessen (substitution)
4%: jedoch (contrast, concession)
4%: statt (substitution)
33%: <other >

Figure 3: Most frequent alignments of *invece*

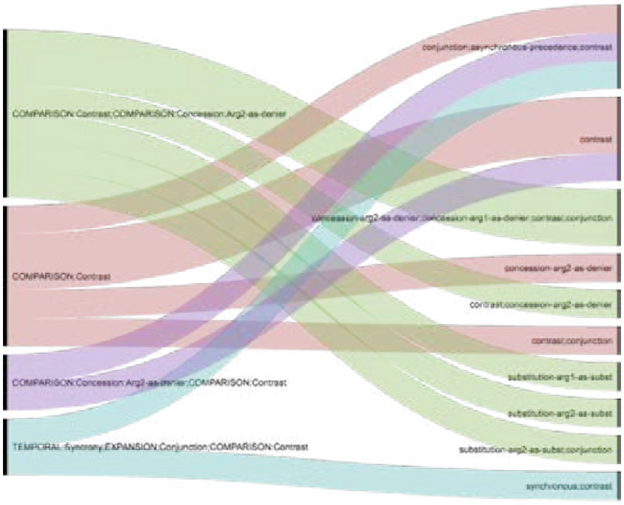


Figure 4: Mapping of connective senses from Italian to German

statt dessen as an orthographic variant of the more canonical *stattdessen*.

Finally, we identified two interesting cases that are DiMLex candidates: *umgekehrt* and *(ganz) im Gegenteil*, which we found aligned to the Italian *viceversa* and *al contrario*, respectively, but more corpus evidence is required to decide whether they can indeed serve as connective in the German language.

As an example visualisation, consider Figure 3, showing the most frequent alignments of *invece*, which always has a connective reading.

For Italian–German, we repeated the steps above with the candidates found using the German seed list (projecting the resulting Italian list back to German) to see if any additional connectives or orthographic variants would be found. We again found *im Gegenteil* through alignment of *al contrario* and a few alternative lexicalisations for DiMLex connectives[4], but no new candidates.

German connective (frequency)	Top 3 Italian alignments
aber (105413)	ma//_(0.24)//tuttavia
alldieweil (3)	finché//perché
allein (6973)	_(0.30)//solo//soltanto
allerdings (16232)	tuttavia//_(0.22)//ma
andererseits (6354)	_(0.30)//dall' altro//d' altro canto
bloß dass (117)	_(0.10)//solo che//che solo
dafür (36895)	_(0.70)//per//per aver
dafür // dass (42)	che//_(0.19)//per
dagegen (5423)	_(0.34)//contro//contrario
dahingegen (24)	_(0.17)//invece//al contrario
dementgegen (0)	
demgegenüber (121)	_(0.25)//invece//contro
doch (37423)	_(0.47)//ma//tuttavia
einerseits (4221)	da un lato//_(0.31)//da una parte
freilich (159)	_(0.30)//naturalmente//certo
gleichzeitig (13293)	_(0.35)//al contempo//allo stesso tempo
hingegen (1909)	invece//_(0.26)//tuttavia
immerhin (1360)	_(0.44)//comunque//dopo tutto
indessen (280)	invece//_(0.19)//tuttavia
jedoch (47667)	tuttavia//_(0.27)//ma
nur dass (21617)	che//solo che
sosehr (14)	malgrado tutto
unterdessen (193)	nel frattempo//_(0.21)//intanto
wiederum (2450)	_(0.55)//a sua volta//ancora una volta
wogegen (111)	mentre//_(0.19)//contro cosa
wohingegen (218)	mentre//_(0.14)//ma
während (20388)	_(0.28)//mentre//durante
währenddessen (78)	nel frattempo//_(0.17)//mentre
zugleich (3576)	_(0.41)//al contempo//allo stesso tempo
zum anderen (4299)	_(0.09)//altri//altre
zum einen (8848)	un//_(0.10)//una

Table 1: German connectives and their Italian alignments

Italian connective (frequency)	Top 3 German alignments
al contrario (3641)	im gegenteil//_(0.10)//im gegenteil
bensì (7107)	sondern//_(0.12)//sondern vielmehr
contrariamente a (661)	_(0.08)//entgegen//im gegensatz zu
da un canto (352)	einerseits//_(0.11)//andererseits
da un lato (4612)	einerseits//_(0.08)//einerseits die
da una parte (10194)	_(0.07)//und//eine
invece (18778)	_(0.48)//anstatt//stattdessen
ma (135218)	aber//sondern//_(0.15)
mentre (15773)	während//_(0.19)//und
per contro (13468)	gegen//und//_(0.06)
però (22687)	aber//jedoch//_(0.24)
viceversa (522)	umgekehrt//_(0.19)//hingegen

Table 2: Italian connectives and their German alignments

4 Related work

Parallel corpora have been successfully exploited before in order to automatically generate or induce connective lexicons in different languages. In particular, Versley (2010) projected discourse connectives across an English–German parallel corpus to train a discourse parser capable of disambiguating connective and non-connective readings. Similarly, Zhou et al. (2012) used an English–Chinese parallel corpus in order to build a Chinese connective lexicon via cross-lingual pro-

[4] Not listed here for reasons of space.

jection, and Hajlaoui and Popescu-Belis (2013) relied on parallel data to automatically retrieve Arabic counterparts for a subset of English connectives.

Since our goal was not to build a connective lexicon from scratch, but to extend the connective lists and refine the inventory of senses for the already existing lexicons, the closest approach to ours is the one adopted by Laali and Kosseim (2014), who aimed at automatically inducing a French connective lexicon via English–French parallel corpora using additional filtering rules. Similar to ours, their results have shown that using parallel translations can improve the coverage of the connective lists in both languages; however, since their lexicons used different sets of discourse relations, they were not able to extend their connective database in respect to senses, as opposed to our work.

5 Toward a bilingual connective database

Our study is meant as a step toward moving from single-language connective lexicons to a *bilingual* one that provides information about the relationships between the language-specific entries. Both monolingual lexicons are already publicly available on GitHub and in addition an interface allowing bilingual search has been made public in a related project[5]. Below we sketch additional plans for providing this information on the levels of connective tokens, and senses (coherence relations).

5.1 Connective mappings

One central purpose of a bilingual database is to assist translators (human or machine) or (human) language learners. For most connectives, there is a complicated m:n mapping between languages, which standard dictionaries do not cover, and the relevant features for making choices are not systematically known yet. A corpus-based inventory of mappings – ideally supplemented by pointers to the corpus instances and their context – can be a very useful resource for undertaking contrastive lexical investigations.

5.2 From connectives to phrases

The PDTB (Prasad et al., 2008) makes a distinction between connectives (a closed set) and "alternative lexicalizations" (AltLex), which are a non-demarcated set of phrases used to express a coherence relation. Such phrases are so far not part of DiMLex nor LICo. Obviously, they are much harder to detect: Corpus annotation (as done in PDTB) is one way, and we regard our cross-lingual projection method as another promising way. Quite often, connectives in language A have been translated to an AltLex in language B. We plan to study this more systematically by a closer inspection of the alignments and their contexts, in order to extract AltLex candidates as a supplement to the connective lexicons.

5.3 Senses and their distributions

A bilingual connective database can shed light on the distribution of senses over different languages and the degree of ambiguity that individual connectives exhibit. While we consider such conclusions premature for the current stage of the language-specific resources, we include Figure 4, which shows groups of connectives that share the same sense (or group of senses for ambiguous connectives) and their alignment to similar groups on the target side. The 12 Italian connectives (on the left), when grouped together based on their sense(s), form 4 sets, whereas for German (right side), fewer connectives (11 that were found in DiMLex among the top 3 alignments of the 12 source connectives) group into more sets (10). This suggests more ambiguity in Italian connectives, with less different senses represented by a larger set of connectives.

In addition, we observed that Italian connectives with a sense Contrast or Concession are frequently aligned to their German counterparts with a sense Substitution, such as *anstelle-invece.* Having examined the parallel examples more closely, we conclude that assigning both senses would be valid for both German and Italian, although they are placed distantly in the PDTB hierarchy of senses. These findings are confirmed by Feltracco et al. (2016), who acknowledge that the distinction between the two senses was one of the main cases of the inter-annotator disagreement. We conclude that both lexicons could benefit from adding additional senses gained via comparing parallel translations.

6 Summary

We present, to the best of our knowledge, the first Italian–German investigation of discourse connective lexicons. For the subclass of Contrast (in

[5] http://connective-lex.info/

a wide sense), we were able to identify several missing entries in both lexicons, and provided a start on identifying AltLex items for the two languages (future work). Once the information is organized in a complete bilingual database, it can assist translation and conclusions can be drawn regarding connective distribution, sense distribution and ambiguity in the different languages.

As prominent steps for future work, we note the disambiguation of connective- and non-connective readings, the implementation of more sophisticated filtering strategies to retrieve more reliable connective candidates and repeating this study for different languages pairs.

Acknowledgments

We are grateful to the Deutsche Forschungsgemeinschaft (DFG) for funding this work in the project 'Anaphoricity in Connectives'. We would like to thank the anonymous reviewers for comments on an earlier version of this manuscript. We also thank Flavia Adani for her help with translation and interpretation of the Italian results, and Tatjana Scheffler for the recent work on DiMLex.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Anna Feltracco, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. 2016. Lico: A lexicon of italian connectives. In *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it)*, Napoli, Italy.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *University of the Aegean-14th International Conference on Intelligent Text Processing and Computational Linguistics*. Springer.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *COLING*, pages 610–619.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi, 2008. *Sense annotation in the Penn Discourse Treebank*, pages 275–286. Springer Berlin Heidelberg, Berlin, Heidelberg.

Renate Pasch, Ursula Brauße, Eva Breindl, and UlrichH̃errmann Waßner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Tatjana Scheffler and Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In Nicoletta Calzolari et al., editor, *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

Manfred Stede. 2002. Dimlex: A lexical approach to discourse markers. In *Exploring the Lexicon - Theory and Computation*. Edizioni dell Orso, Alessandria.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*. Northern European Association for Language Technology (NEALT).

Lanjun Zhou, Wei Gao, Bin Li, Zhong Wei, and Kam-Fai Wong. 2012. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. In *Proceedings of COLING*.

Developing a large scale FrameNet for Italian: the IFrameNet experience

Roberto Basili° Silvia Brambilla§ Danilo Croce° Fabio Tamburini§

° Dept. of Enterprise Engineering
University of Rome Tor Vergata
{basili,croce}@info.uniroma2.it

§ Dept. of Classic Philology and Italian Studies
University of Bologna
fabio.tamburini@unibo.it,
silvia.brambilla@studio.unibo.it

Abstract

English. This paper presents work in progress for the development of IFrameNet, a large-scale, computationally oriented, lexical resource based on Fillmore's frame semantics for Italian. For the development of IFrameNet linguistic analysis, corpus-processing and machine learning techniques are combined in order to support the semi-automatic development and annotation of the resource.

Italiano. *Questo articolo presenta un work in progress per lo sviluppo di IFrameNet, una risorsa lessicale ad ampia copertura, computazionalmente orientata, basata sulle teorie di Semantica dei Frame proposte da Fillmore. Per lo sviluppo di IFrameNet sono combinate analisi linguistica,* corpus-processing *e tecniche di* machine learning *al fine di semi-automatizzare lo sviluppo della risorsa e il processo di annotazione.*

1 Introduction

Firstly developed at the University of Berkeley (California) in 1997, FrameNet adopts theories from Frame Semantics (Fillmore 1976, 1982, 1985) to NLP and explains words' meanings according to the semantic frames they evoke. It illustrates semantic *frames* (i.e. schematizations of prototypical events, relations or entities in the reality), through the involved participants (called *frame elements*, FEs) and the evoking words (or, better, the *lexical units*, LUs). Moreover, FrameNet aims to give a valence representation of the lexical units and underline the relations between frames and between frame elements (Baker et al. 1998).

The initial American project has since been extended to other languages: French, Chinese, Brazilian Portuguese, German, Spanish, Japanese, Swedish and Korean.

All these projects are based on the idea that most of the Frames are the same among languages and that, thanks to this, it is possible to adopt Berkeley's Frames and FEs and their relations, with few changes, once all the language-specific information has been cut away (Tonelli *et al.* 2009, Tonelli 2010).

With regard to Italian, over the past ten years several research projects have been carried out at different universities and Research Centres. In particular, the ILC-CNR in Pisa (e.g. Lenci et al. 2008; Johnson and Lenci 2011), FBK in Trento (e.g. Tonelli *et al.* 2009, Tonelli 2010) and the University of Rome, Tor Vergata (e.g. Pennacchiotti *et al.* 2008, Basili *et al.* 2009) proposed automatic or semiautomatic methods to develop an Italian FrameNet. However, as of today, a resource even remotely equivalent to Berkeley's FrameNet (BFN) is still missing.

As a lexical resource of this kind is useful in many computational applications (such as Human-Robot interaction), a new effort is currently being jointly made at the universities of Bologna and Roma, Tor Vergata. The IFrameNet project aims to develop a large-coverage FrameNet-like resource for Italian, relying on robust and scalable methods, in which the automatic corpus processing is consistently integrated with manual lexical analysis. It builds upon the achievements of previous projects that automatically harvested FrameNet LUs exploiting both distributional and WordNet based models (Pennacchiotti *et al.* 2008). Since the LUs induction is a noisy process, the data thus obtained need to be manually refined and validated.

The aim is also to provide Sample Sentences for LUs with the highest corpus frequency. On the one side, they will be derived from already existing resources such as the HuRIC corpus (Bastianelli 2014) or the EvalIta2011 FLaIT task data: FBK set (Tonelli, Pianta 2008) and ILC set (Lenci *et al.* 2012). On the other side, candidate sentences will also be extracted through semi-automatic distributional analysis of a large corpus - i.e. CORIS (Rossini Favretti *et al.* 2002) - and refined through linguistic analysis and manual validation of data thus obtained.

2 The development of the large scale IFrameNet resource

The need for a large-scale resource cannot be satisfied without resorting to a semi-automatic process for the gathering of linguistic evidence, selection of lexical examples as well as the annotation of the targeted texts. This work is thus at the cross roads of linguistic theoretical investigation, corpus analysis and natural language processing.

On the one hand, the matching between LUs and frames is always granted through manual linguistic validation applied to the data in the development stage. For every Frame the correctness of the inducted LUs is analysed and the 'missing' LUs, that is the BFN LUs' translations, which are absent in the inducted LU's list, are detected.

On the other hand, most choices rely on large sets of corpus examples, as made available by CORIS. Finally, the scaling to large sets of textual examples is supported by automatically searching candidate items through semantic pre-filtering over the corpus: frame phenomena are here used as queries while intelligent retrieval and ranking methods are applied to the corpus material to minimize the manual effort involved.

In the following section, we will sketch the main stages of the process that integrate the above paradigms.

2.1 Integrating corpus processing and lexical analysis for populating IFrameNet

The beneficial contribution of the interaction between corpus processing techniques and lexical analysis for the semi-automatic expansion of the FrameNet resource has been discussed since (Pennacchiotti *et al.* 2008), where LU induction is presented as the task of assigning a generic lexical unit not yet present in the FrameNet database (the so-called unknown LU) to the correct frame(s). The number of possible classes (i.e. frames) and the problem of multiple assignment make it a challenging task. This task is discussed in (Pennacchiotti *et al.* 2008, De Cao *et al.* 2008, Croce and Previtali 2010), where different models combine distributional and paradigmatic lexical information (i.e. derived from WordNet) to assign unknown LUs to frames. In particular, distributional models are used to select a list of frames suggested by the corpus' evidence and then the plausible lexical senses of the unknown LU are used to re-rank proposed frames.

In order to rely on comparable representations for LUs and sentences for transferring semantic information from the former to the latter, we exploit Distributional Models (DM) of Lexical Semantics, in line with (Pennacchiotti *et al.* 2008) and (De Cao *et al.* 2008). DMs are intended to acquire semantic relationships between words, mainly by looking at the word usage. The foundation for these models is the Distributional Hypothesis (Harris 1954), i.e. words that are used and occur in the same "contexts" tend to be semantically similar. A context is a set of words appearing in the neighborhood of a target predicate word (e.g. a LU). In this sense, if two predicates share many contexts then they can be considered similar in some way. Although different ways for modeling word semantics exist (Sahlgren 2006; Pado and Lapata 2007; Mikolov *et al.* 2013; Pennington *et al.* 2014), they all derive vector representations for words from more or less complex processing stages of large-scale text collections. This kind of approach is advantageous in that it enables the estimation of semantic relationships in terms of vector similarity. From a linguistic perspective, such vectors allow for some aspects of lexical semantics to be geometrically modelled, and to provide a useful way to represent this information in a machine-readable format. Distributional methods can model different semantic relationships, e.g. topical similarities (if vectors are built considering the occurrence of a word in documents) or paradigmatic similarities (if vectors are built considering the occurrence of a word in the (short) contexts of another word (Sahlgren 2006)). In such models, words like *run* and *walk* are close in the space, while *run* and *read* are likely to be projected in different subspaces. Here, we concentrate on DMs mainly devoted to modelling paradigmatic relationships, as we are more interested in capturing phenomena of quasi synonymy, i.e. semantic similarity that tends to preserve meaning.

2.2 The development cycle

In the following paragraphs, we outline the different stages in the development process. Each stage corresponds to specific computational processes.

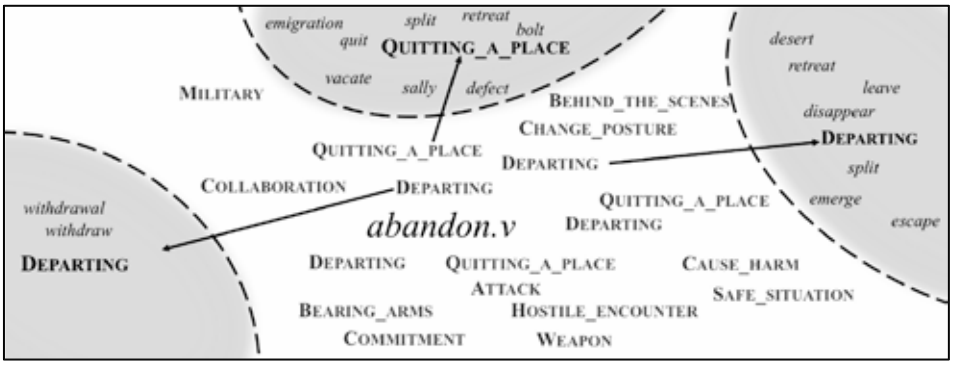


Figure 1: Three lexical clusters for the frames triggered by the verb *abandon*.v: pairs closed in the map correspond to (paradigmatic) semantic similar words and frames

Validation of existing resources. At this stage, the existing resources, dating back to previous work, are analysed and manually pruned of errors such as lexical units wrongly assigned to frames (e.g. *‘asta’* or *‘colmo’* to the Frame ‘BODY_PARTS’), or words never assigned to their correct frame, for instances the LU ‘*piede*’ or ‘*mano*’ for the Frame ‘BODY_PARTS’.

All the acquired Italian LUs have been compared, frame by frame, to BFN’s ones, using bilingual dictionaries (e.g. Oxford bilingual dictionary) and WordNet in order to verify the correctness of matching between lexical and frames. Over the 15,134 automatically acquired ⟨*LU, frame*⟩ pairs (6,670 nouns and 8,464 verbs and adjectives), 7,377 LUs have been considered correctly assigned (2,506 verb and adjective and 4,871 noun pairs).

In addition, bilingual dictionaries, ItalWordNet and MultiWordNet have been used to manually insert a list of missing lexical entries for each frame. At the end of the process, the resulting validated and refined ⟨*LU,frame*⟩ amount to 7,902 (5,128 nouns and 2,774 verbs and adjectives).

Corpus processing and lexical modeling. At this stage, the LUs made available from manual validation are used to model distributionally the individual frames. Firstly, distributional corpus analysis is applied to map individual LUs into distributional vectors. A distributional model will be acquired from the CORIS corpus by applying the neural method presented in (Mikolov *et al.* 2013). It will enable the acquisition of geometrical representations for words in a high dimensional space where distance reflects the paradigmatic relation among words. This model can also be adopted to build a representation for sentences, as traditionally carried out by Distributional Semantic models, e.g. (Landauer and Dumais 1997) or (Mitchell and Lapata, 2010).

Lexical clustering is important here as specific space regions enclosing the instance vectors of some considered LUs correspond to semantically coherent lexical subsets. This is a priming function for mapping unseen word vectors to frames, as applied in (De Cao *et al.* 2008): the centroids of the possibly multiple clusters generated by the known LUs of a given frame *f* are used to detect all regions expressing *f* and thus predict the predicate *f* over previously unseen words and sentences. Examples of semantically coherent regions evoked by the verb *abandon* for the English Framenet are reported in Fig. 1. Here different lexical clusters for a given frame (i.e. DEPARTING) are depicted while different frames (e.g. DEPARTING, QUITTING_A_PLACE, COLLABORATION) are also evoked by the verb. It should be noted that in the figure distances in the two-dimensional plot correspond to distances between the word embedding vectors, while each lexical cluster is expressed as the centroid of its member vectors.

The distributional information has been acquired for the considered 7,902 LUs from CORIS and used to support the LU mapping and the sentence validation. In fact, given a sentence *s* containing a target LU *l*, a specific geometrical representation for *s* can be derived by linearly combining all vectors representing words *w* surrounding *l* in sentence *s*. This duality property allows the embedding space to represent sentences *s*, lexical units *l* as well as generic words *w*. This enables to model the relevance of a frame *f* for an incoming sentence *s* through the distance *d*(*f,s*) between vectors $\underline{f}$ related to a centroid for a frame *f* and the vector $\underline{s}$ of the sentence *s*. It corresponds to a confidence measure computed for a rule such as:

“*s* is a valid example of the usage of frame *f*”

The open aspects of the above semi-automatic process are the following:

I. How to design a suitable representation (centroid or model) for a frame *f*
II. How to define the vector for a sentence *s*
III. How to compute the distance function *d*(*f,s*)

The current research activity is focusing on the best solution for these issues and part of our experimental activity is devoted to assess these design choices, as discussed in Section 3.

First Lexical Analysis and Validation. A further stage for the resource development focuses on the selection of a significant sample of LUs, chosen on the basis of their high semantic salience and for their high number of occurrences in the corpus (*primary* LUs). By relying on the method described above, we use the distributional representation of words, lexical units and sentences, to gather CORIS sentences *s* where a LU occurs and evaluate its suitability as an example for the evoked *f*. This decision function is based on the geometric distance *d*(*f,s*) that can be computed over a large number of sentences *s*. When this step is carried out in CORIS, the validation of the acquired candidate sentences allows for positive examples of a frame *f* to develop quickly*:* this is used to trigger supervised learning of *f*.

The manually validation in fact confirms the proper correspondence between automatically selected sentences and LUs that evoke a targeted frame *f*. It produces novel seed examples for *f*: these will serve as a training set for a semi-automatic stage of resource expansion.

Semi-automatic resource expansion. The acquired distributional model will support the semi-automatic expansion of the seed set, by selecting the most semantically similar word to the seed set and assigning them to frames by applying the methodologies suggested in (Pennacchiotti *et al.* 2008, De Cao *et al.* 2008, Croce and Previtali 2010). Moreover, the same distributional model will support the assignment process of sentences to frames. We will in fact investigate semi-supervised models based on clustering techniques (Pennacchiotti *et al.* 2008) or other supervised approaches such as Support Vector Machines as in (Croce and Previtali 2010).

Final Validation and Release. The extracted sentences will be ordered by decreasing probability, according to their distributional collocation, and a list of 15 to 20 candidates per LU will be provided. This list will be manually validated. The aim is to provide at least 4 sample sentences for each of the primary LUs.

3 Status of the Project and Perspective Views

Although the general software architecture for the project progress is available, the overall process described above has not been fully accomplished.

Current material covers a set of 554 frames and 7,902 lexical units, of which 2,604 verbs, 5,128 nouns and 170 adjectives. The average number of occurrences for each of these selected words is higher than 9,400, although there are still 508 words not present in CORIS.

All these occurrences correspond to a number of about 70 millions non validated and unsorted sentences. In the rest of the paper, we describe the outcome of the First Lexical Analysis and Validation stage: its aim is to trigger the semi-automatic learning and tagging of the whole corpus, according to the methods suggested in section 2.2.

3.1 Empirical Investigation: First Lexical Analysis and Validation

The stage *First Lexical Analysis and Validation* has been currently accomplished. The three research questions posed above: (I) the modelling of a frame *f*, (II) the sentence representation and (III) the definition of a distance function able to model similarity between sentences.

About the problem (I) two approaches are possible. We can model a frame via clustering its lexical units and applying the method described in (Pennacchiotti *et al.* 2008, De Cao *et al.* 2008). On the contrary, we can adopt a supervised technique. A frame *f* is represented as the target class of instances corresponding to $\langle s,l \rangle$ pairs, where *s* is an input sentence and *l* is a lexical unit: a statistical classifier is trained to map $\langle s,l \rangle$ into a confidence value and its output *h*(*s,l,f*) corresponds to the system's confidence that the sentence

"*f* is the frame evoked by *l* in *s*"

is true. Notice that the pair $\langle s,l \rangle$ can be expressed as an instance by combining the embedding vector $\underline{l}$ of its lexical unit *l* with a vector $\underline{s}$ for *s*.

As a solution for the problem (II) we define $\underline{s}$ as the linear combination of vectors $\underline{w}$, for each word *w* in *s,* i.e. $\underline{s} = \sum_{w \in s} \underline{w}$.

The above formulation allows to define the classification task as follows:

Given a sentence *s* including a word *l* as a potential frame evoking LU, *Find* the frame *f* that characterizes *l* in *s*.

The solution of the above problem over a $\langle s,l \rangle$ pair would also be a useful solution for the problem (III), as the confidence *h*(*s,l,f*) in the classification of a sentence s in a frame f for l can be retained as the inverse of the target distance function $d(f,l)$ local to the sentence.

The major problem with the above formulation is that the training of the statistical classifier is not possible without the availability of useful examples of different frame f. The idea is thus to develop ways to derive from CORIS the proper candidates s for f through the knowledge of some of its LUs. In the bootstrapping stage, we define as virtual examples the pairs $\langle l, \{l\} \rangle$ that are retained as positive examples for the frame f, for every l that is a known lexical unit for f. In our approach, an example is thus obtained by modelling the sentence s as a singleton $\{l\}$, i.e. the lexical unit l.

A statistical classifier considers every known LU as an individual (positive) example and can be applied to every LU in our initial resource (i.e. 7,902 for the 554 frames).

In synthesis, the method works as follows. First, for every lemma w in the corpus, an n-dimensional embedding vector $\underline{w}$ is derived, according to (Mikolov *et al.* 2013). As a side effect, for every LU l of each known frame f, the lexical embedding vector $\underline{l}$ is used to build the example ($\underline{l}$, $\underline{l}$) for the LU sentence pair: $\langle l, \{l\} \rangle$.

A multiclass-statistical categorizer is trained for every frame f for which at least 5 examples (i.e. 5 different LUs) where available.

When applied to an incoming sentence s including a LU l, the classifier outcome $h(l,s,f)$ is said to accept the frame f if:

- f belongs to the set of frames evoked by l
- $f = \text{argmax}_{f'} \{ h(l,s,f') \}$

For every sentence s including a frame evoking lexical unit l, the above function suggests one candidate frame among the possibly multiple ones. When the scoring function h is negative everywhere (e.g. with the SVM formulation of a classification task), the sentence is rejected and is not considered a valid example for future iterations.

The application of this method to the CORIS corpus has been carried out applying a multi-classifier SVM with linear kernel to the $2n$-dimensional vectors of each pair $\langle l, \{l\} \rangle$. Starting from the lexicon validated in the first stages, the SVM has been able to label over 2 million sentences.

3.2 Empirical Investigation: Current Results

In order to evaluate the proposed supervised classification method for the stage "*First Lexical Analysis and Validation*" we run and experimental evaluation over a set of 326[1] frames, the ones with more than 5 lexical units in the initial lexicon. In this way, we selected 1,095 different LUs, represented as an embedding vector in the wordspace. On average, we have 12 LU per frame, and every individual lexical entry l appears in about 1.88 frames. The baseline of a classification task that maps a sentence s including a lexical unit into its own frame is about 35%, as for the ambiguity characterizing most frequent entries.

We asked three annotators to evaluate individual triples $\langle l, s, f \rangle$ validating the system proposal. Four main cases where possible:

- MISSING FRAME. The sentence s is not manifesting any of the frames f evoked by the lexical unit l, but corresponds to a frame not yet present in the lexicon for l. In this case the algorithm cannot provide the suitable frame, as it cannot generate a novel frame.
- NOT APPLICABLE. The sentence s does not contain an occurrence of the lexical unit l in one of its proper senses: this case is typical for phraseological uses of a verb such as *morire di freddo*, *andare di fretta, ...* that do not directly correspond to lexical predicates and thus cannot be treated through the lexical embedding vectors.
- CORRECT/INCORRECT, when the outcome $argmax_{f'} \{ h(l,s,f') \}$ is correct (or incorrect) as the frame evoked by l in s is exactly (or not) f.

According to the above method annotators validated 667 sentences for 113 frames and 212 different *verbal* lexical units. The analysis resulted into a *precision* (i.e. the number of correct candidate frames emitted by the algorithm w.r.t. the number of valid cases, that is all but the MISSING FRAME or NOT APPLICABLE cases) is 75,2%, well beyond the 35% baseline. The method could be applied onto the 74,5% of the sentences, including CORRECT cases and MISSING FRAME cases. We neglected in this coverage score the NOT APPLICABLE cases that amount to 44 sentences, i.e. about 6,4%.

Examples of the correct assignment of the algorithm on quite ambiguous verbs, such as *finire* (i.e. *to end*, in frames ACTIVITY_FINISH, CAUSE_TO_END and KILLING) or *rivelare* (i.e. *to reveal*, in frames REVEAL_SECRET, OMEN, EVIDENCE) are the following:

La vicenda avrebbe potuto [*finire*]$_{\text{ACTIVITY\_FINISH}}$ *lì , ma il prefetto di Nuoro fece presentare ...*

In prova si è [*rivelato*]$_{\text{EVIDENCE}}$ *ad altissimo livello sia sull' asciutto sia sul ...*

[1] By keeping the frames that include at least 4 lexical units the number of targeted frames grows to 371.

An example of Missing Frame is BEAT_OPPONENT for the verb *battere* in

... impegnato a fornire quante più informazioni possibili, anche per [*battere*]$_{\text{BEAT\_OPPONENT}}$ *la concorrenza dei siti Ipsoa e il ...*

as the lexicon of the verb *battere* only includes the frames CAUSE_HARM, CORPORAL_PUNISHMENT and EXPERIENCE_BODI-LY_HARM.

The experiments only run over verbal lexical units will be extended soon to nouns and adjectives. However, the encouraging precision reached by the method allows for direct use it in an iterative active learning schema, where the more ambiguous sentences found and annotated within a specific training stage are used to train the system at the next stage. We expect this to speed up the lexicon development process and to allow bootstrapping with fewer resources. The lexicon will be made available for crowdsourcing further annotations and delivered incrementally in the next few months.

References

Baker C. F., Fillmore, C. J., Lowe, J. B.. (1998). The Berkeley FrameNet project. In: COLING '98 Proceedings of COLING '98, 1. Canada, 86-90.

Basili R., De Cao D., Croce D., Coppola B., Moschitti A. (2009). Cross-Language Frame Semantics Transfer in Parallel Corpora. In: Proceedings of the CICLing 2009, Best Paper Award. Mexico

Bastianelli, E., Castellucci, G., Croce, D., Iocchi, L., Basili, R., & Nardi, D. (2014). HuRIC: a Human Robot Interaction Corpus. In Proceeings of *LREC* 2014, 4519-4526.

Croce, D. and Previtali, D. (2010). Manifold learning for the semi-supervised induction of framenet predicates: an empirical investigation. In Proceedings of GEMS '10, pages 7–16, Stroudsburg, PA, USA.

De Cao D., Croce D., Pennacchiotti M., Basili R. (2008). Combining word sense and usage for modeling frame semantics. In Proceedings of STEP 2008, Italy

De Cao D., Croce D., Basili R. (2010). Extensive Evaluation of a FrameNet-WordNet mapping resource. In: Proceedings of the LREC 2010, Malta.

Fillmore, C.J. (1985). Frames and the semantics of understanding. Quaderni di Semantica, VI(2), 222-254.

Fillmore, Charles J. (1976). Frame semantics and the nature of language, Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, vol. 280, pp. 20-32

Fillmore, C. J. (1982). Frame semantics. Linguistics in the morning calm, pp. 111-137.

Harris, Z. (1954). Distributional structure. In Jerrold J. Katz and Jerry A. Fodor, editors, The Philosophy of Linguistics, New York. Oxford University Press.

Johnson, M. And Lenci, A. (2011). Verbs of visual perception in Italian FrameNet, Advances in Frame Semantics, 3(1), 9–45

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104.

Lenci, A, Johnson, M, Lapesa, G. (2010). Building an Italian FrameNet through Semi-automatic Corpus Analysis. Proceedings of LREC 2010. Malta.

Lenci, A., Montemagni, S., Venturi, G, Cutrullà, M. G. (2012). Enriching the ISST-TANL Corpus with Semantic Frames in Proceedings of LREC 2012, Istanbul, Turkey

Mikolov, T., Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient estimation of word representations in vector space. CoRR abs/1301.3781. http://arxiv.org/abs/1301.3781.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. Cognitive Science, 34(8):1388–1429.

Pado, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. Computational Linguistics, 33(2):161–199.

Pennacchiotti M., De Cao D., Basili R., Croce D., Roth M. (2008). Automatic induction of FrameNet lexical units. In: Proceedings of the EMNLP 2008, Hawaii

Pennington, J., Socher, R. and Manning, C. (2014). GloVe: Global Vectors for Word Representation, In Proceedings of EMNLP 2014, 1532-1543.

Rossini Favretti R., Tamburini F., De Santis C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, Lincom-Europa, Munich, 27-38.

Sahlgren, M.. (2006). The Word-Space Model. Ph.D. thesis, Stockholm University.

Tonelli, S. and Pianta, E. (2008). Frame information transfer from English to Italian. In Proceedings of LREC, Marrekech, Morocco

Tonelli, S, Pighin, D, Giuliano, C, Pianta, E. (2009). Semi-automatic Development of FrameNet for Italian. In Proceedings of the FrameNet Workshop and Masterclass, Milano, Italy. Milan, Italy

Tonelli, S. and Pighin, D. (2009). 'New Features for FrameNet - WordNet Mapping', in Proceedings of CoNLL 2009, Boulder, Colorado, 219–227

Tonelli, S. (2010). "Semi-automatic techniques for extending the FrameNet lexical database to new languages", Università Ca' Foscari, Venezia

Venturi G., Lenci A., Montemagni S., Vecchi E., Sagri M., Tiscornia D., Agnoloni T. (2009). Towards a FrameNet Resource for the Legal Domain. In Proceedings of LOAIT 2009. Barcelona, Spain

-io Nouns through the Ages.
Analysing Latin Morphological Productivity with Lemlat

Marco Budassi
Università degli Studi di Pavia
Corso Strada Nuova 65
Pavia, Italy 27100
marcobudassi@hotmail.it

Eleonora Litta, Marco Passarotti
Università Cattolica del Sacro Cuore
Largo Gemelli 1
Milan, Italy 20123
e.littamodignani@gmail.com
marco.passarotti@unicatt.it

Abstract

English. This paper aims at examining the diachronic distribution of one of the richest classes of nouns in Latin, namely those ending in *-io*. The work is performed through the combined use of a morphological analyser for Latin (Lemlat), and a database collecting all word forms occurring through different periods of Latin language (TF-CILF).

Italiano. *Questo articolo presenta un'analisi della distribuzione diacronica di una delle più ricche classi di nomi in latino, ossia quelli che terminano in -io. Metodologicamente, il lavoro viene condotto attraverso l'uso incrociato di un analizzatore morfologico per il latino (Lemlat) e di una risorsa lessicale contenente tutte le forme di parole latine che occorrono in testi che vanno dall'antichità al neo-latino (TF-CILF).*

1 Introduction

The investigation of lexical data of Classical languages through the use of linguistic resources and Natural Language Processing (NLP) tools has witnessed a surge of interest in the past decade. As far as Latin is concerned, today several textual and lexical resources, as well as NLP tools, are being used in lexicographic research.[1] One of the bedrocks of this type of research is the use of morphological analysers, that is, tools that, given an input word form, output its corresponding lemma(s) and morphological features.

First released at the beginning of the 1990s and recently made freely available in its version 3.0 (Passarotti et al., 2017), Lemlat is one of the best performing morphological analysers and lemmatisers for Latin.[2] Lemlat is currently in the process of being enriched with all lemmas contained in the glossary of Medieval Latin *Glossarium mediae et infimae latinitatis* compiled by Charles Du Cange et alii in 1883-1887 (Glorieux, 2010).

One of the first groups of lemmas from Du Cange which was included into the lexical basis of Lemlat was that collecting all 3rd declension nouns ending in *-io*, one of the most productive affixes in all periods of Latin, up to Romance languages (Fruyt, 2011). The aim of this study is to perform a diachronic quantitative evaluation of 3rd declension nouns ending in *-io*. To do so, first we use Lemlat to lemmatise all word forms of such nouns contained in *Thesaurus formarum totius latinitatis a Plauto usque ad saeculum XXum* (TF-CILF) (Tombeur, 1998). Then we evaluate the results of the lemmatisation in both quantitative and qualitative terms.

2 Lemlat and Du Cange

Lemlat relies on a lexical basis resulting from the collation of three Classical Latin dictionaries,[3] for a total of 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be included in one lexical entry). In the context of the development of Lemlat version 3.0, its lexical basis was further enlarged by adding semi-automatically most of the Onomasticon (26,415 lemmas out of 28,178) provided by the 5th edition of the Forcellini dictionary for Latin (Budassi and Passarotti, 2016). Furthermore, the inflectional information provided by Lemlat has been enhanced with information on derivational morphology taken from the *Word For-*

[1] See (Bamman and Crane, 2008), (McGillivray and Passarotti, 2009), (McGillivray, 2013) and (Passarotti et al., 2016).

[2] `www.lemlat3.eu`. See (Springmann et al., 2016) for a comparative evaluation of the morphological analysers currently available for Latin.

[3] (Georges and Georges, 1913-1918), (Glare, 1982) and (Gradenwitz, 1904).

mation Latin (WFL) lexicon (Litta et al., 2016).[4]

However, being based on dictionaries for Classical Latin, one of the current limitations of Lemlat is the fact that its lexical basis is not large enough yet to provide a wide coverage of the word forms occurring in Late and Medieval Latin texts. For this reason an upgrade of Lemlat 3.0 with the Medieval Latin lemmas contained in the Du Cange glossary (Glorieux, 2010), made available online by the École National des Chartes,[5] is underway.

3 Nouns Ending in *-io*

In the Lemlat lexical basis, nouns of the 3rd declension ending in *-io* (with genitive in *-ionis*) are mostly feminine. Only 294 out of 3,065 *-io* nouns in Lemlat are masculine, more than half of which are proper names.[6] Most frequently, nouns in *-io* derive from verbs. WFL contains 2,510 deverbal nouns in *-io*, 87 denominal, and 36 deadjectival. There are also not derived *-io* nouns, like for instance *bacrio* 'trowel'.

Resulting from one of the main mechanisms for Latin nominalisation (Rosén, 1983), deverbal nouns in *-io* are generally called processes or verbal nouns. Semantically, they can be either "nomina actionis", referring to the process of the action expressed by the input verb (e.g. *aberro* 'to wander from the way' > *aberratio* 'diversion', as the process of wandering from the way), or "nomina rei actae", referring to the result of such process (e.g. *aberratio* as the result of wandering from the way).[7]

An investigation on productivity in affixal derivation performed on the data extracted from WFL has proved that deverbal nouns in *-io* are the most numerous formations in Classical Latin (Litta et al., 2017). Such a high presence of nouns in *-io* in Latin lexicon motivates the choice of them as the object of this work.

For this study, we have grouped the nouns in *-io* as follows:

1. Group D: nouns that are only contained in Du Cange (tot. no. 1,416);
2. Group L: nouns that are only contained in Lemlat (tot. no. 2,246);
3. Group L&D: nouns that are contained in both Du Cange and Lemlat (tot. no. 1,494).

Du Cange contains a total of 2,910 nouns ending in *-io*. One of the characteristics of the Du Cange glossary is indeed that no Classical Latin lemma is included in its lexical basis, and if the same lemma is contained in both lexical bases, it means that it has undergone a major semantic or morphological change. 1,416 *-io* nouns out of 2,910 are listed only in Du Cange (Group D), which means that they were absent in the Classical Latin dictionaries used for compiling Lemlat.

Group L contains all those *-io* nouns whose meaning (or morphology) did not change from Classical Latin throughout time, or that were not used anymore in Medieval Latin. Such words are then exclusive only of the Lemlat lexical basis. Even if they were used in Medieval times, they did not undergo a semantic or morphological change, hence they were not included in Du Cange.

Group L&D contains all those *-io* nouns that are recorded both in Lemlat and Du Cange. These are mostly words that have undergone a semantic change, but there are also cases of words that are spelled differently in Medieval sources (e.g. Med. *adsumtio* or *assumtio* for Cl. *assumptio* 'acquisition'), or that in Medieval times acquired a different inflection (e.g. Cl. *beneficium* 'kindness', 2nd declension > Med. *beneficio*, 3rd declension). Because Du Cange treats different meanings in different entries, there is also a number of words appearing more than once (e.g. *defensio* 'defense' x4, *invocatio* 'invocation' x2).

4 Methodology

In order to perform a diachronic evaluation of the frequency of distribution of these three groups, we have used data extracted from the TF-CILF database (Tombeur, 1998). TF-CILF is a database collecting the vocabulary of the entire Latin world drawn from (a) the ancient Latin literature, (b) the literature of the patristic period, (c) a vast body

[4] Funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 658332-WFL, *Word Formation Latin* is a derivational morphology resource for Latin that links lemmas on the basis of word formation processes (`http://wfl.marginalia.it`).

[5] `http://ducange.enc.sorbonne.fr/doc/sources`.

[6] Because at the moment of writing there is no implemented distinction between onomastic and non-onomastic lemmas for what lemmas in Du Cange are concerned, we have taken into consideration onomastic data also in the Lemlat lexical basis.

[7] An ample bibliography on *-io* nouns in Latin is available. See for example (Fruyt, 1995) and (Fruyt, 2011).

of Medieval material and (d) collections of Neo-Latin works. Word forms are assigned their number of occurrences in each of these four periods.

Lemlat has been already proven to perform very efficiently on the TF-CILF dataset, as it is able to analyse 98.345% of the approximately 63 millions textual occurrences of the word forms it contains (Budassi and Passarotti, 2017).

We extracted from TF-CILF a list including those word forms that feature one of the possible inflectional endings of *-io* nouns (*-io*, *-ionis*, *-ionem*, *-ioni* etc.), together with data on their frequency of occurrence in the four periods of Latin mentioned above. In total we extracted 25,510 candidate word forms.

Then we processed these word forms with both Lemlat 3.0 and an enhanced version of it containing nouns ending in *-io* taken from Du Cange. This version of Lemlat was able to analyse 17,775 word forms out of the 25,510 extracted from TF-CILF. Such a low word coverage (69.79%) is consistent with the overall coverage of TF-CILF word forms provided by Lemlat 3.0 (72.25%) (Budassi and Passarotti, 2017). However, if we look at the number of textual occurrences of these unknown forms, they are extremely rare, which makes the textual coverage of Lemlat 3.0 largely reliable. The automatic processing allows not only to match each word form with a lemma, but also to exclude homographs like *capio* 'to seize' (verb). The resulting output (lemmas + frequency) can be graphically mapped on a temporal axis in order to have a complete view on the distribution of *-io* nouns through the ages.

5 Distribution of *-io* Nouns in Latin

Table 1 offers an overview of the total number of occurrences by period.[8] The vast majority of *-io* nouns are attested in the Middle Ages.

However, any evaluation of these results is going to be biased by the fact that the datasets for each period are not balanced. The size of the subsets covering respectively the Patristic and the Medieval period is bigger than that for Classical Latin. The subset for Neo-Latin is considerably smaller than those for the other periods. To give an idea of the difference in size between the four chronological subsets, Table 2 reports the total number of word forms and lemmas in TF-CILF by period.

	L	**L&D**	**D**
Antiquity	30,282	36,570	1,638
Patres	133,042	255,235	5,740
Medieval	216,220	541,049	14,299
Neo-Latin	19,551	45,145	1,812

Table 1: Absolute frequencies by period.

	Word Forms	**Lemmas**
Antiquity	5,726,051	229,587
Patres	21,982,097	310,348
Medieval	33,285,740	359,262
Neo-Latin	2,184,025	105,857
Total	**63,177,913**	**554,828**

Table 2: Number of word forms and lemmas in TF-CILF by period.

In order to flatten the difference in size between the subsets, relative values need to be used instead of absolute. Table 3 displays the distribution of *-io* nouns in Latin texts in terms of relative frequencies of occurrence by period.

	L	**L&D**	**D**
Antiquity	0.528%	0.638%	0.028%
Patres	0.605%	1.161%	0.026%
Medieval	0.649%	1.625%	0.042%
Neo-Latin	0.895%	2.067%	0.082%

Table 3: Relative frequencies by period.

For instance, looking at Table 3, it turns out that *-io* nouns that are only contained in Lemlat are 0.649% of the total number of occurrences in Medieval texts. Those contained in both Lemlat and Du Cange are 1.625%, and those contained in Du Cange (hence exclusively Medieval) are only 0.042%. An overview of the diachronic distribution of relative frequencies of occurrence of *-io* nouns is provided in Figure 1.

Figure 1 clarifies the variation of the presence of *-io* nouns in different chronological phases of Latin. The distribution of the occurrences of those *-io* nouns that were in the lexicon of Classical Latin (Lemlat line) remains fairly constant across all the diachronic phases of the language. In Neo-

[8] L stands for Lemlat only, L&D stands for Lemlat and Du Cange, D stands for Du Cange only. 'Antiquity' (i.e. up to the end of 2nd century AD), 'Patres' (i.e. 3rd century - 735 AD), 'Medieval' (i.e. 736 - 1499 AD) and 'Neo-Latin' (i.e. 1499 AD henceforth) are chronological parameters adopted by TF-CILF.

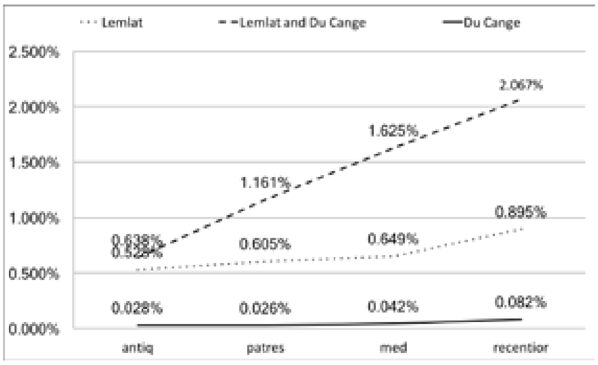


Figure 1: Distribution of relative frequencies of occurrence of *-io* nouns.

Latin times, however, a sharp increase is registered (from 0.649% to 0.895% in terms of relative frequencies). This peak is observable also as far as Medieval Latin *-io* nouns are concerned (Du Cange line). From a value of 0.042% in the Medieval period, the relative frequency raises until 0.082% in Neo-Latin. Nevertheless, the majority of *-io* nouns stored in both Lemlat's and Du Cange's lexical bases (which mostly underwent some semantic change across centuries) are the ones that live the best fate (Lemlat and Du Cange line): they constantly keep growing from the relative frequency value of 0.638% in the Antiquity to the relative frequency value of 2.067% in Neo-Latin.

The odd presence of words from Du Cange in Classical times is due to non-disambiguated homography. For instance, this is the case of the word *dubio*, which is analysed by Lemlat both as a form of the first class adjective *dubius* 'uncertain' (recorded in the original lexical basis of Lemlat, hence here left out) and as the nominative/vocative singular of the *-io* noun *dubio* (a type of hooked tool) from the Du Cange lexical basis.

6 General Discussion

The distribution of *-io* nouns reflects Zipf's law (Zipf, 1949), stating that the frequency of any word in a corpus is inversely proportional to its rank in the frequency table. To put it another way, there are a few *-io* nouns that are massively used, and a lot of *-io* nouns that are used only a few times.

The top most used nouns in *-io* throughout all periods are *ratio* 'reckoning', *passio* 'passion (of Christ)',[9] *oratio* 'speech' and *actio* 'action'. The most used words in Antiquity are *ratio*, *oratio*, *legio* 'legion' and *regio* 'region'. The top most frequent *-io* nouns in Patristic and Medieval times can all be found both in Lemlat and Du Cange. In Patristic literature, the most frequent words (from now on, after *ratio*) are *oratio*, *actio*, *passio* and *resurrectio* 'resurrection'. In Medieval times, they are *passio*, *oratio*, *operatio* 'activity' and *perfectio* 'perfection/completion'.

On another note, the high peak in the relative frequency of *-io* nouns in Neo-Latin texts suggests that these were used more often than others in more recent times. This can be explained by looking at the kind of texts included in the corpus. The texts contained in the Neo-Latin subset are mainly scientific and philosophical treatises, judicial texts, and the text of the Second Vatican Council. When these texts were written, Latin was not the spoken language anymore, as its place was mainly taken by Italian and French, two languages that inherited the suffix *-io* straight from Latin, especially for what learned vocabulary was concerned.[10] The assumption is that learned texts contained a large number of words resembling those used in Italian and French learned language, at least for what *-io* nouns are concerned. A look at the most used *-io* nouns in Neo-Latin texts confirms that once again *ratio* was the most used, followed by *propositio* 'statement of facts', *actio*, *notio* 'judicial enquiry', *definitio* 'definition' and *cognitio* 'examination'. These are also all contained in the Lemlat + Du Cange group.

7 Conclusions and Future Work

In this paper, we presented a study of the diachronic distribution of Latin nouns ending in *-io* by processing word forms from the TF-CILF corpus with the morphological analyser Lemlat. We demonstrated that the *-io* suffix is very productive across all periods of Latin language, showing a particularly high frequency in both Medieval and Neo-Latin texts. *Ratio* remains always the most used *-io* noun across the entire diachronic span covered by the corpus used in our work.

One step further in the study of *-io* nouns would be to establish derivational relationships for each lemma and to verify which of the two lexical groups (Lemlat or Du Cange) the input lemma belongs to. Also, an evaluation of the unknown word

[9] *Passio* is absent in Antiquity texts.

[10] See (Thornton, 1990), (Thornton, 1991) and (Štichauer, 2015).

forms after the lemmatisation process should be performed.

Given the wide lexical coverage provided by Lemlat, our work represents a positive example of how much NLP tools can help to investigate diachronic aspects of language. The wide diachronic as well as diatopic span over which Latin texts are spread opens an appealing challenge for research in NLP, which has to address the problem of portability of NLP tools across time, place and genre. In this sense, Latin texts represent a perfect dataset both for developing and for evaluating techniques of domain-adaptation of NLP tools.

References

David Bamman and Gregory Crane. 2008. Building a Dynamic Lexicon from a Digital Library, In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008, Pittsburgh)* ACM: New York.

Marco Budassi and Marco Passarotti. 2016. Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, 90-94, Association for Computational Linguistics: Berlin.

Marco Budassi and Marco Passarotti. 2017. The Impact of Unassimilated Loanwords on the Latin Lexicon. A Qualitative and Quantitative Analysis. In *Proceedings of DATeCH2017, Göttingen, Germany, June 01-02, 2017*, 85-90. DOI: http://dx.doi.org/10.1145/3078081.3078083.

Charles du Fresne Du Cange 1678-1887. *Glossarium Mediae et Infimae Latinitatis*, éd. augm., Niort, L. Favre http://ducange.enc.sorbonne.fr/.

Michèle Fruyt. 2011 Word-Formation in Classical Latin. In *A Companion to the Latin Language*, ed. James Clackson, 157-175, Wiley-Blackwell: Malden, Mass.

Michèle Fruyt. L'accusatif et les noms en-tio chez Plaute. *De usu, Études de syntaxe latine offertes en hommage à Marius Lavency*, 131-141.

Karl Ernst Georges and Heinrich Georges. 1913-1918. *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hahn: Hannover.

Peter G.W. Glare. 1982. *Oxford Latin Dictionary*. Oxford University Press: Oxford.

Thuillier Glorieux. 2010. Pourquoi informatiser un vieux glossaire? Présentation du Du Cange en ligne. *ÉLA* n°156, octobre-décembre 2009, Klincksieck.

Otto Gradenwitz. 1904. *Laterali Vocum Latinarum*. Hirzel: Leipzig.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC–it 2016). Napoli, aAccademia University Press*, 185-189.

Eleonora Litta, Marco Passarotti and Paolo Ruffolo. 2017. Node Formation. Using Networks to Inspect Productivity in Affixal Derivation in Classical Latin. In *Proceedings of DATeCH2017, Göttingen, Germany, June 01-02, 2017*, 103-108. DOI: http://dx.doi.org/10.1145/3078081.3078092.

Barbara McGillivray. 2013. *Methods in Latin Computational Linguistics* Brill: Leiden.

Barbara McGillivray and Marco Passarotti. 2009. The Development of the Index Thomisticus Treebank Valency Lexicon. In *Proceedings of LaTeCH-SHELT&R Workshop 2009, Athens, Greece*, 43-50, ACL.

Marco Passarotti, Berta González Saveedra and Christophe Onambele. 2016. Latin vallex. A treebank-based semantic valency lexicon for latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) Portorož, Slovenia*, 2599–2606.

Marco Passarotti, Marco Budassi, Eleonora Litta and Paolo Ruffolo 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 24–31.

Hannah Rosén. 1993. The mechanisms of Latin nominalization and conceptualization in historical view. In *ANRW 11I29. 1*, 178-211. De Gruyter: Berlin.

Uwe Springmann, Helmut Schmid and Dietmar Najock. 2016. LatMor: A Latin Finite-State Morphology Encoding Vowel Quantity. In Giuseppe Celano and Gregory Crane (eds.), *Treebanking and Ancient Languages: Current and Prospective Re-search (Topical Issue), Open Linguistics vol. 2*, 386–392.

Pavel Štichauer. 2015. From emergent availability to full profitability: The diachronic development of the Italian suffix -zione from the 16th to the 20th century. In Augendre S., Couasnon-Torlois G., Lebon D., Michard C., et al. *Proceedings of the Décembrettes 8th International conference on morphology*, 319-326, Université de Toulouse: Toulouse.

Anna Maria Thornton. 1990. Sui deverbali italiani in -mento e -zione (I). *Archivio glottologico italiano, LXXV/II*, 169-207. Le Monnier: Torino.

Anna Maria Thornton. 1991. Sui deverbali italiani in -mento e -zione (II). *Archivio glottologico italiano, LXXVI/I*, 79-102. Le Monnier: Torino.

Paul Tombeur. 1998. *Thesaurus formarum totius latinitatis a Plauto usque ad saeculum XXum* Brepols: Turnhout.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press: Cambridge, Mass.

Gender Stereotypes in Film Language: A Corpus-Assisted Analysis

Lucia Busso §
CoLingLab-Università di Pisa*
l.busso0@fileli.unipi.it

Gianmarco Vignozzi §
Università di Pisa**
gianmarco.vignozzi@fileli.unipi.it

§ The research and the writing were carried out by both authors equally. G. V. is responsible for sections 1, 2, 3.1 and 3.2., L.B. for sections 3.3., 4 and 5.

* on leave at the Department of English and Applied Linguistics, University of Birmingham
** on leave at the Department of Linguistics, University of Sydney

Abstract

English: The present study concentrates on the representation and the reception of gender stereotypes. The analysis was first carried out on an ad hoc corpus of cult *romantic comedies and dramedies* of Anglo-American pop contemporary culture and secondly with a perception test. Both the corpus-driven analysis and the test results provide useful insights into the representation, recognition and entrenchment of gender stereotypes in language and in western culture. The preliminary findings generally confirm and validate the scientific literature, although showing some notable new elements.

Italiano: *Il lavoro si incentra sulla rappresentazione e la percezione degli stereotipi di genere. La ricerca è stata prima condotta su un corpus costruito ad hoc di film cult della cultura pop contemporanea anglo-americana appartenenti ai generi* romantic comedy e dramedy*, ed in seguito con un test di percezione. Il duplice approccio utilizzato fa luce sulla rappresentazione, il riconoscimento e il radicamento degli stereotipi di genere nella lingua e nella cultura occidentale. I risultati si trovano in linea con la letteratura, sebbene mostrino alcuni nuovi elementi.*

1 Introduction

In the era of digital revolution and screen proliferation, movies have undoubtedly acquired, thanks to their significance, a pivotal role in shaping our worldviews. In fact, popular films have the power to sway our collective imagination and influence our attitudes on crucial issues related to race, class, gender, etc. Characters in films reflect and perpetuate the status and options of them in today's society and culture, and thus play an active part in creating symbolic role models (Kord 2005, Bednarek 2015). Accordingly, it is interesting to examine the ways in which both females and males are represented on celluloid to better understand the ideologies they bear, and how gender identities are idealized. There seems to be wide agreement on the fact that characterization in filmic discourse heavily relies on archetypes and simplification (Culpeper 2001; Bednarek 2010). This is especially true in gender representation, as stereotypical roles simplify characterization in a way that it is easier to be received by the viewing audience. This, however, often results in an extreme polarization of gender roles. Film dialogues are therefore an ideal ground on which to study gender stereotypes and their linguistic representation and reception. Hence, this paper aims to fathom the discursive representation and the perception of well-established gender stereotypes in the dialogues of a sample of cult British and American romantic comedies, by integrating the tools of discourse analysis, corpus linguistics and perception analysis.

2 Films, language and gender

The nature of film language is still an object of debate. Movie scripts can be classified as texts that are "written-to-be-spoken-as-if-not-written" (Gregory & Carroll 1978: 42). Dialogues, in fact, portray a sort of "prefabricated orality" in that they are carefully written to be performed and sound natural to the audience, who longs for authenticity (Chaume 2012: 81). Corpus-based studies have proved that spontaneous conversation and scripted dialogues are very similar in nature, sharing almost the same array of lexico-grammatical features (Quaglio 2009, Bednarek 2010, Forchini 2012, Baker 2014, amongst others), but due to the evident need for clarity and speed in audio-visual texts, there may be changes in terms of their frequency. In fact, film scripts, sometimes tend to over-use features of spontaneous conversation (e.g.: greetings and leave-takings, Bruti & Vignozzi (2016)) both for dramatic reasons and to render the speech of characters as natural-sounding as possible.

Starting from the premises that gender is socially constructed (Cameron 2010) and that a large part of its perception relies on the observation of pre-established models, television and films provide the perfect field for examining generalized western social representation of accepted human behaviour (Shrum 2008). In this vein, verbal language becomes one of the pivotal means to create, reinforce and most importantly perpetuate stereotypical representations. Canonical research on language and gender has shown that traits such as hedges, empty adjectives, excessively polite forms, intensifiers, troubles talk etc. are more typical of women (Lakoff, 1975; Tannen 1994; Coates 1993), whereas males are associated with substandard and diatopically marked registers (Trudgill 1972; Tannen 1991) and a use of language that is aimed at retaining status and attention. However, nowadays many of these ideas have been partially rejected and framed as stereotypical norms around feminity and masculinity, which do not leave space for diversity (Cameron 2010, Mullany 2007; Bednarek, 2015). In recent times, corpus linguistics and computational linguistics have shown interest in analysing differences in language between genders (Argamom et al, 2003, Baker 2006, Herring & Paolillo 2006, McEnery 2006, Monroe et al. 2008, amongst others). This body of literature represents the backbone structure of our work, which aims to put together "corpus linguistics and gender analysis: two strands of linguistic research that do not go together frequently*"* (Kreyer 2014: 570).

3 Data and corpus driven analysis

The corpus. We compiled a corpus out of the orthographic transcriptions of eight English and American romantic comedies, using the web software *SketchEngine* (Kilgarriff et al. 2004, 2014). The films were chosen not only for their themes, but also for chronological coherence, as they cover approximately the first decade of the 21st century (table 1).

Title	Year	Nation
Sliding Doors	1998	UK
Billy Elliot	2000	UK
Bridget Jones' Diary	2001	UK/USA
Bend It Like Beckham	2002	UK
The Devil Wears Prada	2006	USA
Juno	2007	USA
Eat, Pray, Love	2010	USA
Letters to Juliet	2010	USA

Table 1: corpus rationale

The resulting corpus is therefore a synchronic *ad hoc* corpus of 95,036 tokens. We further subdivided it into two subcorpora consisting of the turns of female and male characters – respectively 55,766 (58.7%) and 39,270 (41.3%) tokens (henceforth: *M* and *F*). We chose to gather a new corpus – instead of relying on existing ones – to obtain a higher control on the data. Moreover, popular romantic comedies are the perfect humus for a polarized representation of gender roles, because of their content and intrinsic structure. As will be seen, however, our results are comparable with the ones extracted from much the larger film corpus *Cornell Movie-Dialogs Corpus*.[1]

Keywords and semantic domains clouds analysis. We used the online text analysis software *WMatrix* (Rayson 2003, 2004) to compare *M* and *F* both against each other and a reference corpus – the *BNC-spoken*. *WMatrix* performs automatic semantic analysis (of English) texts. This semantic analysis is carried out by a first POS tagging phase; the output is then semantically tagged from a set of 21 predefined semantic fields, further subdivided into 232 category labels for more fine-grained classification. Thus, from the comparative analyses starting from males and females' subcorpora, keywords and semantic domains clouds (calculated with log-likelihood statistic). Statistically significant items are the ones with LL values near or over 7, since 6.63 is the cut-off for 99% confidence of significance. The automatically obtained clouds were manually analysed to filter possible errors and select the more significant semantic domains associated with our sub-corpora. From the comparisons of the two sub-corpora against each other and against the *BNC Spoken,* we selected the most relevant semantic domains and keywords (i.e. with the higher LL values) for more qualitative-like evaluation. Tables 2 and 3 report the domains and the keywords that we selected.

[1] The fact that *F* is bigger than *M* should not come as a surprise. The film genre of romantic comedy is generally addressed to women and has therefore more female leading characters.

Sem. domains *F*	Sem. domains *M*
Business: Selling	*Industry*
Evaluation: Authentic	*Evaluation_Inaccurate*
Clothes and Personal Belongings	*Sports*
Time: New and Young	*Money_Generally*
Judgments of Appearance	*Greedy*
People: Female	*People: Male*
Kin	*Foolish*
Informal/Friendly	*Able:Intelligent*
Anatomy and Physiology	***Anatomy and Physiology***
Intimacy and Sex	***Intimacy and Sex***

Keywords *F*	Keywords *M*
Feelings (in_love, love)	*Friendship (lads, man, mate)*
God, oh God, my God	*Swearing (fuck, fuck off, fucking)*
Swearing and Euphemisms (Shit, Shagging)	*Right, all_right*
Mom	*Dad*
Politeness (Thank You, Sorry)	*sorry*
People (Me, My, You)	

Table 2 and 3: WMatrix semantic domains and keywords used in the test

As it can be seen, in our corpus women tend to speak about shopping, cleaning, personal care, and family, whereas men appear to discuss money, sports, work and male friendship. In table 2 are also present semantic domains which were relevant for both *M* and *F* speech, i.e. "Anatomy and Physiology" and "Intimacy and Sex" (in bold). These last two domains may emerge as strongly relevant due to corpus-specific reasons.

Romantic comedies, in fact, are most often centred around romantic and quite physical relationships. However, what we think is of interest when analysing the overlapping between semantic domains between females and males is the different wording. Women and men refer to their bodies and their relationships in different ways, which are consistent with a polarization of gender roles (E.g.: *breasts* vs. *boobs*). Keywords are also worth mentioning. Their evaluation showed that women make larger use of politeness forms, while men resort to more swearwords and interjections, such as "right, all right".

Interestingly, the tendencies that emerged from our small corpus are in line with Schofield and Mehr (2016)'s analysis of the *Cornell Movie-Dialogs Corpus* (Danescu-Niculescu-Mizil et al. 2012a), a vast corpus of more than 600 films of different genres. The similarity of the results gave us confidence in using the stereotypical representations of genders' speech to investigate its reception by means of a test.

The test. With the aim of testing the reception and entrenchment of gender stereotypes in speakers, we developed a perception test based on the results of our corpus-driven analysis. We manually extracted 18 lines per subcorpus[2], each containing one or more of the stereotypical semantic domains and keywords that emerged from the previous *WMatrix* analysis. The resulting 36 extracted lines were used as stimuli in the perception test[3]. The choice of such limited number of sentences was determined by two reasons. The first, theoretically motivated, was not to repeat the same keywords and stereotypes too many times. Such repetition, in our opinion, could have influenced or biased the participants. The second reason, of a more practical nature, was to construct a reasonably-sized test to maintain participants' attention and avoid fatigue, which could have influenced the responses. We extracted film lines containing a variable concentration of stereotypes, ranging from sentences referring to only one to several stereotypical domains. The selection was done manually, based on the rather obvious hypothesis that sentences more "stereotypically dense" would be recognised more easily. The stimuli-sentences were also chosen as deprived of context as possible, in order not to give any clue about the film of origin. Proper names were omitted, and when this was not possible, substituted with the string [XXX]. For example, in (1) the name of the male romantic partner was obscured so that the only clue to the gender of the speaker would be the linguistic stereotypes (shopping, mitigated swearwords, weaving).

1) *When [XXX] and I broke up for two weeks, I bought a loom, a frigging loom*

The test was presented to 22 native, bilingual or highly proficient speakers of English, 15 women and 7 men (mean age: 39.5). The task was to decide whether a given sentence had been uttered by

[2] The stimuli-sentences were chosen to be as representative as possible of the entire corpus: they are evenly distributed among all the films of the corpus, with two or three instances from each film for each subcorpus.

[3] For reasons of space we do not include the complete list of the sentences extracted and used for the test. Several examples are reported in the text and in following footnotes.

a man or a woman. In order not to force participants to a necessarily binary choice, the option "I don't know" was also included. We additionally asked speakers to specify words, expressions or general concepts that influenced their answers. This provided us with interesting insights into participants' process of thinking and categorizing.

4 Results

Several interesting considerations arise from the analysis of the data. Firstly, it appears that overall the stereotypes were correctly spotted and categorized.

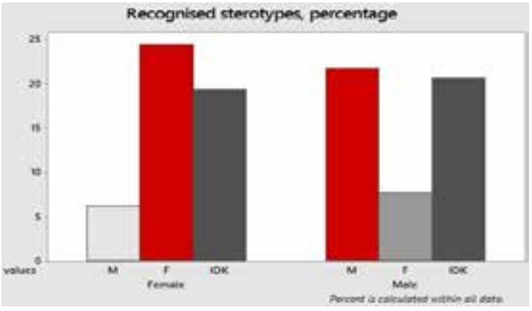


Chart. 1: Percentage of recognised stereotypes (in red)

However, it also emerges that female stereotypes were more unambiguously recognisable, with fewer answers assigned to the other gender or to the "I don't know" category (chart.1).

By examining more closely the results, a subdivision of the data can be made to account for the differences in it: recognised (more than 50% correct), ambiguous (between 25-50% correct) and completely misunderstood (less than 25% correct) stereotypes. Table 4 illustrates the distribution of answers in the three frequency slots.

	> 50%	**25-50%**	**< 25%**
***F* LINES**	**61,1 %**	27,8%	11,1%
***M* LINES**	33,3%	38,9 %	27,8%

Table 4: distribution of participants' answers

As was firstly hypothesized, sentences with a higher "density" of stereotypical keywords or semantic domains were usually the ones that speakers better recognised. Stimuli in the first group, therefore, consist of clear-cut and well recognisable clusters of linguistic and conceptual stereotypes[4]. The second group is instead formed by stereotypes that were recognised by a substantial part of the informants, but not by the majority. This, in our opinion, may be due to several factors: some concepts, for example, could be perceived as less prototypical than others. In addition, some linguistic features (e.g. discourse markers) were not fully recognised as stereotypical due to our limitation to the written dimension. Prosody, contextual information and multimodality are in fact fundamental aspects of language that were inevitably excluded from our experimental design[5]. Finally, the last group consists of stereotypes that were not perceived as such by speakers (e.g.: family as a typical argument of women's speech), and of what we called *reverse stereotypes*. That is, utterances that conceptually represented ambiguous events or anti-prototypical situations: a woman swearing, a man talking about his feelings.[6] As predicted, these stereotypes were not recognised at all by participants, who tended to assign them to the opposite gender. It is interesting to note that also some male-produced sentences were not recognized by our informants, perhaps due to the composition of our corpus. Several predominant keywords and domains in *M*, in fact, may be strictly related to the chosen film genre. For example, the massive presence of the *WMatrix* domain *Evaluation_inaccurate* -- i.e. apologies --reflects the archetypical situation in romantic comedies of men apologizing for their mistakes to women. Being so context-related, however, speakers were not able to correctly locate sentences containing expressions from this domain.[7]

Another aspect that was taken into consideration in our analysis was the gender of the informants, to see if a relation with the data could be recognised. There was a statistically significant difference between the gender of the participant and the answer to the test (H (2) = 9.2388, p-value = 0.0024, *Kruskal-Wallis* test with *Wilcoxon post-hoc*, *Bonferroni* p-value correction).

A chi-square test of independence was performed as well to examine the relation between gender of the speaker and responses given.

[4] E.g.: "*Give me the bag! I've got to get some proper shoes for the wedding now*" (71%) (f); "*What are you doing, eh? You're me best mate!*" (82%) (m).

[5] E.g.: "*God! My mum had a fit when she saw the boots!*" (47%) (f); "*He's a kid. He's just a fucking little kid.*" (47%) (m).

[6] The reverse stereotypes utterances are the following.

I. *Oh, shit! I stubbed my foot on the side of the shagging bath! (f)*
II. *This is the first time in 18 years I'm going to be able to call the shots in my own life*! *(m)*

[7]- *I made a mistake, such a big, BIG mistake and I'm sorry. I'm truly, truly sorry.*
- *We accept that we fight a lot, and we hardly have sex anymore, but we don't wanna live without each other.*

The relation between these variables was significant. (χ^2 = 10.298, p-value= 0.0058).

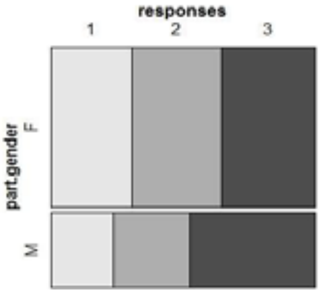


Chart. 2: mosaic plot of the results divided by gender.

Chart 2 shows the difference in male and female informants' answers. The numbers of the variable "responses" indicate the three possible answers of the test: "male" (1), "female" (2), "I don't know" (3). As it can be seen, men assigned overall more utterances to the "I don't know" option rather than to one of the two genders. Women, instead, show a fairly equal distribution of responses among the three conditions. Furthermore, both men and women assigned more utterances to female characters than to male ones (see table 5). This result is in line with the fact that women stereotypes were better recognised overall, in the sense that fewer answers were assigned to the other gender.

	MEN	WOMEN
m	23%	30%
f	29%	34%
idk	48%	36%

Table 5: distribution of informants' answers divided by gender of the speaker

Other useful insights into the data came from the words our informants identified as relevant to their decision. In fact, two tendencies emerged: speakers either indicated specific words, collocations or phrases, or answered with abstract concepts and pragmatic inferences based on the utterances. Interestingly, words and expressions exactly replicated keywords, while general and abstract concepts reflected the semantic domains that emerged in the corpus analysis. In addition, several speakers performed actual pragmatic inferences based on the stereotypical concepts contained in the sentences. For example, to (2) subjects reacted either with a specific word like in a) or with a more general consideration as in b).

2) *Ooh, you must feel like you're about to find your long-lost soul mate!*

a) "soul mate"
b) talking about feelings in general

5 Conclusions

The present paper proposes an original take on investigating gender stereotypes in language. The novelty in our approach lies in the hybrid methodology that falls neither in the tradition of the literature on "gendered discourse" nor in the more recent field of corpus linguistics, but combines the two and adds insights from psycholinguistics as well. This kind of integrated analysis provided us with preliminary results that help identify gender archetypical roles, behaviours and linguistic representations in modern western culture. What is interesting to note is that the gender representations coming to light from our corpus of pop-culture films are based on features that are now dismissed as clichéd and stereotypical by the literature (see Cameron 2005, 2010; Bexter 2006), but which seem to be nonetheless entrenched in our interpretation of reality.

The archetypical depiction of characters is particularly evident in popular comedies, which do not examine characters' psychology in depth. The test validated our assumption that film language stereotypically portrays the way in which men and women talk drawing on recognisable traits attached to femininity and masculinity in our culture. In fact, speakers were mostly able to correctly assign the utterances to the right gender.

In addition, all our informants showed metalinguistic –or second-level –awareness about stereotypical concepts and linguistic clues, and several of them also provided us with insightful and creative inferences based on the event described in the utterance. We interpret this as a sign of stereotypes being conceptual in nature, deeply entrenched in our representation of the world and accessed via linguistic clues. The "reverse stereotypes" also reinforce this idea.

Acknowledgements: We would like to thank professors Silvia Bruti, Alessandro Lenci, Belinda Crawford and Monika Bednarek for useful comments. We also gratefully acknowledge the reviewers for their valuable comments.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. Text & Talk, 23(3), 321–346.

Paul Baker. 2006. Using Corpora in Discourse Analysis. London: Continuum.

Paul Baker. 2014. Using Corpora to Analyze Gender. London; New York: Bloomsbury Academic.

Monika Bednarek. 2010. The Language of Fictional Television: Drama and Identity. London: Continuum.

Monika Bednarek. 2015. Corpus-Assisted Multimodal Discourse Analysis of Television and Film Narratives. In P. Baker, T. McEnery (Eds.), Corpora and Discourse Studies: Integrating Discourse and Corpora, 63-87. Basingstoke, UK: Palgrave Macmillan.

Silvia Bruti and Gianmarco Vignozzi. 2016. Routines as social pleasantries in period dramas: a corpus linguistic analysis. in R. Ferrari,S. Bruti (eds), A Language of One's Own: Idiolectal English, pp. 207-239, Bologna: I libri di Emil.

Deborah Cameron. 2010. Gender, Language and the New Biologism. Constellations, 17 (4), 526–39.

Frederic Chaume. 2012. Audiovisual Translation: Dubbing. Manchester, St Jerome.

Jennifer Coates. 1993. Women, Men and Language. London: Longman.

Jonathan Culpeper. 2001.Language and characterisation: people in plays and other texts. Harlow: Longman.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In Proceedings of ACL, 892–901.

Penelope Eckert and Sally McConnell-Ginet. 2003. Language and Gender. Cambridge: Cambridge University Press.

Pierfranca Forchini. 2012. Movie Language Revisited. Evidence from Multi-Dimensional Analysis and Corpora. Bern: Peter Lang.

Michael Gregory and Susanne Carroll. 1978. Language and Situation: TV Heroines: Contemporary Screen Images of Women. Lanham: Rowman & Littlefield.

Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. Journal of Sociolinguistics, 10(4), 439–459.

Janet Holmes. 2006. Gendered Talk at Work: Constructing Gender Identity through Workplace Discourse. Oxford: Blackwell.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1, 7-36.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, David Tugwell. The Sketch Engine. 2004. Information Technology. Available online at: *www.sketchengine.co.uk*

Susanne Kord and Elisabeth Krimmer. 2005. Hollywood Divas, Indie Queens, and Language Varieties and their Social Contexts. London/New York: Routledge.

Rolf Kreyer. 2014. Review: P. Baker. 2014. Using Corpora to Analyze Gender. London/New York: Bloomsbury. International Journal of Corpus Linguistics 19, 570-575.

Robin Lakoff. 1975. Language and woman's place. New York, NY: Harper & Row.

Tom McArthur. 1981. Lexicon of Contemporary English. London: Longman.

Anthony M. McEnery, Richard Z. Xiao & Yukio Tono. 2006.Corpus-based Language Studies: An Advanced Resource Book. London/New York: Routledge.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372–403.

Louise Mullany. 2007. Gendered Discourse in the Professional Workplace. Basingstoke, NY: Palgrave Macmillan.

Paulo Quaglio. 2009. Television Dialogue: The Sitcom Friends vs. Natural Conversation. Philadelphia: John Benjamins.

Paul Rayson. 2009. Wmatrix: A Web-based Corpus Processing Environment. Computing Department, Lancaster University. Available online at: http://ucrel.lancs.ac.uk/wmatrix/

Paul Rayson, Dawn Archer, Scott Piao, & Anthony M. McEnery. 2004. The UCREL semantic analysis system. Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop, Lisbon, Portugal, 2004.

Alexandra Schofield and Leo Mehr. 2016. Gender distinguishing features in film dialogue. NAACL CLfL.

L.j. Shrum. 2008. Media consumption and perceptions of social reality. In J. Bryant &M.B. Oliver (eds.), Media Effects: Advances in Theory and Research, 3rd Edition. New York, NY: Routledge.

Mary M Talbot. 2003. "Gender Stereotypes: Reproduction and Challenge". In Holmes, J. & Meyerhoff, M. (eds.), The Handbook of Language and Gender. Oxford: Blackwell, 468-86.

Deborah Tannen. 1991. You just don't understand: Women and men in conversation. Virago London.

Deborah Tannen. 1994. Gender and Discourse. New York: Oxford University Press.

Peter Trudgill. 1972. Sex, covert prestige and linguistici change in the urban British English of Norwich. Language in Society 1, 179-195.

How “deep” is learning word inflection?

Franco Alberto Cardillo
Istituto di Linguistica Computazionale
ILC-CNR, Pisa (Italy)
francoalberto.cardillo@ilc.cnr.it

Marcello Ferro
Istituto di Linguistica Computazionale
ILC-CNR, Pisa (Italy)
marcello.ferro@ilc.cnr.it

Claudia Marzi
Istituto di Linguistica Computazionale
ILC-CNR, Pisa (Italy)
claudia.marzi@ilc.cnr.it

Vito Pirrelli
Istituto di Linguistica Computazionale
ILC-CNR, Pisa (Italy)
vito.pirrelli@ilc.cnr.it

Abstract

English. Machine learning offers two basic strategies for morphology induction: lexical segmentation and surface word relation. The first one assumes that words can be segmented into morphemes. Inducing a novel inflected form requires identification of morphemic constituents and a strategy for their recombination. The second approach dispenses with segmentation: lexical representations form part of a network of associatively related inflected forms. Production of a novel form consists in filling in one empty node in the network. Here, we present the results of a recurrent LSTM network that learns to fill in paradigm cells of incomplete verb paradigms. Although the process is not based on morpheme segmentation, the model shows sensitivity to stem selection and stem-ending boundaries.

Italiano. *La letteratura offre due strategie di base per l’induzione morfologica. La prima presuppone la segmentazione delle forme lessicali in morfemi e genera parole nuove ricombinando morfemi conosciuti; la seconda si basa sulle relazioni di una forma con le altre forme del suo paradigma, e genera una parola sconosciuta riempiendo una cella vuota del paradigma. In questo articolo, presentiamo i risultati di una rete LSTM ricorrente, capace di imparare a generare nuove forme verbali a partire da forme gï¿œ note non segmentate. Ciononostante, la rete acquisisce una conoscenza implicita del tema verbale e del confine con la terminazione flessionale.*

1 Introduction

Morphological induction can be defined as the task of singling out morphological formatives from fully inflected word forms. These formatives are understood to be part of the morphological lexicon, where they are accessed and retrieved, to be recombined and spelled out in word production. The view requires that a word form be segmented into meaningful morphemes, each contributing a separable piece of morpho-lexical content. Typically, this holds for regularly inflected forms, as with Italian *cred-ut-o* ’believed’ (past participle, from CREDERE), where *cred-* conveys the lexical meaning, and *-ut-o* is associated with morpho-syntactic features. A further assumption is that there always exists an underlying *base* form upon which all other forms are spelled out. In an irregular verb form like Italian *appes-o* ’hung’ (from APPENDERE), however, it soon becomes difficult to separate morpholexical information (the verb stem) from morpho-syntactic information.

A different formulation of the same task assumes that the lexicon consists of fully-inflected word forms and that morphology induction is the result of finding out implicative relations between them. Unknown forms are generated by redundant analogy-based patterns between known forms, along the lines of an analogical proportion such as: *rendere* ‘make’ :: *reso* ‘made’ = *appendere* ‘hang’ :: *appeso* ‘hung’. Support to this view comes from developmental psychology, where words are understood as the foundational elements of language acquisition, from which early grammar rules emerge epiphenomally (Tomasello, 2000; Goldberg, 2003). After all, children appear to be extremely sensitive to subregularities holding between inflectionally-related forms (Bittner et al., 2003; Colombo et al., 2004;

Dąbrowska, 2004; Orsolini and Marslen-Wilson, 1997; Orsolini et al., 1998). Further support is lent by neurobiologically inspired computer models of language, blurring the traditional dichotomy between processing and storage (Elman, 2009; Marzi et al., 2016). In particular we will consider here the consequences of this view on issues of word inflection by recurrent Long Short Term Memory (LSTM) networks (Malouf, in press).

2 The cell-filling problem

To understand how word inflection can be conceptualised as a word relation task, it is useful to think of this task as a *cell-filling problem* (Blevins et al., 2017; Ackerman and Malouf, 2013; Ackerman et al., 2009). Inflected forms are traditionally arranged in so-called *paradigms*. The full paradigm of CREDERE 'believe' is a labelled set of all its inflected forms: *credere, credendo, creduto, credo* etc. In most cases, these forms take one and only one *cell*, defined as a specific combination of tense, mood, person and number features: e.g. *crede*, PRES IND, 3S. In all languages, words happen to follow a Zipfian distribution, with very few high-frequency words, and a vast majority of exceedingly rare words (Blevins et al., 2017). As a result, even high-frequency paradigms happen to be attested partially, and learners must then be able to generalise incomplete paradigmatic knowledge. This is the cell-filling problem: given a set of attested forms in a paradigm, the learner has to guess other missing forms in the same paradigm.

The task can be simulated by training a learning model on a number of partial paradigms, to then complete them by generating missing forms. Training consists of <lemma_paradigm cell, inflected form> pairs. A lemma is not a form (e.g. *credere*), but a symbolic proxy of its lexical content (e.g. CREDERE). Word inflection consists of producing a fully inflected form given a known lemma and an empty paradigm cell.

2.1 Methods and materials

Following Malouf (in press), our LSTM network (Figure 1) is designed to take as input a lemma (e.g. CREDERE), a set of morpho-syntactic features (e.g. PRES_IND, 3, S) and a sequence of symbols (<*crede*>)[1] one symbol s_t at a time, to output a probability distribution over the upcoming symbol s_{t+1} in the sequence: $p(s_{t+1}|s_t,\text{CREDERE, PRES\_IND, 3, S})$. To produce the form <*crede*>, we take the start symbol '<' as s_1, use s_1 to predict s_2, then use the predicted symbol to predict s_3 and so on, until '>' is predicted. Input symbols are encoded as mutually orthogonal one-hot vectors with as many dimensions as the overall number of different symbols used to encode all inflected forms. The morpho-syntactic features of tense, person and number are given different one-hot vectors, whose dimensions equal the number of different values each feature can take.[2] All input vectors are encoded by trainable dense matrices whose outputs are concatenated into the projection layer $z(t)$, which is in turn input to a layer of LSTM blocks (Figure 1). The layer takes as input both the information of $z(t)$, and its own output at *t–1*. Recurrent LSTM blocks are known to be able to capture long-distance relations in time series of symbols (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997; Jozefowicz et al., 2015), avoiding classical problems with training gradients of Simple Recurrent Networks (Jordan, 1986; Elman, 1990).

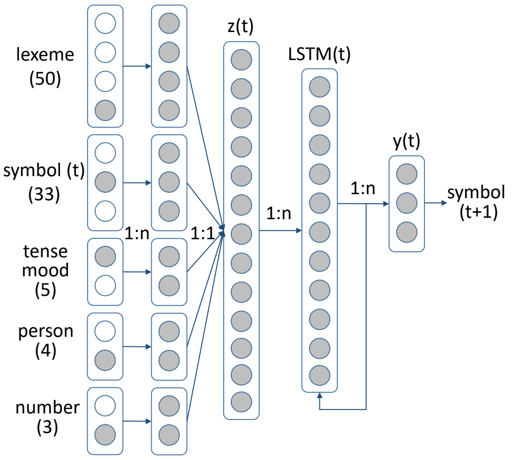


Figure 1: The network architecture. The input vector dimension is shown in brackets. Trainable dense projection matrices are shown as $1:n$, and concatenation as $1:1$.

We tested our model on two comparable sets of Italian and German inflected verb forms (Table 1), where paradigms are selected by sampling the highest-frequency fifty paradigms in two reference corpora (Baayen et al., 1995; Lyding et al., 2014). For both languages, a fixed set of cells was

[1] '<' and '>' are respectively the start-of-word and the end-of-word symbols

[2] Note that an extra dimension is added when a feature can be left uninstatiated in particular forms, as is the case with person and number features in the infinitive.

language	alphabet size	max len	cells	reg / irreg paradigms	forms
German	27	13	15	16 / 34	750
Italian	21	14	15	23 / 27	750

Table 1: The German and Italian datasets.

chosen from each paradigm: all present indicative forms (*n*=6), all past tense forms (*n*=6), infinitive (*n*=1), past participle (*n*=1), German present participle/Italian gerund (*n*=1).[3] The two sets are inflectionally complex: they exhibit extensive stem allomorphy and a rich set of affixations, including circumfixation (German *ge-mach-t* 'made', past participle). Most importantly, the distribution of stem allomorphs is entirely accountable in terms of equivalence classes of cells, forming morphologically heterogenous, phonologically poorly predictable, but fairly stable sub-paradigms (Pirrelli, 2000). Selection of the contextually appropriate stem allomorph for a given cell thus requires knowledge of the form of the allomorph and of its distribution within the paradigm.

3 Results and discussion

To meaningfully assess the relative computational difficulty of the cell-filling task, we calculated a simple baseline performance, with 695 forms of our original datasets selected for training, and 55 for testing.[4]For this purpose, we used the baseline system for Task 1 of the CoNLL-SIGMORPHON-2017 Universal Morphological Reinflection shared task.[5]The model changes the infinitive into its inflected forms through rewrite rules of increasing specificity: *e.g.* two Italian forms such as *badare* 'to look after' and *bado* 'I look after' stand in a BASE :: PRES_IND_3S relation. The most general rule changing the former into the latter is *-are* -> *-o*, but more specific rewrite rules can be extracted from the same pair: *-dare* -> *-do*, *-adare* -> *-ado*, *-badare* -> *-bado*. The algorithm then generates the PRES_IND_3S of - say - *diradare* 'thin out', by using the rewrite rule with the longest left-hand side matching *diradare* (namely *-adare* > *-ado*). If there is no matching rule, the base is used as a default output.

German test	all	regs	irregs
CoNLL baseline	0.4	0.81	0.23
128-blocks	0.68	0.79	0.64
256-blocks	**0.75**	**0.89**	**0.69**
512-blocks	0.71	0.84	0.66

Italian test	all	regs	irregs
baseline	0.65	0.9	0.5
128-blocks	0.61	0.84	0.47
256-blocks	0.63	0.83	0.51
512-blocks	**0.69**	**0.92**	**0.54**

Table 2: Per-word accuracy in German and Italian. Overall scores for the three word classes are averaged across 10 repetitions of each LSTM type.

The algorithm proves to be effective for regular forms in both languages (Table 2). However, per-word accuracy drops dramatically on German irregulars (0.23), and Italian irregulars (0.5). The same table shows accuracy scores on test data obtained by running 128, 256 and 512 LSTM blocks. Each model instance was run 10 times, and overall per-word scores are averaged across repetitions.[6]

The CoNLL baseline is reminiscent of Albright and Hayes' (2003) Minimal Generalization Learner, inferring Italian infinitives from first singular present indicative forms (Albright, 2002). In the present case, however, the inference goes from the infinitive (base) to other paradigm cells. The inference is much weaker in German, where stem allomorphy is more consistently distributed within each paradigm. In Appendix, Table 3 contains a list of all German forms wrongly produced by the CoNLL baseline, together with per-word accuracy of our models. Most wrong forms are inflected forms requiring ablaut, which turn out to be over-regularised by the CoNLL baseline (e.g. **stehtet* for *standet*, **beginntet* for *begannt*). It appears that, in German, a purely syntagmatic approach to word production, deriving all inflected forms from an underlying base, has a strong bias towards over-regularisation. Simply put, the orthotactic/phonotactic structure of the German stem

[3] The full data set is available at `http://www.comphyslab.it/redirect/?id=clic2017_data`. Each training form is administered once per epoch, and the number of epochs is a function of a "patience" threshold. Although a uniform distribution is admittedly not realistic, it increases the entropy of the cell-filling problem, to define some sort of upper bound on the complexity of the task.

[4] Test forms were selected to constitute a benchmark for evaluation. We made it sure that a representative sample of German and Italian irregulars were included for evaluation, provided that they could be generalised on the basis of the training data available.

[5] `https://github.com/sigmorphon/conll2017`(written by Mans Hulden).

[6] The per-word score is 1 (correct), or 0 (wrong).

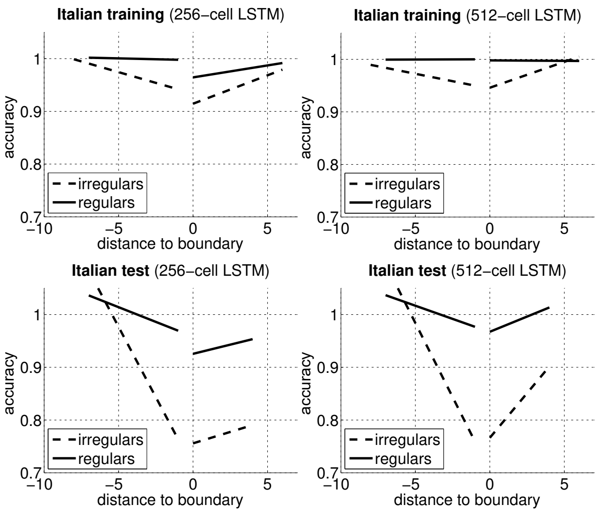


Figure 2: Marginal plots of the interaction between distance to morpheme boundary, stem/inflectional ending, inflectional regularity, stem length and suffix length (fixed effects) in a LME model fitting per-symbol accuracy by a 256-block (left) and a 512-block (right) RNN on training (top) and test (bottom) Italian data. Random effects are model repetitions and word forms.

is less criterial for stem allomorphy than the Italian one. LSTMs are considerably more robust in this respect. Memory resources allowing, they can keep track of local syntagmatic constraints as well as more global, paradigmatic constraints, whereby *all* paradigmatically-related forms contribute to fill in gaps in the same paradigm. For example, knowledge that a paradigm contains a few stem allomorphs is good reason for an LSTM to produce a stem allomorph in other (empty) cells. The more systematic the distribution of stem alternants is across the paradigm, the easier for the learner to fill in empty cells. German conjugation proves to be paradigmatically well-behaved.

An LSTM recurrent network has no information about the morphological structure of input forms. Due to the predictive nature of the production task and the LSTM re-entrant layer, however, the network develops a left-to-right sensitivity to upcoming symbols, with per-symbol accuracy being a function of the network confidence about the next output symbol. To assess the correlation between per-symbol accuracy and "perception" of the morphological structure, we used a Linear Mixed Effects (LME) model of how well structural features of German and Italian verb forms interpolate the "average" network accuracy in producing an upcoming symbol (1 for a hit, 0 for a miss) in both training and test. The marginal plots of Figure 2 show that there is a clear structural effect of the distance to the stem-ending boundary of the symbol currently being produced, over and above the length of the input string. Besides, stems and suffixes of regulars exhibit different accuracy slopes compared with stems and suffixes of irregulars. Intuitively, production of an inflected form by a LSTM network is fairly easy at the beginning of the stem, but it soon gets more difficult when approaching the morpheme boundary, particularly with irregulars. Accuracy reaches the minimum value on the first symbol of the inflectional ending, which marks a point of structural discontinuity in an inflected verb form. From that position, accuracy starts increasing again, showing a characteristically V-shaped trend. Clearly, this trend is more apparent with test words (Figure 2, bottom), where stems and endings are recombined in novel ways. The same results hold for German. On the other hand, no evidence of structure sensitivity was found in a LME model of the baseline output for both German and Italian.

The cell-filling problem is an ecological, developmentally motivated task, based on evidence of fully inflected forms. Although other (simpler) models have been proposed to account for form-meaning mapping in Morphology (Baayen et al., 2011; Plaut and Gonnerman, 2000, among others), we do not know of any other artificial neural networks that can simulate word inflection as a cell-filling task. Unlike more traditional connectionist architectures (Rumelhart and McClelland, 1986), recurrent LSTMs do not presuppose the existence of underlying base forms, but they learn possibly alternating stems upon exposure to full forms. Admittedly, the use of orthogonal one-hot vectors for lemmas, unigram temporal series for inflected forms, and abstract morpho-syntactic features as a proxy of context-sensitive functional agreement effects, are crude representational short-hands. Nonetheless, in tackling the task, LSTMs prove to be able to orchestrate "deep" knowledge about word structure, well beyond pure surface word relations: namely stem-affix boundaries, paradigm organisation and degrees of regularity in stem formation. Acquisition of different inflectional systems may require a different balance of all these pieces of knowledge.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*. Oxford University Press.

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

Adam Albright. 2002. Islands of reliability for regular morphology: Evidence from italian. *Language*, pages 684–709.

Harald R. Baayen, P. Piepenbrock, and L. Gulikers, 1995. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Harald R. Baayen, Petar Milin, Dusica Filipović Ðurđević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, 118(3):438–481.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Dagmar Bittner, Wolfgang U. Dressler, and Marianne Kilani-Schoch, editors. 2003. *Development of Verb Inflection in First Language Acquisition: a cross-linguistic perspective*. Mouton de Gruyter, Berlin.

James P. Blevins, Petar Milin, and Michael Ramscar. 2017. The zipfian paradigm cell filling problem. In Ferenc Kiefer, James P. Blevins, and Huba Bartos, editors, *Morphological Paradigms and Functions*. Brill, Leiden.

Lucia Colombo, Alessandro Laudanna, Maria De Martino, and Cristina Brivio. 2004. Regularity and/or consistency in the production of the past participle? *Brain and language*, 90(1):128–142.

Ewa Dąbrowska. 2004. Rules or schemas? evidence from polish. *Language and cognitive processes*, 19(2):225–271.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Jeffrey L Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4):547–582.

Adele E Goldberg. 2003. Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Michael Jordan. 1986. Serial order: A parallel distributed processing approach. Technical Report 8604, University of California.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell' Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisá corpus of italian web texts. Proceedings of the 9th Web as Corpus Workshop (WaC-9)@ EACL 2014, pages 36–43. Association for Computational Linguistics.

Robert Malouf. in press. Generating morphological paradigms with a recurrent neural network. *Morphology*.

Claudia Marzi, Marcello Ferro, Franco Alberto Cardillo, and Vito Pirrelli. 2016. Effects of frequency and regularity in an integrative model of word storage and processing. *Italian Journal of Linguistics*, 28(1):79–114.

Margherita Orsolini and William Marslen-Wilson. 1997. Universals in morphological representation: Evidence from italian. *Language and Cognitive Processes*, 12(1):1–47.

Margherita Orsolini, Rachele Fanari, and Hugo Bowles. 1998. Acquiring regular and irregular inflection in a language with verb classes. *Language and cognitive processes*, 13(4):425–464.

Vito Pirrelli. 2000. *Paradigmi in morfologia. Un approccio interdisciplinare alla flessione verbale dell'italiano*. Istituti Editoriali e Poligrafici Internazionali, Pisa.

David C Plaut and Laura M Gonnerman. 2000. Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, 15(4/5):445–485.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of english verbs. In David E. Rumelhart, James L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing. Explorations in the Microstructures of Cognition*, volume 2 Psychological and Biological Models, pages 216–271. MIT Press.

Michael Tomasello. 2000. The item-based nature of children's early syntactic development. *Trends in cognitive sciences*, 4(4):156–163.

Appendix A. Comparative test results

base form	target	CoNNL **baseline**	LSTM 128	LSTM 256	LSTM 512
bleiben	bliebt	**blieb**	0	0	0
dï¿œrfen	gedurft	**gedï¿œrfen**	0.2	0.2	0
sein	seiend	**seind**	0	0	0
mï¿œssen	gemusst	**gemï¿œssen**	0.2	0.1	0.1
bestehen	bestandet	**bestehtet**	0.5	0.6	0.2
sprechen	spricht	**sprecht**	0	0.5	0.2
geben	gibt	**gebt**	0	0.3	0.3
sehen	siehst	**sehst**	0	0	0.3
tun	tatet	**tut**	0.3	0	0.3
stehen	standet	**stehtet**	0.2	0.1	0.4
fahren	fï¿œhrst	**fahrst**	0.2	0.6	0.5
finden	fandet	**findet**	0.8	0.6	0.6
dï¿œrfen	darf	**dï¿œrfe**	0.5	0.7	0.7
fahren	fuhrst	**fahrtest**	0.4	0.4	0.7
beginnen	begannt	**beginntet**	0.6	0.9	0.8
kommen	kamst	**kommst**	1	1	0.8
liegen	lagt	**liechtet**	0.5	0.9	0.8
sehen	saht	**sehtet**	0.9	0.8	0.8
bringen	brachtet	**brinchtet**	1	1	0.9
fragen	fragtet	**frugt**	1	1	0.9
gehen	gingt	**gehtet**	0.9	1	0.9
haben	hattet	**habt**	1	1	0.9
nehmen	nahmt	**neht**	0.9	1	0.9
nennen	nanntet	**nenntet**	0.8	1	0.9
sagen	sagtet	**sugt**	1	1	0.9
tragen	trï¿œgst	**tragst**	0.9	0.9	0.9
bitten	baten	**bitten**	1	1	1
denken	dachtest	**denkest**	1	1	1
geben	gabst	**gebst**	1	1	1
scheinen	schienst	**scheintest**	0.8	1	1
setzen	setztet	**setzet**	1	1	1
sprechen	sprachst	**sprechtest**	0.8	1	1
werden	wurdet	**werdet**	0.9	1	1

Table 3: Comparative results for the 33 German verb forms that are wrongly inflected by the CoNNL baseline (highlighted in bold). In most cases, forms are over-regularised. Results are ordered by increasing accuracy of the 512-block LSTM model. Accuracy scores are given per word, and averaged across repetitions of each LSTM model in the [0, 1] range: '0' means that the output is wrong in all model repetitions, '1' that it is always correct. The most accurate results are provided by the 256-block LSTM model.

Tree LSTMs for Learning Sentence Representations

Héctor Cerezo-Costas
AlantTic, Gradiant
Universidade de Vigo, Spain
Edificio CITEXVI, local 14
Vigo, Pontevedra 36310, SPA
hcerezo@gradiant.org

Manuela Martín-Vicente
Gradiant
Edificio CITEXVI, local 14
Vigo, Pontevedra 36310, SPA
mmartin@gradiant.org

F.J. González-Castaño
Dept. Enxeñaría Telemática
E.E. de Telecomunicación
Universidade de Vigo, SPA
javier@det.uvigo.es

Abstract

English. In this work we obtain sentence embeddings with a recursive model using dependency graphs as network structure, trained with dictionary definitions. We compare the performance of our recursive Tree-LSTMs against other deep learning models: a recurrent version which considers a sequential connection between sentence elements, and a bag of words model which does not consider word ordering at all. We compare the approaches in an unsupervised similarity task in which general purpose embeddings should help to distinguish related content.

Italiano. *In questo lavoro produciamo sentence embedding con un modello ricorsivo, utilizzando alberi di dipendenze come struttura di rete, addestrandoli su definizioni di dizionario. Confrontiamo le prestazioni dei nostri alberi-LSTM ricorsivi con altri modelli di apprendimento profondo: una rete ricorrente che considera una connessione sequenziale tra le parole della frase, e un modello bag-of-words, che non ne considera l'ordine. La valutazione dei modelli viene effettuata su un task di similarit non supervisionata, in cui embedding di uso generale aiutano a distinguere i contenuti correlati.*

1 Introduction

Word embeddings have succeeded in obtaining word semantics and projecting this information in a vector space. (Mikolov et al., 2013) proposed two methodologies for learning semantic abstractions of words from large volumes of unlabelled data, Skipgram and CBOW, comprised in the word2vec framework. Another approach is GloVe (Pennington et al., 2014), which learns from statistical co-occurrences of words. The two conceptually similar algorithms employ a sliding window of words, the context, with the intuition that words appearing frequently together are semantically related and thus should be represented closer in $\mathbb{R}^n$. The resulting vectors have shown strong correlation with human annotations in word-analogy tests (Griffiths et al., 2007).

Despite the success of word embeddings in capturing semantic information, they cannot obtain on its own the composition of longer constructions, which is essential for natural language understanding. Thus, several methods using deep neural networks combine word vectors for obtaining sentence representations with linear mappings (Baroni and Zamparelli, 2010) and deep neural networks, which make use of multiple network layers to obtain higher levels of abstraction (Socher et al., 2012). One of the first approaches of obtaining generic embeddings was Paragraph2Vec (Le and Mikolov, 2014). Paragraph2Vec can learn unsupervised sentence representations, analogous to word2vec models for word representation, by adding an extra node, indicating the document contribution, to the model.

Attending to the way the nodes of the network link with each other, two approaches are frequent in NLP: recurrent neural networks and recursive neural networks (RNN) [1]. Recurrent models consider sequential links among words, while recursive models use graph-like structures for organizing the network operations. They process neighbouring words by parsing the tree order (dependency or syntactic graphs), and compute node representations for each parent recursively from the previous step until they reach the root of the tree, which gives the final sentence abstraction.

In this work, we train a variant of Tree-LSTM models for learning concept abstractions with dic-

[1] We use the same classification as in (Li et al., 2015).

tionary descriptions as an input. To the best of our knowledge, this is the first attempt to embed dictionaries using such approach. Our model takes complex graph-like structures (e.g. syntactic or dependency graphs) as opposed to the most common approaches which employ recurrent models or unordered distributions of words as representation of the sentences. We use an unsupervised similarity benchmark with the intuition that better sentence embeddings will produce more coincidences with human annotations (comparably to the word analogy task in word embeddings).

2 Related Work

The following recurrent models are capable of obtaining general purpose embeddings of sentences: Skip-thought Vectors, and DictRep.

Skip-thought Vectors (Kiros et al., 2015) learns general semantic sentence abstractions with unsupervised training. This concept is similar to the learning of word embeddings with the skipgram model (Mikolov et al., 2013). Skip-thoughts tries to code a sentence in such a way that it maximises the probability of recovering the preceding and following sentence in a document.

DictRep (Hill et al., 2015) trains RNN networks and BoW models mapping definitions and words with different error functions (cosine similarity and ranking loss). Whilst the RNN models take into account the word orderings, the BoW models are just a weighted combination of the input embeddings. The simplest BoW approach offered competitive results against its RNN counterparts, beating them in most tests (Hill et al., 2016).

Recurrent models have achieved good performance results in different tasks such as polarity detection (e.g. bidirectional LSTMs in (Tai et al., 2015)), machine translation (Cho et al., 2014) or sentence similarity detection (e.g. Skip-thoughts), just to name a few.

Despite being less explored for building general purpose sentence embeddings, in several classification tasks, tree-structured RNNs represent the current state of the art. In their seminal paper, (Socher et al., 2013) captured complex interactions among words with tensor operations and graph-like links among network nodes. Recursive Neural Tensor Networks (RNTN) networks have been used to solve a simplified version of a QA system in (Iyyer et al., 2014).

In (Bowman, 2013), the authors built a natural language inference system using RNTN in a simplified scenario with basic sentence constructions. Although the results show that the system is able to learn inference relationships in most cases, it is unclear if this model could be generalised for more complex sentences. RNTNs were subsequently improved by (Tai et al., 2015), using LSTMs in the network nodes instead of tensors. With tree-structures the network can capture language constructions which greatly affect the polarity of sentences (e.g. negation, polarity reversal, etc.).

A more complete benchmark was conducted by (Li et al., 2015). There, sequential and recursive RNNs were tested in different tasks: sentiment analysis, question-answer matching, discourse parsing and semantic relation extraction. Recursive models excelled in tasks with enough available supervised data, when nodes different from the root are labelled, or when semantic relationships must be extracted from distant words in a sentence.

3 Approach

Learning models that build a dictionary of embeddings have solid advantages over other supervised approaches, since they take advantage of large volumes of data that are already available online. The training data of the system are pairs of definition/target word which can be built with dictionaries or encyclopedia descriptions (e.g. picking the first sentences of a description as training data). We follow previous work of (Hill et al., 2015) that employed dictionaries with sequential connections but using tree structures instead.

We used the Tree-LSTM as the starting point to build our system. The input to the system are the words conforming a definition together with the structure of the graph with the syntactic/dependency relationships, and the word closer to this definition, i.e. the target. Typically the LSTM nodes are intended for strictly sequential information propagation. Our variant is based in the previous work of (Tai et al., 2015).

The main differences with the original LSTM node are the presence of two forget gates instead of one and the operation over two previous nodes of the system which modify node states and inhibitor gates. Hence, sub-indexes 1 and 2 are reserved for left and right child nodes of the graph, respectively. In this LSTM node there are no peephole connections between memory states and the inhibitor gates.

The state value in the root node is fed to the last layer of the system. Then, a non-linear transformation is applied to obtain the sentence embedding. In the basic configuration of the model, the error is measured by calculating the cosine similarity between target and predicted embeddings. The target is the embedding of the word result of the definition. Pre-trained word embeddings or random initialised embeddings might be employed. In the second case, the error is also propagated to the leaf nodes of the graph and thus the word embeddings are updated during training. We did not initialise randomly embeddings because this has consistently produced poorer results in comparison with the same model using pre-trained word embeddings.

In the network configurations of the tree-LSTM models, we added an extra backward link between the root node and the leaves reversing the uplink path (as hinted in (Socher et al., 2011; Paulus et al., 2014)). In these settings, the error to minimise is a combination of the target word similarity and the leaves word similarity modulated by a smoothing parameter.

We implemented our model with Theano (Theano Development Team, 2016) and trained it with minibatch (30) and Adam (Kingma and Ba, 2014) as optimisation algorithm (with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate $l = 0.002$). This configuration has achieved state of the art performance in other NLP tasks (Kumar et al., 2015).

4 Experiments

We compared DictRep (BoW and RNN) and our Tree-LSTM variant in a benchmark of unsupervised text similarity tasks and a supervised task (sentiment polarity). These tasks greatly benefit from a good representation of sentences and it requires a lot of human effort to build a dataset.

DictRep models were trained using available data and online code. For a fair comparison, all models employed the pre-trained word embeddings and training data provided by (Hill et al., 2015) and cosine similarity as error metric. The configuration setting was similar for all the models.

Our model employs two connection configurations: The Tree-LSTM with transformed dependency graphs and the sequential mapping of connections, which is conceptually similar to the DictRep-RNN model.

For SkipThoughts we used the code available online (ski,) and the pre-trained model with a sentence representation of 4800 dimensions. Additionally, we trained a compressed model with sentence and word representation dimensions of 1200 and 320 respectively in about three weeks. Like in the available model, the 80 million registers of the BookCorpus dataset (Zhu et al., 2015) were used during the training process.

The objective of the semantic similarity task benchmark is to measure the similarity between a pair of sentences. SemEval STS 2014 (Agirre et al., 2014) and SICK (Marelli et al., 2014) datasets were used for benchmarks. In both datasets, each example was gold-standard ranked between 0 (totally unrelated sentences) and 5 (completely similar). Furthermore, SICK dataset considers three different types of semantic relatedness (Neutral, Entailment and Contradiction). We tested the models against the three relations to check if recursive and recurrent models exhibited different behaviour.

This is the same dataset used in previous work (Hill et al., 2016) but excluding the WordNet set, since it was used as part of the training.

For the sentiment polarity, we used as training/validation data the Sentiment Penn Treebank dataset [2]. In this dataset, each sentence node is labelled with a 5-tag intensity tag from 0, the most negative, to 4. Sentences are already binarised in the same format of our TreeDict approach so that no preprocessing is needed in this task for TreeModels. We used for training and test the labels at the root node which is the the overall sentence polarity. For completeness, we repeat the analysis for a 3-label annotations over the same dataset. We used the same SVM classifier for all the models and we trained it with the sentence vectors as input.

5 Results and conclusion

The DictRep BoW model was undeniably better than the recurrent and recursive models achieving the best position in all cases (Table 1). The TreeDict-Dep model ranked second [3].

[2] http://nlp.stanford.edu/sentiment/treebank.html

[3] The character “-” indicates that some vectors for a sentence could not be obtained (e.g. due to a malformed dependency graph)

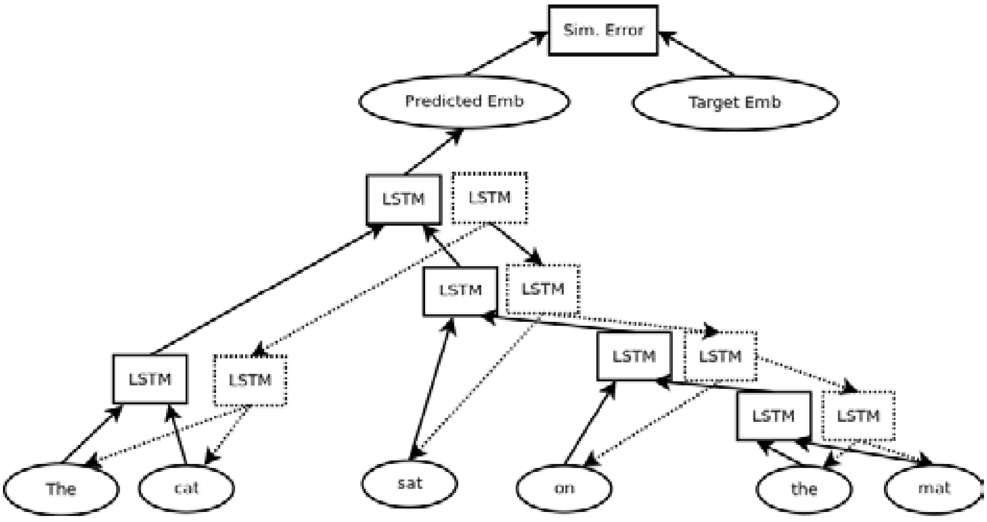


Figure 1: Tree-LSTM schema employed. Dotted blocks and lines depict the optional reverse channel.

All models capture the correlations with human annotations better in neutral contexts. If there are contradictions and entailment relationships, the agreement with human annotations is less evident. Nevertheless, this behaviour is expected and also desirable, as this is an unsupervised benchmark and the system has no way of learning a *similar but conflicting* relationship without external help.

It is clear that BoW models offered the best performance in all the datasets. The Tree-LSTM model, which is consistently better than the sequential models, ranked second. Table 2 shows the correlation among models over the SICK similarity dataset. All the models experience strong cross-correlations between them but the Tree-LSTM with dependency parsing showed the closest correlation with the BoW and recurrent models.

The Table 3 shows the performance of the models in the supervised polarity tasks. BoW and SkipThoughts models experience similar outcomes for the 5 and 3 label task. Models trained with dictionary definitions (DictRep and TreeDict) lag behind those models. However, all the networks using dependency structures have consistently beaten its sequential counterparts. This is a strong indicative of the benefits of using this more complex network structure. The difference between the different network configurations of the same model are less pronounced that in the similarity tasks but in our tests, the models that used the extra link backwards achieved small gains (at least in the 3-label task).

In previous work, (Hill et al., 2016) compared other models in this same similarity benchmark achieving comparable results. Not only DictRep-BoW models outperformed the DictRep-RNNs but also the Skip-thought model, which considers the order of the words in a sentence, was beaten by FastSent, its counterpart that employs BoW representation of a sentence.

The effect of word orderings is not clear. BoW models are far from being ideal as they cannot obtain which parts are negated or the dependencies among the different elements of the sentence (e.g. *the black dog chases the white cat* and *the black cat chases the white dog* cannot be differentiated by only using BoW models).

It is important to mention that the similarity was tested only at the root node when using Tree-LSTM. Notwithstanding, recursive models allow to use more elaborated strategies, taking advantage of the dependencies used to build the relationships of the nodes in the deep network. These strategies could combine similarities at different levels of the sentence to obtain a more approximate value of similarity (e.g. using a pooling matrix with all the nodes of the parse tree (Socher et al., 2011)).

The errors during training time in held-out data were 0.57 for BoW models versus the 0.51 achieved by recurrent and recursive models. Nevertheless, better dictionary embeddings do not seem to directly translate into better performance at inferring general purpose sentence embeddings in the benchmarks. Results in the test also show that we need better mechanisms to infer sentence level representations.

	STS 2014					Sick			
Model	News	Forum	Twitter	Images	Headlines	Neu	Ent	Con	All
DictRep-BoW	.67/.74	.42/.39	.60/.65	.71/.74	.58/.62	.60/.70	.58/.56	.12/.18	.62/.72
DictRep-RNN	.45/.52	.06/.04	.30/.32	.57/.57	.39/.42	.52/.59	.22/.23	.09/.10	.48/.56
TreeDict-Seq	.48/.54	.24/.23	.40/.45	.60/.64	.46/.51	.51/.59	.24/.27	.07/.10	.51/.59
TreeDict-Seq 250	.50/.58	.20/.21	.44/.47	.61/.66	.46/.49	.56/.62	.27/.30	.08/.11	.54/.64
TreeDict-Seq 250BL	.47/.47	.23/.21	.52/.59	.51/.51	.43/.45	.48/.52	.29/.33	.10/.14	.51/.56
TreeDict-Dep	.48/.55	.29/.28	-	.61/.67	-	.56/.64	.35/.39	.08/.13	.55/.65
TreeDict-Dep 250	.50/.56	.31/.30	-	.56/.63	-	.55/.61	.36/.41	.09/.12	.56/.63
TreeDict-Dep 250BL	.43/.45	.30/.28	-	.56/.58	-	.52/.56	.34/.38	.09/.11	.55/.60
SkipThoughts-4800	.43/.23	.13/.13	.42/.40	.48/.51	.36/.37	.49/.49	.19/.25	.10/.15	.48/.50
SkipThoughts-1200	.55/.54	.22/.23	-	.55/.61	.39/.41	.56/.56	.21/.24	.09/.15	.53/.56

Table 1: Performance of the models measured with Spearman/Pearson correlations against golden standard annotations in the similarity benchmarks.

Model	D.BoW	D.RNN	T.Seq	T.Penn
D.BoW	1.0/1.0	.70/.71	.74/75	.80/.82
D.RNN	.70/.71	1.0/1.0	.77/.75	.73/.72
T.Seq	.74/.75	.77/.75	1.0/1.0	.79/.78
T.Dep	.80/.82	.73/.72	.78/.78	1.0/1.0

Table 2: Spearman/Pearson correlations among the different models in the SICK dataset.

Model	**F_1-score**	
	(5-label)	**(3-label)**
DictRep-BoW	.40	.56
DictRep-RNN	.32	.49
TreeDict-Seq	.31	.49
TreeDict-Seq 250	.32	.48
TreeDict-Seq 250BL	.32	.49
TreeDict-Dep	.35	.53
TreeDict-Dep 250	.35	.51
TreeDict-Dep 250BL	.35	.53
SkipThoughts-4800	.40	.56
SkipThoughts-1200	.38	.55

Table 3: Performance of the models in the polarity detection task

In this paper we introduced the use of recursive models for the generation of general purpose embeddings once they are trained by embedding dictionary definitions. We compare recurrent and recursive models in the embedding dictionary task and we test the validity of these embeddings for their use as general purpose codification of sentences with both similarity.

Results demonstrate slight advantages of the Tree recursive variant over recurrent models that learn from dictionaries, which are more frequently employed. Recursive models are more expensive computationally and have a more complex implementation but they exhibit better performance in longer sentences. However, with current learning techniques recurrent and recursive models cannot offer better results than simpler models such as BoW representations of sentences in unsupervised similarity benchmarks. The results of these findings shall be confirmed in the future in more complex scenarios, such as large scale QA.

Acknowledgments

This work has been funded by the Spanish Ministerio de Economa y Competitividad through the project INRISCO (TEC2014-54335-C4-4-R).

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Samuel R Bowman. 2013. Can recursive neural tensor networks learn logical reasoning? *arXiv:1312.6192*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078*.

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in Semantic Representation. *Psychological review*, 114(2):211.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2015. Learning to Understand Phrases by Embedding the Dictionary. *Transactions of the Association for Computational Linguistics*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. *arXiv:1602.03483*.

Mohit Iyyer, Jordan L Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. 2014. A Neural Network for Factoid Question Answering over Paragraphs. In *EMNLP*, pages 633–644.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. *arXiv preprint arXiv:1506.07285*.

Quoc V Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *ICML*, volume 14, pages 1188–1196.

Jiwei Li, Minh-Thang Luong, Dan Jurafsky, and Eudard Hovy. 2015. When Are Tree Structures Necessary for Deep Learning of Representations? *arXiv:1503.00185*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Romain Paulus, Richard Socher, and Christopher D Manning. 2014. Global belief recursive neural networks. In *Advances in Neural Information Processing Systems*, pages 2888–2896.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–43.

Sent2Vec encoder and training code from the paper "Skip-Thought Vectors". `https://github.com/ryankiros/skip-thoughts`. Accessed: 2017-07-07.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Advances in Neural Information Processing Systems*, pages 801–809.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic Compositionality through Recursive Matrix-vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved Semantic Representations from Tree-structured Long Short-term Memory Networks. *ACL*.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *arXiv preprint arXiv:1506.06724*.

Irony Detection: from the Twittersphere to the News Space

Alessandra Cervone, Evgeny A. Stepanov, Fabio Celli, Giuseppe Riccardi
Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Trento, Italy
{alessandra.cervone,evgeny.stepanov}@unitn.it
{fabio.celli,giuseppe.riccardi}@unitn.it

Abstract

English. Automatic detection of irony is one of the hot topics for sentiment analysis, as it changes the polarity of text. Most of the work has been focused on the detection of figurative language in Twitter data due to relative ease of obtaining annotated data, thanks to the use of hashtags to signal irony. However, irony is present generally in natural language conversations and in particular in online public fora. In this paper, we present a comparative evaluation of irony detection from Italian news fora and Twitter posts. Since irony is not a very frequent phenomenon, its automatic detection suffers from data imbalance and feature sparseness problems. We experiment with different representations of text – bag-of-words, writing style, and word embeddings to address the feature sparseness; and balancing techniques to address the data imbalance.

Italiano. *Il rilevamento automatico di ironia è uno degli argomenti più interessanti in sentiment analysis, poiché modifica la polarità del testo. La maggior parte degli studi si sono concentrati sulla rilevazione del linguaggio figurativo nei dati di Twitter per la relativa facilità nell'ottenere dati annotati con gli hashtags per segnalare l'ironia. Tuttavia, l'ironia è un fenomeno che si trova nelle conversazioni umane in generale e in particolare nei forum online. In questo lavoro presentiamo una valutazione comparativa sul rilevamento dell'ironia in blogs giornalistici e conversazioni su Twitter. Poiché l'ironia non è un fenomeno molto frequente, il suo rilevamento automatico risente di problemi di mancanza di bilanciamento nei dati e feature sparseness. Per ovviare alla feature sparseness proponiamo esperimenti con diverse rappresentazioni del testo – bag-of-words, stile di scrittura e word embeddings; per ovviare alla mancanza di bilanciamento nei dati utilizziamo invece tecniche di bilanciamento.*

1 Introduction

The detection of irony in user generated content is one of the major issues in sentiment analysis and opinion mining (Ravi and Ravi, 2015). The problem is that irony can flip the polarity of apparently positive sentences, negatively affecting the performance of sentiment polarity classification (Poria et al., 2016). Detecting irony from text is extremely difficult because it is deeply related to many out-of-text factors such as context, intonation, speakers' intentions, background knowledge and so on. This also affects interpretation and annotation of irony by humans, often leading to low inter-annotator agreements.

Twitter posts are frequently used for the irony detection research, since users often signal irony in their posts utilizing hashtags such as *#irony*, *#justjoking*, etc. Despite the relative ease of collecting the data, Twitter is a very particular kind of text. In this paper we experiment with different representations of text to evaluate the utility of Twitter data for the detection of irony in text coming from other sources such as news fora. The representations of text – bag-of-words, writing style, and word embeddings – are chosen such that they are not dependent on the resources available for the language. Due to the fact that irony is less frequent than literal meaning, the data is usually imbalanced. We experiment with balancing techniques such as random undersampling, random oversampling and cost-sensitive training to observe its effects on a supervised irony detection.

The paper is structured as follows. In Section 2 we introduce related work on irony. In Section 3 we describe the corpora used throughout experiments. In Sections 4 and 5 we describe the methodology and the result of the experiments. In Section 6 we provide concluding remarks.

2 Related Works

The detection of irony in text has been widely addressed. Carvalho et al. (2009) showed that in Portuguese news blogs, pragmatic and gestural text features such as emoticons, onomatopoeic expressions and heavy punctuation marks work better than deeper linguistic information such as n-grams, words or syntax. Reyes et al. (2013) addressed irony detection in Twitter, using complex features like temporal expressions, counter-factuality markers, pleasantness or imageability of words, and pair-wise semantic relatedness of terms in adjacent sentences. This rich feature set enabled the same authors to detect 30% of the irony in movie and book reviews in (Reyes and Rosso, 2014).

Ravi and Ravi (2016), on the other hand, exploited resources such as LIWC (Tausczik and Pennebaker, 2010) to analyze irony in two different domains: satirical news and Amazon reviews; and found out that LIWC's words related to sex or death are good indicators of irony.

Charalampakis et al. (2016) addressed irony detection in Greek political tweets comparing semi-supervised and supervised approaches, with the aim to analyze whether irony predicts election results or not. In order to detect irony, they use as features: spoken style words, word frequency, number of WordNet SynSets as a measure of ambiguity, punctuation, repeated patterns and emoticons. They found that supervised methods work better than semi-supervised in the prediction of irony (Charalampakis et al., 2016).

Poria et al. (2016) developed models based on pre-trained convolutional neural networks (CNNs) to exploit sentiment, emotion and personality features for a sarcasm detection task. They trained and tested their models on balanced and unbalanced sets of tweets retrieved searching the hashtag #sarcasm. They found that CNNs with pre-trained models perform very well and that, although sentiment features are good also when used alone, emotion and personality features help in the task (Poria et al., 2016).

Sulis et al. (2016) investigated a new set of features for irony detection in Twitter with particular regard to affective features; and studied the difference between irony and sarcasm. Barbieri et al. (2014) were the first ones to propose an approach for irony detection in Italian.

Irony detection is a popular topic for shared tasks and evaluation campaigns. Among others, SemEval-2015 (Ghosh et al., 2015) task on sentiment analysis of figurative language in Twitter, and SENTIPOLC 2014 (Basile et al., 2014) and 2016 (Barbieri et al., 2016) tasks on irony and sentiment classification in Twitter. SemEval considered three broad classes of figurative language: irony, sarcasm and metaphor. The task was cast as a regression as participants had to predict a numeric score (crowd-annotated). The best performing systems made use of manual and automatic lexica, term-frequencies, part-of-speech tags, and emoticons.

The SENTIPOLC campaigns on Italian tweets, on the other hand, included three tasks: subjectivity detection, sentiment polarity classification and irony detection (binary classification). The best performing systems utilized broad sets of features ranging from the established Twitter-based features, such as URL links, mentions, and hashtags, to emoticons, punctuation, and vector space models to spot out-of-context words (Castellucci et al., 2014). Specifically, in SENTIPOLC 2016, the best performing system exploited lexica, hand-crafted rules, topic models and Named Entities (Di Rosa and Durante, 2016). In this paper, on the other hand, we address irony detection from features not dependent on language resources such as manually crafted lexica and source-dependent features such as hashtags and emoticons.

3 Data Set

The experiments reported in this paper make use of two data sets: SENTIPOLC 2016 (Barbieri et al., 2016) and CorEA (Celli et al., 2014). While SENTIPOLC is a corpus of tweets, CorEA is a data set of news articles and related reader comments collected from the Italian news website *corriere.it*. The two corpora consist of inherently different types of text. While tweets have a limit on the length of the post, news articles comments are not constrained. The length limitation does not only impact the number of tokens per post, but also the style of writing, since in Tweets authors

SENTIPOLC 2016	CorEA
@gadlernertweet Se #Grillo fosse al governo, dopo due mesi lo Stato smetterebbe di pagare stipendi e pensioni. E lui capeggerebbe la rivolta	bravo, escludi l'universitá .... restare ignoranti non fa male a nessuno, solo a sé stessi. questi sono i nostri.... geni. non mi meraviglierei se votasse grillo
#Grillo,fa i comizi sulle cassette della frutta,mentre alcune del #Pdl li fanno senza,cassetta...solo sulle banane. #ballaró @Italialand	beh dipende da come la guardi..A campagna elettorale all'inverso: rispettano ció che avevano promesso
@MissAllyBlue Non mi fido della compagnia.. meglio far finta di stare sveglio.. sveglissimo O_o	Saranno solo 4 milioni (comunque dimentichi i 42 mil di rimborsi) peró pochi o tanti li hanno restituiti. Gli altri invece , probabilmente politici a te "simpatici" continuano a gozzovigliare con i soldi tuoi . Sveglia volpone

Table 1: Examples of ironic posts from SENTIPOLC 2016 and CorEA.

naturally try to squeeze as much content as possible within the limits.

This difference can be seen also in the type of irony used across the two corpora, as shown in the examples reported in Table 1. While in Tweets we observe much more the presence of external 'sources' (such as URL links, mentions, hashtags and emoticons) to signal the irony and make it interpretable (for example by disambiguating entities using hashtags); news fora users tend to use style much more similar to natural language, where entities are not specifically signaled and there are no emojis to mark the non-literal meaning of a sentence. Thus, CorEA presents a more difficult, but also a more interesting, dataset for automatic irony detection, given the closer similarity to the language used in other genres.

Both corpora have been annotated following a version of the scheme of SENTIPOLC 2014 (Basile et al., 2014). According to the scheme, the annotator is asked to decide whether the given text is subjective or not, and in case it is considered subjective, to annotate the polarity of the text and irony as binary values. The CorEA corpus (Celli et al., 2014) was annotated for irony by three annotators specifically for this paper, and has an inter-annotator agreement of $\kappa = 0.57$.

Since SENTIPOLC 2016 is composed of different data sets, which used various agreement metrics (Barbieri et al., 2016), it is not possible to directly compare the inter-annotator agreements between the corpora. The two component data sets of SENTIPOLC 2016 for which a comparable metric is reported have an inter-annotator agreement of $\kappa = 0.538$ (TW-SENTIPOLC14) and $\kappa = 0.492$ (TW-BS) (Stranisci et al., 2016).

Despite the differences in the number of posts (9,410 for SENTIPOLC and 2,875 for CorEA; see Table 2); due to the length constraint of the former, the corpora have comparable numbers of tokens: 159K for SENTIPOLC and 164K for CorEA. Consequently, there are drastic differences in the average number of tokens per post: 21 for SENTIPOLC and 57 for CorEA. As shown in Table 2, we also observe a major difference in the percentages of ironic posts between the corpora: 12% for SENTIPOLC and 20% for CorEA.

	Non-Ironic		Ironic		Total
	SENTIPOLC 2016				
Training	6,542	(88%)	868	(12%)	7,410
Test	1,765	(88%)	235	(12%)	2,000
CorEA	2,299	(80%)	576	(20%)	2,875

Table 2: Counts and percentages of ironic and non-ironic posts in SENTIPOLC 2016 training and test set and CorEA corpus.

4 Methodology

In this paper we address irony detection in Italian making use of source independent and 'easily' obtainable representations of text such as lexical (bag-of-words), stylometric, and word embedding vectors. The models are trained and tested using Support Vector Machines (SVM) (Vapnik, 1995) with linear kernel and defaults parameters, implemented in the scikit-learn (Pedregosa et al., 2011) python library.

To obtain the desired representations of text, the data is pre- For the bag-of-word representation, the data is lowercased, and all source-specific entities, such as emoji, URL, Twitter hashtags, and mentions are mapped to a single entity (e.g. $\langle$H$\rangle$ for hashtags); as the objective is to use Twitter models to detect irony in news fora and other kinds of textual data, where presence of such entities is less likely. We also apply a cut-off frequency and remove all the tokens that appear in a single document only.

For the style representation, we use the lexical richness metrics based on type and token frequencies such as type-token ratio, entropy, Guiraud's R, Honores H, etc. (Tweedie and Baayen, 1998) (22 features); and character-type ratios, (including specific punctuation marks) (46 features) that previously were successfully applied to tasks such as agreement-disagreement classification (Celli et al., 2016) and mood detection (Alam et al., 2016).

To extract the word embedding representation (Mikolov et al., 2013), we use skip-gram vectors (size: 300, window: 10) pre-trained on Italian Wikipedia, and a document is represented as a term-frequency weighted average of per-word vectors.

Since our goal is to analyze utility of Twitter data for irony detection in Italian news fora, we first experiment with the text representations and chose models that behave above chance-level baseline on per-class F_1 scores and Micro-F_1 score using a 10-fold stratified cross-validation setting. Even though on imbalanced data the frequently used evaluation metric is Macro-F_1 score, e.g. (Barbieri et al., 2016), which we report for comparison purposes; it is misleading as it does not reflect the amount of correctly classified instances. The majority baseline, on the other hand, is very strong for highly imbalanced data sets, and is provided for reference purposes only.

As data imbalance has been observed to adversely affect irony detection performance (Poria et al., 2016; Ptacek et al., 2014), we experiment with simple balancing techniques such as random under- and oversampling and cost sensitive training. While undersampling balances the data set by removing majority class instances, oversampling achieves that by replicating (copying) minority class instances. Undersampling is often reported as a better option, as oversampling may lead to overfitting problems (Chawla et al., 2002). In cost-sensitive training, on the other hand, the performance on minority class is improved by higher misclassification costs for it. In the paper, the selected representations are analyzed in terms of balancing effects and cross-source performance (Twitter - news fora).

5 Results and Discussion

The results of experiments comparing different document representations – bag-of-words, writing style, and word embeddings – are presented in Table 3 for stratified 10-fold cross-validation on both corpora (SENTIPOLC and CorEA). The document representations behave similarly across corpora, and the only representation that achieves above chance-level per-class and micro-F_1 scores is the bag-of-words. At the same time, it achieves the highest macro-F_1 score. However, none of the representations is able to surpass the majority baseline in terms of micro-F_1.

Model	NI	I	Mic-F_1	Mac-F_1
SENTIPOLC: Training				
BL: Chance	0.8783	0.1183	0.7862	0.4983
BL: Majority	0.9378	0.0000	0.8829	0.4689
BoW	0.8979	**0.2112**	0.8207	**0.5546**
Style	0.8817	0.0892	0.7612	0.4605
WE	0.9361	0.0044	0.8799	0.4702
CorEA				
BL: Chance	0.7952	0.1895	0.6733	0.4923
BL: Majority	0.8886	0.0000	0.7996	0.4443
BoW	0.8414	**0.2951**	0.7411	**0.5682**
Style	0.7116	0.1688	0.6186	0.4402
WE	0.8811	0.1447	0.7912	0.5129

Table 3: Average per-class, micro and macro-F_1 scores for stratified 10-fold cross-validation on SENTIPOLC 2016 training set and CorEA for different **document representations**: bag-of-words (*BoW*), stylometric features (*Style*) and word embeddings (*WE*). *BL: Chance* and *BL: Majority* are chance-level and majority baselines. **NI** and **I** are non-ironic and ironic classes, respectively.

The performance of the bag-of-words representation on data balancing techniques is presented in Table 4. The training with natural distribution (*BoW: ND*) yields the best performance across the corpora. For SENTIPOLC data, it is the only model that produces above chance-level (Table 3: *BL: Chance*) performances for per-class and micro-F_1 scores.

Cost-sensitive training (*BoW: CS*) and random oversampling (*BoW: RO*) perform very close. For CorEA corpus, all balancing techniques except random undersampling (*BoW: RU*) yield above chance-level performances. Random undersampling, however, yields the highest F_1 score for the irony class, which unfortunately comes at the expense of the overall performance. This verifies previous observations in the literature that undersampling leads to negative effect on novel imbalanced data (Stepanov and Riccardi, 2011). Since cost-sensitive training achieves the best performance in terms of macro-F_1 score, which was used as official evaluation metrics in SENTIPOLC 2016 (Barbieri et al., 2016), it is retained for SENTIPOLC training-test and cross-corpora (SEN-

Model	NI	I	Mic-F$_1$	Mac-F$_1$
SENTIPOLC: Training				
BoW: ND	**0.8979**	0.2112	**0.8207**	0.5546
BoW: CS	0.8732	0.2493	0.7861	**0.5612**
BoW: RO	0.8737	0.2375	0.7857	0.5555
BoW: RU	0.7270	**0.2679**	0.6115	0.4974
CorEA				
BoW: ND	**0.8414**	0.2951	**0.7411**	0.5682
BoW: CS	0.8331	0.3202	0.7321	**0.5766**
BoW: RO	0.8302	0.3138	0.7279	0.5720
BoW: RU	0.6882	**0.3599**	0.5810	0.5241

Table 4: Average per-class, micro and macro-F_1 scores for stratified 10-fold cross-validation on SENTIPOLC 2016 training set and CorEA for **balancing techniques**: cost-sensitive training (*CS*), random oversampling (*RO*) and random undersampling (*RU*). *ND* is training with natural distribution of classes (*BoW* in Table 3). **NI** and **I** are non-ironic and ironic classes, respectively.

Model	NI	I	Mic-F$_1$	Mac-F$_1$
SENTIPOLC: Training - Test Split				
BL: Chance	0.8826	0.1155	0.7927	0.4990
BL: Majority	0.9376	0.0000	0.8825	0.4688
SoA	0.9115	0.1710	–	0.5412
BoW: ND	**0.9330**	0.1678	**0.8760**	0.5504
BoW: CS	0.9245	**0.2023**	0.8620	**0.5634**
SENTIPOLC - CorEA: 10-fold testing				
BL: Chance	0.8393	0.1213	0.7286	0.4803
BL: Majority	0.8886	0.0000	0.7996	0.4443
BoW: ND	**0.8164**	0.1755	**0.7001**	0.4959
BoW: CS	0.8109	**0.2020**	0.6945	**0.5065**

Table 5: Average per-class, micro and macro-F_1 scores for SENTIPOLC Training-Test split and 10-fold testing of SENTIPOLC models on CorEA for bag-of-words representation with imbalanced (*ND*) and cost-sensitive (*CS*) training. *SoA* are the state-of-the-art results for SENTIPOLC 2016: the system of (Di Rosa and Durante, 2016). *BL: Chance* and *BL: Majority* are chance-level and majority baselines. **NI** and **I** are non-ironic and ironic classes, respectively.

TIPOLC - CorEA) evaluation along with the models trained on natural imbalanced distribution with equal costs.

The final models make use of bag-of-words representation and are trained on SENTIPOLC training set in cost-sensitive and insensitive settings. The evaluation of models is performed on SENTIPOLC 2016 test set and CorEA's 10-folds. This setting allows us to compare our results to the state of the art on SENTIPOLC data and CorEA's cross-validation setting. From the results in Table 5, we observe that on the SENTIPOLC test set both models outperform the state of the art in terms of macro-F_1 score. The model with cost-sensitive training additionally outperforms it in terms of irony class F_1 score. However, both models fall slightly short of outperforming the majority baseline in terms of micro-F_1.

In the cross-corpora setting the behavior of models is similar – cost-sensitive training favors minority class F_1 and macro-F_1 scores. While both models perform worse than the chance-level baseline generated using the label distribution of SENTIPOLC data in terms of micro-F_1, they both outperform it in terms of irony class F_1 score. However, only the model with cost-sensitive training yields statistically significant difference using paired two-tail t-test with $p = 0.05$.

6 Conclusion

We have presented experiments on irony detection in Italian Twitter and news fora data comparing different document representations – bag-of-words, writing style as stylometric features, and word embeddings. The objective is to evaluate the suitability of Twitter data for detecting irony in news fora. The models were compared for balanced and imbalanced training, as well as cross-corpora performance. We have observed that the bag-of-words representation with imbalanced cost-insensitive training produces the best results (micro-F$_1$) across settings, closely followed by cost-sensitive training.

The models outperform the results on irony detection in Italian tweets (Di Rosa and Durante, 2016) in terms of macro-F_1 scores reported for SENTIPOLC 2016 (Barbieri et al., 2016). However, micro-F_1 is the most informative metric for the downstream application of irony detection, as it considers the total amount of true positives. Given that the highest micro-F_1 is attained by the majority baselines for both corpora (0.8829 for SENTIPOLC and 0.7996 for CorEA), the task of irony detection is far from being solved.

Acknowledgments

The research leading to these results has received funding from the European Union – Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 610916 – SENSEI.

We would like to thank Paolo Rosso and Mirko Lai for their help in annotating CorEA.

References

F. Alam, F. Celli, E.A. Stepanov, A. Ghosh, and G. Riccardi. 2016. The social mood of news: Self-reported annotations to design automatic mood detection systems. In *PEOPLES @COLING*.

F. Barbieri, F. Ronzano, and H. Saggion. 2014. Italian irony detection in twitter: a first approach. In *CLiC-it 2014 & EVALITA*.

F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, and V. Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *CLiC-it - EVALITA*.

V. Basile, A. Bolioli, M. Nissim, V. Patti, and P. Rosso. 2014. Overview of the evalita 2014 sentiment polarity classification task. In *EVALITA*.

P. Carvalho, L. Sarmento, M.J. Silva, and E. De Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-. In *Topic-sentiment analysis for mass opinion*.

G. Castellucci, D. Croce, and R. Basili. 2014. Context-aware convolutional neural networks for twitter sentiment analysis in italian. In *EVALITA*.

F. Celli, G. Riccardi, and A. Ghosh. 2014. CorEA: Italian news corpus with emotions and agreement. In *CLIC-it*.

F. Celli, E.A. Stepanov, and G. Riccardi. 2016. Tell me who you are, I'll tell whether you agree or disagree: Prediction of agreement/disagreement in news blogs. In *NLPJ @IJCAI*.

B. Charalampakis, D. Spathis, E. Kouslis, and K. Kermanidis. 2016. A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence*, 51:50–57.

N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

E. Di Rosa and A. Durante. 2016. Tweet2check evaluation at evalita sentipolc 2016. In *CLiC-it - EVALITA*.

A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *SemEval*.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

S. Poria, E. Cambria, D. Hazarika, and P. Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv:1610.08815*.

T. Ptacek, I. Habernal, and J. Hong. 2014. Sarcasm detection on czech and english twitter. In *COLING*.

K. Ravi and V. Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*.

K. Ravi and V. Ravi. 2016. A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*.

A. Reyes and P. Rosso. 2014. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*.

A. Reyes, P. Rosso, and T. Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

E.A. Stepanov and G. Riccardi. 2011. Detecting general opinions from customer surveys. In *SENTIRE @ICDM*.

M. Stranisci, C. Bosco, D.I. Hernández Farías, and V. Patti. 2016. Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In *LREC*.

E. Sulis, D.I. Hernández Farías, P. Rosso, V. Patti, and G. Ruffo. 2016. Figurative messages and affect in Twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, 108:132–143.

Y.R. Tausczik and J.W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*.

F.J. Tweedie and R.H. Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*.

V.N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

Phase-based Minimalist Parsing and complexity in non-local dependencies

Cristiano Chesi
NETS - IUSS
P.zza Vittoria 15
I-27100 Pavia (Italy)
cristiano.chesi@iusspavia.it

Abstract

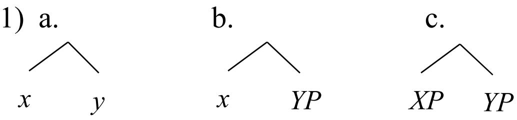
English. A cognitively plausible parsing algorithm should perform like the human parser in critical contexts. Here I propose an adaptation of Earley's parsing algorithm, suitable for Phase-based Minimalist Grammars (PMG, Chesi 2012), that is able to predict complexity effects in performance. Focusing on self-paced reading experiments of object clefts sentences (Warren & Gibson 2005) I will associate to parsing a complexity metric based on cued features to be retrieved at the verb segment (Feature Retrieval & Encoding Cost, FREC). FREC is crucially based on the usage of memory predicted by the discussed parsing algorithm and it correctly fits with the reading time revealed.

Italian. *Un algoritmo di parsing cognitivamente plausibile dovrebbe avere una performance paragonabile a quella umana in contesti critici. In questo lavoro propongo un adattamento dell'algoritmo di Earley che utilizza Grammatiche Minimaliste basate sul concetto di Fase (PMG, Chesi 2012). Associata all'algoritmo, verrà discussa una funzione di costo (Feature Retrieval & Encoding Cost, FREC) capace di misurare la difficoltà relativa al recupero dei referenti coinvolti in dipendenze a distanza. La funzione si basa sui tratti morfosintattici archiviati nel memory buffer utilizzato dal parser. Concentrandosi sulle strutture scisse ad estrazione dell'oggetto, si mostrerà come il FREC risulti predittivo dei dati sperimentali ricavati da studi classici di lettura autoregolata (Warren & Gibson 2005).*

1 Introduction

The last twenty years of formal linguistic research have been deeply influenced by Chomsky's minimalist intuitions (Chomsky 1995, 2013). In a nutshell, the core Minimalist proposal is to reduce phrase structure formation to the recursive application of a binary, bottom-up, structure-building operation dubbed *Merge*. Merge creates hierarchical structures by combining two lexical items (1.a), one lexical item and an already built (by previous application of Merge operations) phrase (1.b) or two already built phrases (1.c).

(1)



Phrases are not linearly ordered by Merge. Only when they are spelled-out (i.e. sent to the Sensory-Motor interface, aka Phonetic Form, *PF*), linearization is required: assuming that *x* and *y* are terminal nodes (i.e. words), either <*x*, *y*> or <*y*, *x*> can both be proper linearizations of (1.a). Hierarchical structure (and linearization) is also determined by another structure building operation: *Move* (or *Internal Merge*, Chomsky 1995); Move re-arranges phrases in the structure by re-merging an item (already merged in the structure) to the edge of the current, top-most, phrase: for instance [*XP* [*YP* [*ZP*]]] can lead to [*ZP* [*XP* [*YP* (*ZP*)]] if *XP* (the *probe*) has a feature triggering movement (e.g. *+f*) and *ZP* (the *goal*) has the relevant feature qualifying it as a plausible target for movement (e.g. *-f*). At the end, the element displaced (*ZP*) will occupy the edge of the structure. When the items within an already built phrase, for instance *XP*, are delivered to PF, they get properly linearized according to their hierarchical structure (e.g. Linear Correspondence Axiom, Kayne 1994), intrinsic phonetic properties (e.g. cliticization), as well as economy conditions (e.g. an items should not be pronounced twice). Such a (cyclic) spell-out happens at *phases*: *XP* will be delivered to PF only if it qualifies as a phase (Chomsky 2013). In this sense, a phase should be a constituent/phrase with some degree of completeness with respect to semantic interpretation (Logic Form, aka LF). Most minimalist linguists agree on the fact that a full-fledged sentence (aka Complementizer Phrase, *CP*) is a phase, the highest argumental shell of a predicate qualifies as a phase (aka little-*v* Phrase, *vP*) and also a full argument is a phase (aka Determiner Phrase, *DP*). Such a simple (and computationally appealing) model has been fully formalized (Stabler 1997, Collins & Stabler 2016) and some parsing algorithm that implements main minimalist insights has been discussed in literature (e.g. Harkema 2001, Chesi 2012 a.o.).

In these pages, I will present some of the advantages of retaining such a simplified computational approach to syntactic derivation. Crucially, I will try to overcome some clear disadvantages in assuming the just presented standard, bottom-up, structure building operations, while obtaining, at the same time, a better empirical fit: on the one hand, I will avoid any non-efficient deductive-parsing perspective (that is a consequence of the assumed bottom-up nature of the Merge and Move operations); on the other, I will promote a more transparent relation between formal competence, parsing and psycholinguistic performance by presenting a simple adaptation of Earley's Top-Down parsing algorithm (Earley 1970) and a complexity metric that refers directly to parsing memory usage: this metric will be able to account for complexity in retrieving the correct item while processing specific non-local dependencies. By "non-local" dependencies I refer to those relations involving movement, namely constructions where the very same item occurs in two distinct, non-adjacent, positions: for instance, *wh*-dependencies in English require the *wh-* item (*who*, in (1)) to be interpreted both in a the left peripheral (focalized) position (the Criterial position, in the sense of Rizzi 2007) and in the thematic lower position (right next to the verb *meet* in (1))[1]:

(1) *Who$_1$* do you think Mary will meet $\_{1}$?

The critical derivation I will discuss in this paper is that of object clefts (Gordon et al. 2001) that with *wh*-questions share a similar non-local dependency formation:

(2) a. It is [$_{DP1}$ the banker|John|me] that [$_{DP2}$ the lawyer|Dan|you] will meet $\_{DP1}$

In short, the head of the dependency (DP1) should be interpreted both as a focalized item and as the direct object (this is where the name of the construction "object cleft" comes from) of the embedded verb. The difficulty of parsing this structure has been deeply discussed in literature (Gordon et al. 2004). What is considered a crucial factor is the role of the similarity between *DP$_1$* and *DP$_2$* (the subject of the cleft, Belletti and Rizzi 2013, §2). To capture this fact, I will re-adapt Earley's algorithm (§3.1) to operate on a specific version of Minimalist Grammar (§3). This would allow us to subsume the similarity effect by predicting reading differences as revealed in self-paced reading experiments (e.g. Warren & Gibson 2005, §4).

2 Parsing with Minimalist Grammars

Since Merge and Move strictly operate "from bottom to top", we expect sentence structure in (2) to be built in 9 steps (and 5 phases: *ph1*, *ph2* …):

1. [$_{ph1}$ the banker]
2. [$_{ph3}$ meet [$_{ph1}$ …]]]
3. [$_{ph3}$ will [meet [$_{ph1}$ …]]]
4. [$_{ph2}$ the lawyer] (independently built)
5. [$_{ph3}$ [$_{ph2}$ …] will [meet [$_{ph1}$ …]]]
6. [$_{ph4}$ that [$_{ph3}$ [$_{ph2}$ …] will [meet [$_{ph1}$ …]]]]
7. [[$_{ph1}$ …] [$_{ph4}$ that [$_{ph3}$ [$_{ph2}$ …] will [meet ($_{ph1}$ …)]]]]
   (*ph$_1$* moves to *ph$_4$* edge)
8. [$_{ph5}$ is [[$_{ph1}$ …] [$_{ph4}$ that [$_{ph3}$ [$_{ph2}$ …] will [meet [$_{ph1}$ …]]]]]
9. [$_{ph5}$ it [is [[$_{ph1}$ …] [$_{ph4}$ that [$_{ph3}$ [$_{ph2}$ …] will [meet [$_{ph1}$ …]]]]]

With the exception of step 4, all other steps must be strictly ordered. As a consequence, moving the direct object in the relevant position would force the linearization to place *ph$_2$* first at the edge of *ph$_3$*, then at the edge of *ph$_4$*. This is how Minimalism derives the relevant non-local dependencies in (2). Obviously this is not transparent at all with respect to parsing (e.g. Fong 2011), where the processing order is expected to be completely reversed:

1. [$_{ph5}$] is initiated
2. [$_{ph1}$] is fully processed while [$_{ph5}$] is still open

[1] Coreference in non-local dependencies will be indicated by the same subscript placed both on the "displaced" item and on the thematic position (the non-pronounced item in the thematic position is indicated with a co-indexed underscore)

3. [*ph4*] is initiated (a Relative Clause)
4. [ph3] is initiated as well (Verbal Phrase)
5. [*ph2* …] is fully processed while [*ph5*], [*ph4*] and [*ph3*] are open
6. [*ph1*] finally receives a thematic role, hence [*ph5*], [*ph4*] and [*ph3*] can be closed.

Unless we deeply revise Minimalist Grammars (both with respect to movement, Fong 2005, and to thematic role assignment, Niyogi & Berwick 2005), we are left with an asymmetry that can not be explained simply in terms of structure building operations as discussed in the next section.

2.1 The "similarity" problem

Warren & Gibson (2005) show that in clefts constructions like the one discussed in (2), the variation of the two DPs [ph1] and [ph2] produces differences in reading time at the verb segment in self-paced reading experiments with the full-DP matching condition ([ph1 the barber] that [ph2 the banker] *praised* …) and proper nouns matching condition ([ph1 John] that [ph2 Dan] *praised* …) ranking higher in terms of difficulty (greatest slow down at verb segment), while pronouns ([ph1 you] that [ph2 we] *praised* …) are easier (fastest reading time). No CFG-based parsing algorithm (in fact, no classic algorithm implements the non-local dependencies in (2) as presented in §2) or Minimalist deductive parsing (parsing strategies exploit the weak equivalence of MGs with multiple Context Free Grammars, Michaelis 1998) have a chance to compare these cases.

3 A processing-friendly proposal

Phase-based Minimalist Grammars (PMGs, Chesi 2012) suitable for parsing of sentences like the ones in (2) can be formalized as follows:

(3) PMG able to parse cleft sentences

Lexicon

[[+D +Sg John*i*] [N _*i*]], [[+D +Sg Dan*i*] [N _*i*]], [N +Sg banker], [N +Sg lawyer], [+D the], [+D +P1 +Pl +case_acc me [N Ø]], [+D +P2 +Sg +case_nom you [N Ø]], [+T will], [+T that], [=[DP (+case_nom)] =[DP (+case_acc)] V meet], [+exp it], [=rCP BE is]

Phases

DP	→	[*DP* ([+F Ø]/[+S Ø]) +D N]
Cleft	→	[*CP* +Exp BE]
rCP	→	[*CP* +F +FIN (+S) +T V]

Operations

Merge = ([*phH* +f (+f$_n$) (H)], [+f L]) = [*phH* [+f L (+f$_n$) (H)]]
Phase Projection = [*phH* **=phX** H] = [*phH* ~~**=phX**~~ H [***phX***]]
Move = if expected [*phX* +f X] and found
[*phX* [*phY* +f +g Y] X] → MEM([*phY* +g <Y>)])

As in MGs (Stabler 1997), the *Lexicon* is a finite set of lexical items storing phonetic, semantic (here ignored) and syntactic features (functional *+F*, selectional *=S*, categorial *C*); an item bearing a selection feature, e.g. [=XP A], requires an XP ph(r)ase right afterward: [~~=XP~~ A [XP]] (once features are projected in the structure, i.e. [XP], the selection features are deleted, i.e. ~~=XP~~); functional features, e.g. +X express a functional specification like determiner +D, tense +T or topic +S (when placed under brackets, e.g. (+f), functional features are optional; ø indicates phonetically null items).

Merge simply unifies the expected structure built so far with a new incoming item, if and only if, this item bears (at least) the first relevant feature expected (Merge operation is greedy: an item bearing more features in the correct expected order will lexicalize them all):

1. *Merge*([+X +Y +Z W], [+X +Y A])=[[+X +Y A] +Z W]
2. *Merge*([[+X +Y A] +Z W], [+Z B])=[[+X +Y A][+Z B] W]
3. Merge([[+X +Y A][+Z B] W], [W C])=[[+X +Y A][+Z B] [W C]]

Move uses a Last-In-First-Out (LIFO) memory buffer (M) to create non-local dependencies: M is used to store unexpected bundles of features merged in the derivation (below, underlined features, e.g. [+W U], are the unexcepted ones triggering Move):

1'. *Merge*([+X +Y +Z W], [+X +W U A]) = [[+X +W U A] +Z W]
2'. *Move*([+X +W U A]) = *M*[+W U <A>]

Items in the memory buffer M will be re-merged in the structure, before any other item taken from the lexicon, as soon as a coherent selection is introduced by another merged item:

3'. *Merge*([… [W ~~=[+W U]~~ C [+W U]]], *M*[+W U <A>]) =
[… [W ~~=[+W U]~~ C [+W U <A>)]]], *M*[*empty*]

Notice that phonetic features (items under angled brackets, i.e. [<A>]) are not re-merged in the structure (that is, they are not expected to be found in the input) since they are already been pronounced/parsed in the higher position. When the M(emory) buffer is empty and no more selection features must be expanded, the procedure ends.

3.1 Parsing cleft structures with PMGs

The parsing algorithm using the minimalist grammar described in (3) implements an Earley-like procedure composed of three sub-routines:

1. *Ph(ase)P(rojection)* (Earley *Prediction* procedure): the most prominent (i.e. first/left most) select feature is expanded (the sentence parsing starts with a default PhP using one of the *phases* in grammar (3));
2. *Merge* (Earley *Scanning* procedure): if Memory is empty, the first available feature *F* in the expected phase is searched in the input/lexicon and possible items will be retrieved[2] (*search*(F) = [F lex$_1$], [F lex$_2$] … [F lex$_n$]) then unified with the expected structure (e.g. *Merge*([F … X], [F lex$_1$]) = [[F lex$_1$]… X]); items stored in Memory are checked before the sentence input for Merge;
3. *Move*: if more features than the one expected are introduced, those features are clustered and moved in the LIFO Memory buffer:
M[[*slot 1*][*slot 2*] … [*slot n*]].

Given the recursive, cyclic, application of the three subroutines above, this is the sequence of steps needed for parsing a cleft sentence like (2):

1. Default PhP (in this case: *Cleft*): [*CP* +Exp BE]
2. Search(+Exp): *M*[*empty*], Lex[[+exp it]]
3. Merge([*CP* +Exp BE], [+exp it]) = [*CP* [+exp it] BE]
4. Search(BE): *M*[*empty*], Lex[[BE is]]
5. Merge([*CP* [+exp it] BE], [=rCP BE is]) =
   [*CP* [+exp it] [=rCP BE is]]
6. PhP([*CP* [+exp it] [**=rCP** BE is]) =
   [*CP* [+exp it] [~~=rCP~~ BE is [*CP* +F +FIN +S +T V]]
7. Search(+F): *M*[*empty*], Lex[[*DP* [+F Ø] +D N]]
8. Merge([…[*CP* +F +FIN +S +T V]], [*DP* [+F Ø] +D N]) =
   [*CP* [*DP* [+F Ø] +D N] +FIN +S +T V]]
9. Search(+D): *M*[*empty*], Lex[[+D the]]
10. Merge([*DP* [+F Ø] +D N], [+D the]) =
    [*CP* [*DP* [+F Ø] [+D the] N] +FIN +S +T V]]
11. Search(N): *M*[*empty*], Lex[[N banker]]
12. Merge([*DP* [+F Ø] [+D the] N], [N banker]) =
    [*CP* [*DP* [+F Ø] [+D the] [N banker]] +FIN +S +T V]]
13. Move([*DP* [+F Ø] [+D the] N], [N banker]) =
    M[[*DP* +D N <the banker>]]
    (Move is triggered because at step 8 +D N were unexpected; only after full lexicalization [*DP* [+F Ø] +D N] is stored in M, namely at step 13)
14. Search(+FIN): *M*[[*DP* +D N <the banker>]], Lex[[+FIN that]]
15. Merge([*CP* [*DP* [+F Ø] [+D the] [N banker]] +FIN +S +T V]], [+FIN that]) = [*CP* [*DP* [+F Ø] [+D the] [N banker]] [+FIN that] +S +T V]]
16. Search(+S): *M*[[*DP* +D N <the banker>]], Lex[[*DP* [+S Ø] +D N]]
17. Merge([*CP* [*DP* [+F Ø] [+D the] [N banker]] [+FIN that] +S +T V]], [*DP* [+S Ø] +D N]) = [*CP* [*DP* [+F Ø] [+D the] [N banker]] [+FIN that] [*DP* [+S Ø] +D N] +T V]]
18. (repeat 9-13 mutatis mutandis)
19. Search(+T): *M*[[*DP* +D N (the lawyer)],[*DP* +D N (the banker)]], Lex[+T will]
20. Merge([*CP* [*DP* [+F Ø] [+D the] [N banker]] [+FIN that] [DP [+S Ø] [+D the] [N lawyer]] +T V]], [+T will]) =
    ([*CP* [*DP* [+F Ø] [+D the] [N banker]] [+FIN that] [DP [+S Ø] [+D the] [N lawyer]] [+T will] V]]
21. Search(V): *M*[[*DP* +D N (the lawyer)],[*DP* +D N (the banker)]], Lex[=DP =DP V meet]
22. Merge([*CP* [*DP* [+F Ø] [+D the] [N banker]] [+FIN that] [*DP* [+S Ø] [+D the] [N lawyer]] [+T will] V]], [=DP =DP V meet])
    =
    [*CP* [*DP* [+F Ø] [+D the] [N banker]] [+FIN that] [*DP* [+S Ø] [+D the] [N lawyer]] [+T will] [=DP =DP V meet]]
23. PhP([*CP* … [**=DP** =DP V meet]]) = [*CP* … [~~=DP~~ =DP V meet [DP +D N]]]
24. Merge ([*CP* … [~~=DP~~ =DP V meet [*DP* +D N]]], M[[*DP* +D N (the lawyer)]] = ([*CP* … [~~=DP~~ =DP V meet [*DP* +D N (the lawyer)]]]
25. PhP([*CP* … [~~=DP~~ **=DP** V meet]]) = [*CP* … [~~=DP~~ ~~=DP~~ V meet [DP +D N (the lawyer)] [*DP* +D N]]]
26. Merge ([*CP* … [~~=DP~~ ~~=DP~~ V meet … [DP +D N]]], M[[*DP* +D N (the banker)]] = ([*CP* … [~~=DP~~ ~~=DP~~ V meet … [*DP* +D N (the banker)]]]

According to the lexicon and the phase expectations, step 10 and 19 could have found in the input [+D N John], [+D N Dan], [+D +P1 +Pl +case_acc me [N Ø]] or [+D +P2 +Sg +case_nom you [N Ø]], capturing all possible combinations of definite descriptions, correct pronominal DPs and proper nouns. Exactly all the possibilities we want to test.

4 Explaining the "similarity" problem in terms of cue-based feature retrieval

According to Warren & Gibson (2005) revealed reading times (see also Gordon et al. 2004 for very similar results) we can roughly rank on a difficulty scale all the (3x3) tested conditions (D = definite condition, e.g. "the banker", N = nominal condition, e.g. "Dan", P = pronoun condition, e.g. "we"; for instance D-D stands for "it is *the banker* that *the lawyer* will meet…", vs D-P condition "it is *the banker* that *we* will meet…"):

(4) D-D ≥ N-D ≈ N-N ≈ P-D
> D-N ≥ P-N > D-P ≥ N-P ≈ P-P

Building on Gillund & Shiffrin (1984) Search of Associative Memory (SAM) model, and assuming a cue-based retrieval mechanism for items in memory (Van Dyke & McElree 2006), we can define a complexity (*C*) function associated to the features to be retrieved from M (Feature Retrieval

[2] For reason of space, I will not discuss here neither lexical and syntactic ambiguity nor reanalysis (i.e. recovery from wrong expectations); the proposed algorithm here is meant to be a Top-Down complete procedure, that is, all the possible ambiguities will be taken into consideration and stored in the parsing "chart" as in the classic Earley's parser. For ranking of alternatives see Hale (2001).

Cost, *FRC*, Chesi 2016) for each item to be re-merged after the phase projection at verb (*V*):

$$(5)\ \ C_{FRC}(V) = \prod_{i=1}^{m_x} \frac{(1+nF_i)^{m_i}}{(1+dF_i)}$$

In the formula above, *m* is number of items stored in memory at retrieval, *nF* is the number of features characterizing the argument to be retrieved that are non-distinct in memory (i.e. also present in other objects in memory), *dF* is number of distinct cued features (e.g. case features explicitly probed by the verb selection). C_{FRC} will express the cost, in numerical terms, that should fit with the revealed reading time (i.e. higher differences in reading times, higher differences in C_{FRC}). According to the lexicon in (3), the cost for retrieving the correct items in the D-D condition, for instance, is calculated as follows:

1. [$_{\textbf{=[DP (+case\_nom)]}\ \text{=DP(+case\_acc) V}}$ meet] will trigger retrieval of the first item (the last inserted one in the buffer) which is (step 24) the *DP* [$_{\text{+D +Sg N}}$ (the lawyer)]
2. No cued-features are present (the verb selection only asks for an optional nominative case) and the 3 features to be retrieved are in fact shared with the other item in memory ([$_{\text{+D +Sg N}}$ (the banker)])
3. Hence: $\text{C}_{\text{FRC}} = \frac{(1+3)^2}{(1+0)} \text{ x } \frac{(1+0)^1}{(1+0)} = 16$

Notice that retrieving the object when the subject has been removed from memory has a minimal cost since no confounding features are present anymore in memory. As for the other relevant conditions: N-N, as in D-D condition share the same features hence we expect them to have similar cost except for the fact that N feature is not fully lexicalized, but it is a trace of an N-to-D movement (Longobardi 1994). Counting this as 0.5 (further investigation is needed to correctly assign a cost to an emptied lexical position), we obtain 12,25. Same complexity for N-D condition (since the [$_{\text{N}}$] lexical feature in D is compared to the trace [$_{\text{N}\ \_i}$] feature of N counting 0.5). While we would expect slightly smaller cost with the P-D condition (P does have a [$_{\text{N}}$ Ø] empty feature), that is 9, we will correctly predict simpler complexity for retrieving pronouns at the subject position, since they are always bearing person features (which are distinct from default 3rd person of D and N) and they are marked for case (which is cued by the verb, producing the minimal cost in the P-P condition (C_{FRC}= 1) and similar costs in the D-P and N-P conditions (both C_{FRC}=4). Predictions can be further differentiated by adding a cost for encoding the features in the structure (*eF*) which is (to keep the calculation as simple as possible) proportional to the number of lexical features to be encoded once an item is retrieved from memory (the numerator of the C_{FRC} cost function becomes: $(1 + nF_i + eF_i)^{m_i}$). This corresponds to an increase of +1 for D and +0,5 for N at retrieval. The new $C_{FREC}(V)$ in the different conditions becomes:

$$C_{FREC}(V)_{\text{ D-D}} = \frac{(1+3+1)^2}{(1+0)} \text{ x } \frac{(1+0+1)^1}{(1+0)} = 50$$

$$C_{FREC}(V)_{\text{ N-D}} = \frac{(1+3+1)^2}{(1+0)} \text{ x } \frac{(1+0+0{,}5)^1}{(1+0)} = 37{,}5$$

$$C_{FREC}(V)_{\text{ N-N}} = \frac{(1+3+0{,}5)^2}{(1+0)} \text{ x } \frac{(1+0+0{,}5)^1}{(1+0)} = 30{,}37$$

$$C_{FREC}(V)_{\text{ P-D}} = \frac{(1+3+1)^2}{(1+0)} \text{ x } \frac{(1+0+0)^1}{(1+0)} = 25$$

$$C_{FREC}(V)_{\text{ D-N}} = \frac{(1+2+0{,}5)^2}{(1+0)} \text{ x } \frac{(1+0+1)^1}{(1+0)} = 24{,}5$$

$$C_{FREC}(V)_{\text{ P-N}} = \frac{(1+2+0{,}5)^2}{(1+0)} \text{ x } \frac{(1+0+0)^1}{(1+0)} = 12{,}25$$

$$C_{FREC}(V)_{\text{ D-P}} = \frac{(1+1+0)^2}{(1+1)} \text{ x } \frac{(1+0+1)^1}{(1+0)} = 8$$

$$C_{FREC}(V)_{\text{ N-P}} = \frac{(1+1+0)^2}{(1+1)} \text{ x } \frac{(1+0+0{,}5)^1}{(1+0)} = 6$$

$$C_{FREC}(V)_{\text{ P-P}} = \frac{(1+1+0)^2}{(1+1)} \text{ x } \frac{(1+0+0)^1}{(1+0)} = 2$$

Though in some cases *FREC* predicts slightly larger differences (e.g. *D-D* vs *N-D/N-N* condition), it correctly ranks all conditions revealed by the discussed experiment, and it is coherent with specific predictions (e.g. related to feature matching) discussed in literature (Belletti & Rizzi 2013).

5 Conclusion

In this paper I presented an adaptation of Earley's Top-Down parsing algorithm to be used with a simple implementation of a Minimalist Grammar (PMG). The advantages of this approach are both in terms of cognitive plausibility and parsing/performance transparency. From the cognitive plausibility perspective, I showed how a re-orientation of the minimalist structure building operations Merge and Move is sufficient to include such operations directly within a parsing procedure. This is a step toward the "Parser Is the Grammar" (PIG) default hypothesis (Phillips 2006) and a welcome simplification of the linguistic competence description: such a grammar description (i.e. our linguistic competence) is shared both in production (generation) and in comprehension (parsing); this seems trivial from a cognitive perspective (we have a unique Broca's area activated in syntactic processing both in parsing and in generation), but it is far from trivial in computational terms. On the other hand, from the parsing/performance transparency perspective, I presented a complexity metric (FREC), based on cued features stored in memory which better characterize performance in object clefts constructions compared to alternative models: for instance the Depencency Locality Theory (DLT) based on accessibility hierarchy (Gibson 2000) is unable to predict high complexity in N-N condition comparable to N-D or D-D condition, since N should be uniformly more accessible than D, contrary to the facts. The proposed model, obviously should be extended in many respects to capture other critical phenomena (see Lewis & Vasishth 2005) but the first results on specific well-studied constructions, like object clefts, seem very promising.

Reference

Belletti, A., & Rizzi, L. 2013. Intervention in grammar and processing. *From grammar to meaning: The spontaneous logicality of language*, 293-311.

Chesi, C. 2012. *Competence and Computation: toward a processing friendly minimalist Grammar*. Padova: Unipress.

Chesi, C. 2016. *Il processamento in tempo reale delle frasi complesse*. EM Ponti, M. Budassi (eds.), 21-38.

Chomsky, N. 1995. *The Minimalist Program* (Current Studies in Linguistics 28). Cambridge (MA): MIT Press.

Chomsky, N. 2008. On phases. In Robert Freidin, Carlos P. Otero, and Maria Luisa Zubizarreta (eds.). *Foundational issues in linguistic theory*, 133-166.

Chomsky, N. 2013. Problems of projection. *Lingua*, 130, 33-49.

Collins, C., & Stabler, E. 2016. A formalization of minimalist syntax. *Syntax*, 19(1), 43-78.

Earley, J. 1970. An efficient context-free parsing algorithm. *Communications of the Association for Computing Machinery*, 13(2), February.

Fong, S. 2005. Computation with probes and goals. In Di Sciullo, A. M. and R. Delmonte (eds.). *UG and external systems: Language, brain and computation*. 75, 311.

Fong, S. 2011. Minimalist parsing: Simplicity and feature unification. In Proceedings of *Workshop on Language and Recursion*. Mons, Belgium: University of Mons. March.

Gibson, E. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. Image, language, brain, 95-126.

Gillund, G., & Shiffrin, R. M. 1984. A retrieval model for both recognition and recall. Psychological review, 91(1), 1.

Gordon, P. C., Hendrick, R., & Johnson, M. 2001. Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6), 1411.

Gordon, P., Hendrick, R., Johnson, M. 2004. "Effects of noun phrase type on sentence complexity", *Journal of Memory and Language* 51, 97-114.

Hale, J. T. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1-8.

Hale, J. T. 2011. What a rational parser would do. *Cognitive Science*, 35(3), 399-443.

Harkema, H. 2001. Parsing minimalist languages. University of California, Los Angeles.

Kayne, R. S. 1994. *The antisymmetry of syntax*. Cambridge (MA), MIT Press.

Lewis, R. L., & Vasishth, S. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29(3), 375-419.

Longobardi, G. 1994. Reference and proper names: a theory of N-movement in syntax and logical form. *Linguistic Inquiry*, 25, 609–65.

Michaelis, J. 1998. Derivational Minimalism is Mildly Context-Sensitive." In M. Moortgat (ed.), *Logical Aspects of Computational Linguistics*, (LACL '98). Lecture Notes in Artificial Intelligence. Springer Verlag.

Niyogi, S., & Berwick, R. C. 2005. A minimalist implementation of Hale-Keyser incorporation theory. In A. M. Di Sciullo (ed.) *UG and external systems language, brain and computation*, linguistik aktuell/linguistics today, 75, 269-288.

Phillips, C. 1996. *Order and structure*. Doctoral dissertation, Massachusetts Institute of Technology.

Rizzi, L. 2007. On some properties of criterial freezing. *Studies in linguistics*, 1, 145-158.

Stabler, E. 1997. Derivational minimalism. In *International Conference on Logical Aspects of Computational Linguistics* (pp. 68-95). Springer, Berlin, Heidelberg.

Van Dyke, J. A., & McElree, B. 2006. Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2), 157-166.

Warren, T., & Gibson, E. 2005. Effects of NP type in reading cleft sentences in English. *Language and Cognitive Processes*, 20(6), 751-767.

TWITTIRÒ:
a Social Media Corpus with a Multi-layered Annotation for Irony

Alessandra Teresa Cignarella, Cristina Bosco and Viviana Patti
Dipartimento di Informatica, Università degli studi di Torino
alessandra.cignarell@edu.unito.it
{bosco,patti}@di.unito.it

Abstract

English. In this paper we describe our work concerning the application of a multi-layered scheme for the fine-grained annotation of irony (Karoui et al., 2017) on a new Italian social media corpus. In applying the annotation on this corpus containing tweets, i.e. TWITTIRÒ, we outlined both strengths and weaknesses of the scheme when applied on Italian, thus giving further clarity on the future directions that can be followed in the multilingual and cross-language perspective.

Italiano. *In questo articolo descriviamo la creazione di un corpus di testi estratti da social media in italiano e l'applicazione ad esso di uno schema multilivello per l'annotazione a grana fine dell'ironia sviluppato in (Karoui et al., 2017). Nell'applicare l'annotazione a questo corpus composto da messaggi di Twitter, i.e.* TWITTIRÒ*, abbiamo discusso i punti di forza ed i limiti dello schema stesso, in modo da evidenziare le direzioni da seguire in futuro anche in prospettiva multilingue e cross linguistica.*

1 Introduction

The recognition of irony and the identification of pragmatic and linguistic devices that activate it are known as very challenging tasks to be performed by both humans or automatic tools (Mihalcea and Pulman, 2007; Reyes et al., 2010; Kouloumpis et al., 2011; Maynard and Funk, 2011; Reyes et al., 2012; Hernández Farías et al., 2016). Our goal, was to create an annotated Italian corpus through which we could address some issues concerning formalization and automatic detection of irony. This work collocates, therefore, in the context of a multilingual project for studying irony and for developing resources to be exploited in training NLP tools for sentiment analysis.

Providing that irony detection is a field that has been growing very fast in the last few years (Maynard and Greenwood, 2014; Ghosh et al., 2015; Sulis et al., 2016), and also taking into account that generation of irony (whether it is spoken or written) may also depends on the language and culture in which it is expressed, the main aim of this work is that of replying to the following research questions: *Is it possible to formally model irony? If so, how?*

Through the present paper indeed, we aim at contributing to the study of irony not only in Italian, but rather in a multilingual and cross-linguistic perspective. Our hope is that, on the one hand, studying the use of figurative language in Italian social media texts, will help us to better understand the developing of this figure of speech itself -irony- and its relations with humor. On the other hand, the study will lead us to the discovery of features and patterns that can be shared and confronted with similar projects in other languages.

2 Data collection

In this section we describe the methodology applied in the collection of tweets, and the internal structure of the dataset. Our work is part and extends a wider joint project with other research groups working on English and French (Karoui et al., 2017). In the French and English datasets, where the same annotation scheme for irony has been applied, tweets were retrieved by using Twitter APIs and filtered through specific *hashtags* exploited by users to self-mark their ironic intention (#irony, #sarcasm, #sarcastic). Providing that Italian users exploit a series of humorous hashtags, but no long-term single hashtag is established and shared among them, the same procedure could not be applied.

Some corpora from Twitter, where the presence of irony is marked, have been made available for Italian in the last few years, and we extracted from them tweets to be included in TWITTIRÒ according to the distribution presented in Table 1[1].

Corpus	Number of tweets
TW-SPINO	400
SENTIPOLC	600
TW-BS	600
TWITTIRÒ	**1,600**

Table 1: Tweet distribution in TWITTIRÒ

As it is shown in Table 1 the tweets were collected from three different pre-existent datasets.

• **TW-SPINO** is a portion of SENTITUT (Bosco et al., 2013) which contains tweets collected from the satirical blog *Spinoza.it*. The language used is grammatically correct and featured by a high register and style, while the topics are variegate with a clear preference for jokes concerning the world of politics and general news.

1. Pubblicata la classifica mondiale della libertà di stampa. Non possiamo dirvi altro. [giga]
→ *(The world ranking for freedom of printing competition has been published. We cannot say anything else. [giga])*

• **SENTIPOLC** (Basile et al., 2014) contains tweets generated by common users and therefore it is less homogeneous than TW-SPINO, with a frequent use of creative hashtags, mentions, repetitions of laughters. We selected here the political tweets with reference to the government of Monti between 2011 and 2012.

2. Mario Monti? non era il nome di un antipasto? #FullMonti #laresadeiconti #elezioni #308.
→ *(Mario Monti? Wasn't it the name of a starter? #FullMonti #laresadeiconti #elezioni #308.)*

• **TW-BS** (Stranisci et al., 2015; Stranisci et al., 2016) contains tweets on the debate of the reform of Italian School "Buona Scuola".

3. @fattoquotidiano Quest'anno è peggio del solito: oltre all'amianto c'è anche #labuonascuola.
→ *(@fattoquotidiano This year worse than usual: in addition to asbestos there is also #labuonascuola.)*

3 A multi-layered annotation scheme

The main goal of the scheme proposed in (Karoui et al., 2017) is to provide a fine-grained representation of irony and to achieve this goal it includes four different levels of annotation as follows.

LEVEL 1: CLASS. It concerns the classification of tweets into **ironic** or **not ironic**, but it does not apply in principle to our case where the corpus only includes ironic tweets.

LEVEL 2: CONTRADICTION TYPE. As stated from various linguistic theories (Grice, 1975; Sperber and Wilson, 1981; Clark and Gerrig, 1984), irony is often exhibited through the presence of a clash or a contradiction between two elements. In tweets, these elements, henceforth named P1 and P2, can be found both as two lexicalized clues belonging to the internal context, see example below, or can be one in the utterance and the other outside, as part of some pragmatic context external to the tweet.

According to (Karoui et al., 2015), we annotate the contradiction that relies exclusively on the lexical clues internal to the utterance as *explicit*, while the contradiction that combines lexical clues with an additional pragmatic context external to the utterance, as *implicit*.

Explicit contradiction: It can involve a contradiction between proposition P1 and proposition P2 that have e.g. opposite polarities, like in the example below where the opposition is between *liberate* (free) and *processate* (process).

4. [**Liberate**]$_{P1}$ Greta e Vanessa. Saranno [**processate**]$_{P2}$ in Italia. [@maurizioneri79]
→ *(Greta and Vanessa have been [**freed**]$_{P1}$. They will [**undergo trial**]$_{P2}$ in Italy. [@maurizioneri79].)*

Implicit contradiction: The irony occurs because the writer believes that his audience can detect the disparity between P1 and P2 on the basis of contextual knowledge or common background shared with the writer.

5. La [buona scuola e le **sillabe**]$_{P1}$ - http:t.conS42fRjAKp
→ *(The [buona scuola and the **syllables**]$_{P1}$ - http:t.conS42fRjAKp)*[2] 2

LEVEL 3: CATEGORIES. Both forms of contradictions can be expressed through different rhetorical devices, patterns or features that are grouped under different labels.

Analogy: In this category are summoned also other figures of speech that comprehend mechanisms of comparison, such as *simile* and *metaphor*.

[1] A portion of these tweets (400 messages) has already been exploited and analyzed in (Karoui et al., 2017).

[2] The official document that presented the school reform had hyphenation mistakes.

5. Il governo #Monti mi ricorda la corazzata kotiokmin.
→ *(Monti's government reminds me of the Battleship Kotiokmin)*

Hyperbole/exaggeration: It is a figure of speech which consists in expressing an idea or a feeling with an exaggerated way.

6. #M5S #Renzi, se tra un anno non ci saranno 170 mila insegnanti di ruolo in più, te li porto **tutti** a @Palazzo_Chigi #labuonascuola.
→ *(#M5S #Renzi, if in one year at least 170,000 teachers will not be employed, I will bring them* ***all*** *to @Palazzo_Chigi #labuonascuola.)*

Euphemism: It is a figure of speech which is used to reduce the facts of an expression or an idea considered unpleasant in order to soften the reality.

7. Nel 2006 Charlie Hebdo aveva pubblicato delle vignette satiriche su Maometto. Ci hanno messo **un po'** a capirle. [nicodio]
→ *(In 2006 Charlie Hebdo published some satirical comic stips regarding Mohammad. It took them* ***a while*** *to understand them.)*

Rhetorical question: It is a figure of speech in the form of a question asked in order to make a point rather than to elicit an answer.

8. Mario Monti? **non era il nome di un antipasto?** #FullMonti #laresadeiconti #elezioni #308.
→ *(Mario Monti?* ***Wasn't it the name of an appetizer?*** *#FullMonti #laresadeiconti #elezioni #308.)*

Context shift (explicit only): It occurs by the sudden change of the topic/frame in the tweet.

9. @matteorenzi Più che la **#labuonascuola** direi #carascuola visto che ci vogliono più di 800 euro a pischello....quasi quanto **5 kg di gelato**
→ *(More than the* ***#labuonascuola*** *I'd say #carascuola being that more than 800 euros are needed for each kid....almost like* ***5 kilograms of ice-cream****.)*

Register changing: (sub-category of the former) in which the "context shift" is due to a sudden change of linguistic style, exploitation of vulgarities or, on the contrary, a rather pompous style. In Italian tweets, users often recur to the exploitation of dialectal expression:

10. Mario, Monti sulla **#cadrega**.
→ *(Mario, Monti on the* ***#chair.)***

False assertion (implicit only): Indicates that a proposition, fact or an assertion fails to make sense against the reality. The speaker expresses the opposite of what he thinks or something wrong with respect to a context. External knowledge is fundamental to understand the irony (it is, in fact, implicit only).

11. Totoministri per il governo Monti: **Gelmini ai lavori pubblici, farà il tunnel dei neutrini!**
→ *(Footbal pools of ministers for the Monti's government:* ***Gelmini at public works' ministry, she will build the tunnel of neutrinos!)***[3]

Oxymoron/paradox (explicit only): This category is equivalent to the category FALSE ASSERTION except that the contradiction, this time, is explicit.

12. Individuata una mafia tipicamente romana. **Prima di mezzogiorno non prendeva appuntamenti.**
→ *(Identified a typical Rome's mafia.* ***It did not fixed appointments before midday.)***[4]

Other: This last category represents ironic tweets, which can not be classified under one of the other seven previous categories. It can occur in case of humor or situational irony.

13. Sicilia, arriva barcone di migranti e a bordo c'è anche un gatto. Vengono a rubarci i nostri like. [@LughinoViscorto]
→ *(Sicily, a big boat full of refugees arrives. There's also a kitty on board. They come here and steal our likes.)*

LEVEL 4: CLUES. Clues represent words that can help annotators to decide in which category belongs a given ironic tweet, such as **like** for analogy, **very** for hyperbole/exaggeration. Clues include also negation words, emoticons, punctuation marks, interjections, named entity (and mentions). Since the extraction of the information about this level can be done, to a great extent by automatic tools, we did not addressed this specific task by manual annotation.

4 Annotation and Disagreement

Given the complexity of irony attested in literature, it is not surprising that the task of annotating irony often leads to disagreement between annotators, which are connected to their individual experience, sense of humor and situational context (Grice, 1975; Grice, 1978; Sperber and Wilson, 1981; Wilson and Sperber, 2007; Reyes et al., 2010; Fink et al., 2011; Reyes et al., 2012).

In our work, the annotation process involved three people previously trained in similar tasks. Since we are aiming at testing the value of the

[3] Minister Gelmini was never in charge of public work administration. It is a reference to an erroneous statement about neutrinos that the Minister had previously uttered.

[4] It is common knowledge that people from Rome are often late, thus the paradox of creating a criminal organization that is also often late.

annotation scheme, the 1,200 new tweets were tagged by two independent annotators (A1 and A2) and by a third (A3) only where a disagreement is detected between A1 and A2.
According to (Karoui et al., 2017), the annotators were asked to apply the second and third levels of the scheme, thus classifying each tweet as featured by implicit or explicit contradiction and selecting for it a category tag between the eight proposed.

4.1 Disagreement Analysis

The inter-annotator agreement (IAA) between A1 and A2 for the labeling of implicit vs. explicit, calculated with Cohen's coefficient, is $\kappa = 0.41$ (moderate agreement), and the distributions of these labels for each annotator are reported in Table 2. Our data analysis, for the moment, seems to corroborate the results of Karoui et al. (2017) where the annotation for the pair EXPLICIT vs. IMPLICIT, obtained a kappa of *0.65* (substantial agreement).

It is interesting to note that while in French implicit activation is the majority (76.42%), in Italian the majority is represented by the explicit type. This is an important result that shows that annotators are able to identify which are the textual spans that activate the incongruity in ironic tweets, whether explicit or implicit. Further studies are surely needed about the activation type of irony for Italian.

		A2		
		implicit	explicit	TOTAL
A1	implicit	**104**	136	240
	explicit	63	**897**	960
	TOTAL	167	1033	**1200**

Table 2: Inter-annotator agreement on type tags

The IAA regarding category tags is slightly higher, $\kappa = 0.46$ (moderate agreement), as we will examine in detail later. The comparison with the French dataset (Karoui et al., 2017) shows a slightly higher inter-annotator agreement: $\kappa = 0.56$ (still moderate). For the second time a clearer identification of pragmatic devices is encountered in French, overcoming the results obtained between Italian annotators.
It is also interesting to mention that, Karoui et al. (2017) operated some calculations when similar devices were grouped together and the scores showed an increment to $\kappa = 0.60$.

Since our work is mainly focused on category tags, their exploitation and distribution, we will discuss in particular on the tweets where A1 and A2 were in disagreement and the need A3's annotation was required (579 tweets). As support, Table 3 shows the distribution of category tags exploited by A1 and A2.

The analysis of the disagreement detected in this new experimental dataset supports the following ideas. Firstly, observing the tag distribution between A1 and A2, the tag OXYMORON/PARADOX is the more frequently exploited, followed by FALSE ASSERTION (see charts in Fig. 1). Concerning the latter, it is also observed a stronger bias from A1 towards that category tag (15.9%) compared to A2 choices (8.4%).

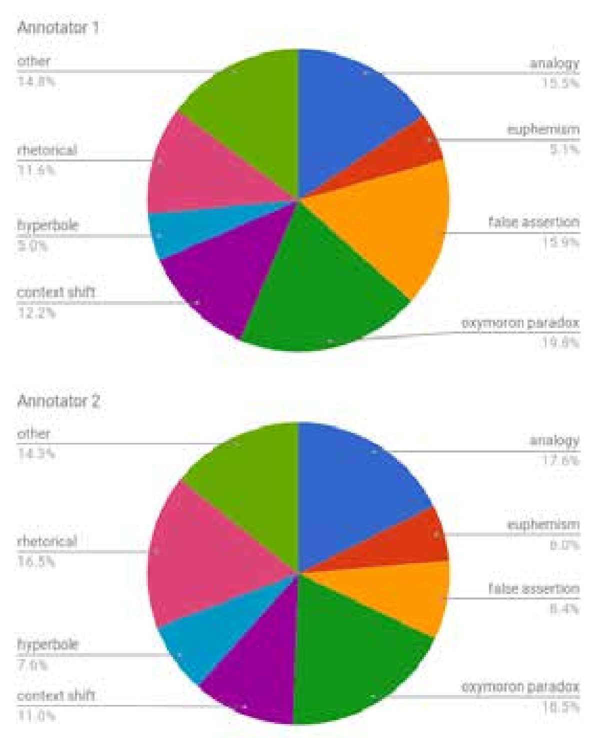


Figure 1: Category tags exploited by the two annotators

The comparison with the annotation results obtained on the French dataset furthermore triggers the need of a deeper research on the application of the scheme in a cross-linguistic perspective.

5 Discussion

Throughout a deeper analysis, the following main issues emerged.

The choice between the category tags OXYMORON/PARADOX and FALSE ASSERTION seems to be strongly influenced by personal biases (see

		A2								
		analogy	euphemism	false assertion	oxymoron paradox	context shift	hyperbole	rhetorical question	other	TOTAL
A1	analogy	131	4	9	13	16	8	7	23	211
	euphemism	4	33	8	7	10	5	1	6	74
	false assertion	6	1	53	21	7	4	0	9	101
	oxymoron paradox	10	8	34	121	21	3	4	21	222
	context shift	9	2	4	31	62	8	2	14	132
	hyperbole	7	4	13	19	4	29	1	14	91
	rhetorical question	8	5	6	25	17	2	127	8	198
	other	19	7	22	10	16	4	3	90	171
	TOTAL	194	64	149	247	153	63	145	185	1200

Table 3: Inter-annotator agreement on category tags

Table 3). In the annotation guidelines it is indeed stated that the labels represent the same category but the former as realized in the context of an explicit contradiction, and the latter when an implicit contradiction happens. For example, in the following tweet A1 tagged as explicit OXYMORON/PARADOX, while A2 as implicit FALSE ASSERTION.

14. Adesso ho capito perché ci son così pochi #presepi in giro. La gente ha paura che il #Governo #Monti faccia pagare l'#ICI anche su quelli...
→ *(Now I get why there are so few Christmas cribs around. People are worried that Monti will put a tax also on them...)*

Another issue we want to address is that of the strong overlapping of RHETORICAL QUESTION with any other tag. As we can see from the following example, it is true that a rhetorical question is made, but the trigger of irony are the paradox and absurdity of the question itself.

15. Ma secondo voi super #Mario #Monti riuscirà a tassare anche la felicità?
→ *(What do you think, will super #Mario #Monti manage to put a tax also on happiness?)*

The problem is caused by the fact that RHETORICAL QUESTION is a category tag that pertains to the linguistic level of pragmatics, which can coexist with semantical or lexical category tags such as ANALOGY or OXYMORON/PARADOX. An improvement in agreement could be that of allowing the presence of one or more categories at the same time.

We have also noticed the exploitation of a common pattern, which we believe should constitute a new category on its own. We named it **false logical conclusion**, most of the time is an EXPLICIT CONTRADICTION, and it expresses which kind of relationship exists between a $P1$ and $P2$. In 45 out of 82 cases, when a false logical conclusion was signaled by at least one annotator (54.88%), the category was tagged as OTHER. We can interpret this as a statistically relevant signal of unsatisfaction of annotators towards the available seven applicable category-tags. Finally, we noticed a high presence of negative words in the whole corpus.

6 Conclusions and future work

The paper describes our work concerning the application of a fine-grained annotation scheme for pragmatic phenomena. In particular, it has been used to annotate the rhetorical device of irony in texts from Twitter. It confirms how this task is challenging, it contributed to shed some light on linguistic phenomena and to significantly extend the resource in (Karoui et al., 2017) with new Italian annotated data to be exploited in future experiments on irony detection in a multi-lingual perspective[5]. The disagreement in the annotation of irony in the three sub-corpora TW-SPINO, SENTIPOLC and TW-BS, which are featured by different characteristics, is a further issue to be addressed. In future work, we plan therefore to investigate the differences in the disagreemente detected across the three portions of TWITTIRÒ providing in-depth analysis of currently available and new linguistic data.

Acknowledgements

The work of Cristina Bosco, Alessandra Teresa Cignarella and Viviana Patti was partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, project S1618_L2_BOSC_01).

References

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the EVALITA 2014 Sentiment Polarity Classification Task. In *Proceedings of the 4th evaluation cam-*

[5] The dataset is available at: `https://github.com/IronyAndTweets/Scheme`

paign of Natural Language Processing and Speech tools for Italian (EVALITA'14).

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Herbert H. Clark and Richard J. Gerrig. 1984. *On the Pretense Theory of Irony.* American Psychological Association.

Clayton R. Fink, Danielle S. Chou, Jonathon J. Kopecky, and Ashley J. Llorens. 2011. Coarse- and Fine-Grained Sentiment Analysis of Social Media Text. *Johns Hopkins APL Technical Digest*, 30(1):22–30.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).*

Paul H. Grice. 1975. Logic and Conversation. *Syntax and Semantics 3: Speech Arts*, pages 41–58.

Paul H. Grice. 1978. Further Notes on Logic and Conversation. *Pragmatics*, 1:13–128.

Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in twitter: The role of affective content. *ACM Trans. Internet Techn.*, 16(3):19:1–19:24.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich Belguith. 2015. Towards a contextual pragmatic model to detect irony in tweets. In *53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015).*

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, and Cristina Bosco. 2017. Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study. In *Proceedings of the European Chapter of the Association for Computational Linguistics.*

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the ICWSM: International AAAI Conference on Web and Social Media.*

Diana Maynard and Adam Funk. 2011. Automatic Detection of Political Opinions in Tweets. In *Proceedings of the ESWC: Extended Semantic Web Conference.*

Diana Maynard and Mark A. Greenwood. 2014. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014).*

Rada Mihalcea and Stephen Pulman. 2007. Characterizing Humour: An Exploration of Features in Humorous Texts. In *International Conference on Intelligent Text Processing and Computational Linguistics.*

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2010. Finding Humour in the Blogosphere: the Role of Wordnet Resources. In *Proceedings of the 5th Global WordNet Conference.*

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering*, 74:1–12.

Dan Sperber and Deirdre Wilson. 1981. Irony and the Use-Mention Distinction. *Philosophy*, 3:143–184.

Marco Stranisci, Cristina Bosco, Viviana Patti, and Delia Irazú Hernández Farías. 2015. Analyzing and Annotating for Sentiment Analysis the Sociopolitical Debate on #labuonascuola. In *Proceedings of the CLiC-it: Italian Conference on Computational Linguistics.*

Marco Stranisci, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. 2016. Annotating sentiment and irony in the online italian political debate on #labuonascuola. In *Proceedings of the Tenth International Conference on Language Resources (LREC 2016).*

Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132 – 143. New Avenues in Knowledge Bases for Natural Language Processing.

Deirdre Wilson and Dan Sperber. 2007. On verbal irony. *Irony in language and thought*, pages 35–56.

Identifying Predictive Features for Textual Genre Classification: the Key Role of Syntax

Andrea Cimino◇, **Martijn Wieling**•,
Felice Dell'Orletta◇, **Simonetta Montemagni**◇, **Giulia Venturi**◇
•University of Groningen - The Netherlands
m.b.wieling@rug.nl
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it

Abstract

English. The paper investigates impact and role of different feature types for the specific task of Automatic Genre Classification with the final aim of identifying the most predictive ones. The goal was pursued by carrying out incremental feature selection through Grafting using different sets of linguistic features. Achieved results for discriminating among four traditional textual genres show the key role played by syntactic features, whose impact turned out to vary across genres.

Italiano. *L'articolo intende indagare il ruolo svolto da diversi tipi di caratteristiche linguistiche nella classificazione automatica del genere testuale al fine di identificare le più efficaci e rilevanti. A questo scopo è stata messa a punto una metodologia basata su un processo incrementale di selezione realizzato mediante un algoritmo di Grafting usando diversi tipi di caratteristiche. I risultati raggiunti mostrano il ruolo chiave delle caratteristiche sintattiche, il cui impatto varia in modo significativo tra generi diversi.*

1 Introduction

Automatic classification of textual genres has always received significant attention from both theoretical and application perspectives. On the one hand, it has been considered relevant by linguists and educators to teach students the correct way of writing in specific communicative scenarios (Biber, 1995; Lee, 2001). On the other hand, the classification of textual genres is seen as a way to cope with the well known problem of information overload: the exploitation of information about document genre can help to develop more accurate Information Retrieval tools. Genre identification has been considered a key factor for reducing irrelevant results of search engines, as users would be able to specify the desired textual genre along with the keywords expressing the content they are looking for (Santini, 2004; Lim, 2004; Santini, 2007). In fact, document genre and document content represent orthogonal dimensions of classifications (Finn, 2003).

A variety of different approaches to Automatic Genre Classification (AGC) has been proposed so far differing at the level of the *genre* and the *typology of features* considered. According to the widely acknowledged fact that no established classification of *genres* exists (see e.g. Sharoff (2010) or Biber (2009)), previous studies focused on 'traditional genres' such as journalism, handbooks, academic prose, see among the others (Kessler, 1997; Stamatatos, 2001; Fang, 2010), and on 'web genres', i.e. genres of web pages, see e.g. (Santini, 2004; Lim, 2004; Mehler, 2010).

Despite the great interest in the investigation of which linguistic features qualify a text genre (Biber, 2009; Fang, 2015), so far little effort has been devoted to use sophisticated NLP techniques, such as syntactic parsing, to capture complex linguistic features for the automatic classification of textual genres. Differently from other application scenarios where the *form* (the style) of a document is investigated, such as e.g. Authorship Attribution (Cranenburgh, 2012), Readability Assessment (Collins, 2014) and Native Language Identification (Tetreault, 2013), AGC approaches proposed so far mainly focus on word level linguistic features, in particular the distribution of function words, word frequency, n–gram models of both characters and Parts–Of-Speech (Santini, 2004; Crossley, 2007; Mehler, 2010) or finer-grained Parts–Of-Speech tags including morpho-syntactic features such as verb tense (Fang, 2010). Very few studies rely on features extracted from syntactically annotated texts, the exception being Stamatatos (2001) who combines lexical features (i.e. word frequency) with features extracted from the output of a chunk boundary detector (e.g. the distribution of noun, verbal, adjectival phrases), the average number of words included in verbal phrases. Similar structural features have been also used by (Lim, 2004) who combined web–specific features (e.g. HTML tags) with lexical information and features aiming at capturing the syntactic structure of a sentence, e.g. the distribution of declarative and imperative sentences, syntactic ambiguities, etc.

In this paper, we tackle the AGC task for traditional genres (namely literary, scientific, educational and journalistic texts) by using different types of linguistic features, i.e. lexical, morpho-syntactic and syntactic. In particular, the following research questions are addressed: i) which are the most effective features to classify a textual genre, and ii) whether and to what extent features identified as most effective remain the same across different genres. These questions have been addressed by carrying out incremental feature selection with the final aim of identifying the most predictive ones. So far, studies focused on the *best set* of features to classify textual genres have been carried out mainly on English. In this paper, this issue is investigated for a typologically different language, Italian.

2 Model training and feature ranking

In order to identify and rank the most important features playing a role in genre classification, we used GRAFTING (Perkins, 2003). This approach allows us to simultaneously train a maximum entropy model while also including incremental feature selection. Grafting uses a gradient-based heuristic to select the most promising feature (which is added to the set of selected features S), and subsequently performs a full weight optimization over all features in S. This process is repeated until a certain stopping condition is reached. The stopping condition integrates l_1 regularization in the grafting approach. This means that only those features are included (with a non-zero weight) if the l_1 penalty is outweighed by the reduction of the objective function. Consequently, overfitting is prevented by excluding noisy features, or those that change value infrequently. In our case, the l_1 penalty was selected on the basis of evaluating maximum entropy models (using 10-fold cross validation) using varying l_1 values (range: 1e-11, 1e-10, ..., 0.1, 1).

For selecting the features and estimating their weights, we used TINYEST[1], a grafting-capable maximum entropy parameter estimator for ranking tasks (De Kok, 2011; De Kok, 2013). Even though our task is not a ranking task, it can be used for binary classification by assigning a high score (1) to the correct class and a low score (0) to the incorrect class. A similar approach was followed by Dell'Orletta (2014) for discriminating between easy–to–read vs difficult–to–read sentences. As the focus of the present study is on the classification of texts belonging to different traditional genres, we created four separate binary classifiers which were trained to distinguish Literature texts from non-Literature (i.e. the three remaining genres) texts, Educational texts from non-Educational texts, etc. A text was assigned the class of the classifier which returned the highest score.

3 Typology of Features

Various types of features have been proposed in the literature for the automatic classification of text genres. Following Stamatatos (2001) and Lim (2004), we combine token–based and structural features. Token–based features were extracted from the top list of the most frequent lemmata in the training corpus and represented in terms of the relative frequency of each lemma in each document. Structural features were extracted from the considered corpora morpho–syntactically tagged by the POS tagger described in (Dell'Orletta, 2009) and dependency–parsed by the DeSR parser using Multi–Layer Perceptron (Attardi, 2009). As shown in Table 1, they range across different linguistic description levels (lexical, morpho–syntactic and syntactic) for a total of 90 features that resulted to be informative “fingerprints” of the form of a text, on issues of e.g. genre, style, authorship or readability.

4 Experiments

4.1 Experimental Setup

We used an Italian corpus including documents representative of four different genres: educational material (Dell'Orletta, 2011), newspaper ar-

[1] http://github.com/danieldk/tinyest

ticles (Marinelli, 2003), literary texts (Marinelli, 2003) and scientific papers (Dell'Orletta, 2014). The whole corpus was split up into a training set (136 documents for the Education genre, 579 for the Journalism genre, 365 for the Literature genre and 317 for the Scientific genre), and a held-out test set (60 documents for each genre).

To assess the influence of including structural features over simply using the most frequent words (lemmata), we used two sets of features (each consisting of about 200 features). The first set of features (taken as the baseline) corresponds to the relative frequency of the 200 top-most frequent words (henceforth referred to as the `tw200` set). [2] The second set combines token-based and structural features: i.e. in addition to the relative frequency of the 100 top-most frequent words, it contains the (90) structural features illustrated above and detailed in Table 1 (this set is henceforth referred to as the `lingtw` set). To guarantee comparability of values, for each feature the values were scaled between 0 and 1 on the basis of the data from the training set. If a (non-scaled) feature value in the held-out test set exceeded the maximum non-scaled value of that feature in the training set, it was set to the maximum value (1).

The feature ranking for each genre was obtained using grafting on the full training data set. The performance (i.e. the percentage of correctly classified documents) of the algorithm was evaluated for an increasing number of features (starting from including only the first (best) feature for each genre to including all features for each genre) against both a 10-fold cross-validation test set and a held-out test set.

The 10-fold cross-validation procedure was performed on the basis of the training set (i.e. the feature weights were determined on the basis of 90% of the training data, whereas the performance was evaluated on the remaining 10% of the training data; this procedure was repeated 10 times). As stated before, the genre of the document in the test set was assigned to the genre whose binary classification model (in this case with the same number of features) resulted in the highest score.

The classification accuracy was assessed with respect to the held-out test set for different numbers of features: *i)* the number of features associated with the best performance on the cross validation set, and *ii)* the lowest number of features such that the performance dropped when a new feature was added (i.e. performance kept increasing for each additional feature up to the selected number of features).

Typology	Feature
Raw Text	Sentence and token length
Lexical	Rate of words in the Basic Italian Vocabulary, Type/Token ratio
Morpho-syntactic	Part-Of-Speech unigrams, Lexical density, Verbal mood
Syntactic	Dependency type unigrams, Parse tree depth features, Arity of verbal predicates, Distribution of subordinate vs main clauses, Length of dependency links

Table 1: Typology of features automatically extracted from linguistically annotated texts.

4.2 Replication

Results reported below can be replicated by downloading the docker image `italianlp-wieling/dockergenreclassification` which contains all data and scripts necessary for the feature extraction and the grafting procedure, and also contains all results. The Docker file including all commands to setup the virtual machine can be found at https://github.com/italianlp-wieling/dockergenreclassification.

5 Results

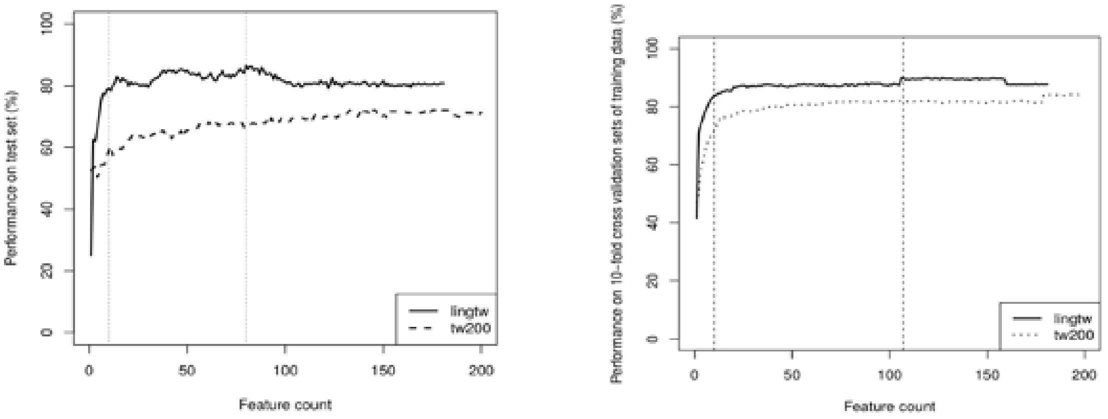
5.1 Genre Classification Results

Figures 1(a) and 1(b) report the classification results using the `lingtw` vs `tw200` features sets: it can be clearly observed that inclusion of structural features is highly beneficial. With only 10 features, the 10-fold cross validation performance is 83.89% for the `lingtw` set, whereas it is only 71.51% for the `tw200` set. The optimal performance for the `lingtw` set is reached with 106 features (89.72%), whereas a significantly lower performance is reached using the `tw200` set (84.17%) despite the much higher number of features used (179). The performance on the held-out test set turned out to be slightly lower: 79.16% for 10 features using the `lingtw` set, against 59.16% with the `tw200` set. The optimal performance for the `lingtw` set is reached with 80 features (86.66%), whereas for the `tw200` set a lower per-

[2] In our preliminary analyses, we also assessed the effect of including the most frequent bigrams as features. However, as the performance was similar to only using unigrams, we did not include bigrams as features.

Genre	First 10 features					First 50 features					First 80 features				
	S	P	L	R	W	S	P	L	R	W	S	P	L	R	W
Journalism	30	40	30	0	0	40	38	10	0	12	38.75	28.75	6.25	1.25	25
Literature	40	30	20	0	10	34	28	8	2	28	33.75	26.25	6.25	1.25	32.50
Education	50	20	10	20	0	44	32	6	4	14	42.5	30	5	2.5	20
Science	60	10	20	10	0	50	32	10	4	4	42.5	30	6.25	2.5	18.75

Table 2: Percentage distribution of different typologies of ranked syntactic (**S**), morpho-syntactic (**P**), lexical (**L**), raw-text (**R**) and token-based (**W**) features selected via GRAFTING on the held-out test set.



(a) Held-out test set.

(b) 10-fold cross-validation test set.

Figure 1: Genre classification results using a held-out test set (a), and a 10-fold cross-validation procedure (b).

formance (72.08%) is obtained using 133 features.

5.2 Feature Ranking Results

In order to investigate the typology of linguistic features most significantly contributing to AGC we focused on the `lingtw` set. In particular, we carried out an in-depth analysis of the grafting-based feature ranking resulting from the classification of the held-out test set. Ranked features were categorized into five classes: syntactic, morpho-syntactic, lexical, raw and token-based features. Figure 2 provides a genre-independent view reporting the percentage average distribution (across genres) of different feature types within the first 10, 50 and 80 ranked feature sets. As shown, *syntactic* features play the most relevant role. They cover the 45% and 42% of the first 10 and 50 features respectively, and remain the most predictive ones also when 80 features are considered (representing 39.38% of the set). On the other hand, the distribution of token-based features increases as far as a wider amount of ranked features is considered (they cover 2.5%, 14.5% and 24.06% in the 10, 50 and 80 feature sets respectively).

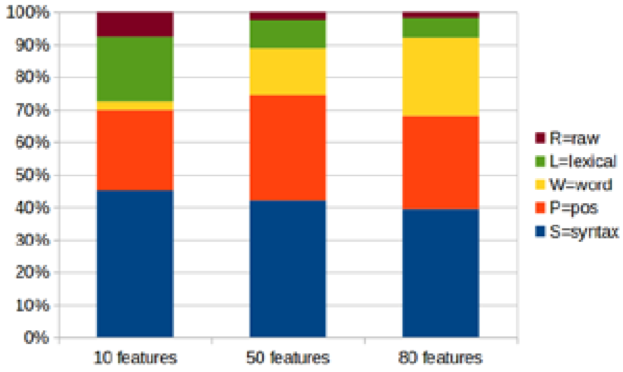


Figure 2: Genre-independent average distribution of different feature types in the top 10, 50 and 80 ranked sets.

Consider now the distribution of different types of features across genres reported in Table 2: notable differences can be observed. In particular, *Literature* and *Scientific prose* represent two opposite poles. Token-based features (*W*) are more predictive for literary texts with respect to other genres (i.e. they represent 10%, 28% and 32.50% in the top 10, 50 and 80 features respectively). On the contrary, syntactic features (*S*) play for *Scientific prose* a more important role than for the other genres (covering respectively 60%, 50% and 42.50% of the top 10, 50 and 80 features).

	Journalism	Literature	Education	Science
Sentence length	–	82	6	32
Word length	56	50	3	3
Type/Token Ratio (forms)	3	95	94	4
Parse tree depth	11	3	37	94
Maximum length of dependency links	42	24	48	56
Post-verbal subject	16	22	90	76
Pre-verbal object	31	21	47	42
Passive subject	7	53	17	5

Table 3: Different ranking positions of a selection of features across genres. Features which were not selected during ranking have no specified rank in the table.

Let's focus now on the role played by individual features across genres. Table 3 reports the different rank positions associated with a selection of features in the classification of the four genres. Raw text features (i.e. sentence and word length) resulted to play a key role in the classification of educational materials (*Education*) with respect to the other genres (e.g. *Literature*). A feature capturing the lexical richness of texts such as Type/Token Ratio (TTR), which refers to the ratio between the number of lexical types and the number of tokens (considered as single forms) within a text, is similarly ranked for *Journalism* and *Science* while it plays a less relevant role in the classification of educational material and literary texts. Moving to syntax, it should be noted that two features characterizing the overall sentence structure, i.e. the depth of the whole parse tree (calculated in terms of the longest path from the root of the dependency tree to some leaf) and the maximum length of dependency links (calculated in terms of the words occurring between the syntactic head and the dependent), play a key role in the classification of the *Literature* and *Journalism* genres. For the latter, it is interesting to contrast the high rank associated with the parse tree depth feature and the irrelevant role played by sentence length (typically taken as a proxy of the underlying grammatical structure): this clearly shows that syntactic features are more effective in discriminating genres. Other features which turned out to play a relevant role in ACG are concerned with the relative ordering of subject and object with respect to the verbal head: their non-canonical orders, i.e. post-verbal subject and pre-verbal object, play a key role in the classification of *Literature* and *Journalism* genres. On the contrary, the use of passive voice (inferred from the presence of passive subjects) is less relevant for the classification of *Literature*, whereas it is highly ranked in the characterization of scientific writing and newspaper articles.

6 Conclusion

In this paper we investigated impact and role of different feature types for Automatic Genre Classification. The goal was pursued by carrying out incremental feature selection through Grafting augmented with TinyEst. Two sets of features were taken into account, token-based and structure-based. Achieved results show the key role played by syntactic features, a result which is new with respect to the AGC literature. Another original contribution is concerned with the role of different feature types which turned out to vary across textual genres, suggesting the specialization of features in binary genre classification tasks (e.g. Literature vs. other genres). The features contributing to AGC for Italian are possibly influenced by the language dealt with. Although it is widely acknowledged that linguistic variation across genres is a language universal, the question is whether similar linguistic features are expected to play a similar role across languages. If this might be the case of features such as e.g. TTR, use of passive voice, tenses or pronouns, on the other hand features concerned with the ordering of sentence constituents or the overall sentence structure (e.g. parse tree depth or dependency length) may be distinctive to a specific language or language family. Further directions of research thus include comparison of results in a multilingual perspective as well as across a wider variety of genres.

Acknowledgements

Reported research started in the framework of a Short Term Mobility program of international exchanges funded by CNR, and continued within the project "Smart News, Social sensing for breakingnews", funded by the Tuscany Region under the FAR-FAS 2014 program.

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi and Joseph Turian 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. *Proceedings of Evalita 2009.*

Douglas Biber. 1995. Dimensions of register variation: A cross-linguistic comparison. *Cambridge University Press* Press, Cambridge, UK.

Douglas Biber and Susan Conrad 2009. Genre, Register, Style. *Cambridge University Press*

Andreas van Cranenburgh 2012. Literary authorship attribution with phrase-structure fragments. *Proceedings of the ACL Workshop on Computational Linguistics for Literature*, 59–63

Kevin Collins-Thompson 2014. Computational assessment of text readability: a survey of current and future research. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, (165–2), 97–135

Scott A. Crossley and Max Louwerse. 2007. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics)*, (12–4) 453–478

Daniël de Kok 2011. Discriminative features in reversible stochastic attribute-value grammars. *Proceedings of the EMNLP Workshop on Language Generation and Evaluation*, 54–63

Daniël de Kok 2013. Reversible Stochastic Attribute-Value Grammars. Rijksuniversiteit Groningen.

Felice Dell'Orletta 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian.*

Felice Dell'Orletta, Simonetta Montemagni, Eva Maria Vecchi and Giulia Venturi. 2011. Tecnologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria. *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, 319–366

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *International Journal of Applied Linguistics*, 165:2, 319–366

Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi and Simonetta Montemagni 2014. Assessing the Readability of Sentences: Which Corpora and Features? *Proceedings of 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014)*, 163–173

Alex Chengyu Fang, and Jing Cao. 2010. Enhanced Genre Classification through Linguistically Fine-Grained POS Tag. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 223–232

Chengyu Alex Fang, and Jing Cao 2015. Text Genres and Registers: The Computation of Linguistic Features. Springer.

Aidan Finn and Nicholas Kushmerick 2003. Learning to classify documents according to genre. *Proceedings of the IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 1–26

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze 1997. Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of ACL (ACL/EACL'97)*, 223–232

David Lee. 2001. Genres, registers, text types, domains, and styles: clarifyng the concepts and navigating a path through the BNC jungle. *Proceedings of ECIR 2004 (26th European Conference on IR Research)*, University of Sunderland (UK), (3), 37–72

Chul Lim, Kong Lee, and Gil Kim. 2004. Multiple sets of features for automatic genre classification of web documents. *Information processing and management* (41) 1263–1276

R. Marinelli, L. Biagini, R. Bindi, S. Goggi, M. Monachini, P. Orsolini, E. Picchi, S. Rossi, N. Calzolari and A. Zampolli. 2003. The italian parole corpus: an overview. *Computational Linguistics in Pisa, Special Issue, XVI-XVII*, 401-421

Alexander Mehler, Serge Sharoff and Marina Santini (Eds.) 2010. Genres on the Web. *Springer Series - Text, Speech and Language Technology*

Simon Perkins, Kevin Lacker and James Theiler 2003. Grafting: Fast, incremental feature selection by gradient descent in function space. *The Journal of Machine Learning Researchs*, (3), 1333–1356

Marina Santini. 2004. Identification of Genres on the Web: a Multi–Faceted Approach. *Language Learning and Technology*, 1–8

Maria Santini 2007. Enhanced Genre Classification through Linguistically Fine-Grained POS Tag. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 223–232

Serge Sharoff 2010. In the garden and in the jungle: Comparing genres in the BNC and internet. in Mehler (2010), 149–166

Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis 2001. Automatic text categorization in terms of genre and author. *Computational Linguistics*, (26) 471–495

Joel Tetreault, Daniel Blanchard and Aoife Cahill 2013. A Report on the First Native Language Identification Shared Task *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 48–57

CorAIt – A Non-native Speech Database for Italian

Claudia Roberta Combei
FiLeLi - University of Pisa
(on leave at FAU Erlangen-Nürnberg)
roberta.combei@fileli.unipi.it

Abstract

English. CorAIt is a non-native speech database for Italian, which is freely accessible online for academic research purposes. It was especially designed to meet the requirements of a larger research project focused on foreign accented Italian speech. The corpus is aimed at providing a uniform collection of speech samples uttered by non-native speakers of Italian. To date, 105 non-native speakers – whose mother tongues are either French, Romanian, Spanish, English, German, or Russian – have been recorded. The corpus includes also a control group made up of 16 Italian speakers. There are almost 8 hours of audio material, both read speech (first and second reading), and spontaneous speech. This paper emphasizes the necessity for this type of database, it describes the steps involved in its construction, and it presents the features of CorAIt.

Italiano. *CorAIt è un corpus audio di l'italiano L2 liberamente consultabile online per scopi di ricerca scientifica. Il corpus è parte integrante di un progetto di ricerca che affronta l'accento straniero nella lingua italiana da una prospettiva più ampia. E' stato ideato e costruito con lo scopo di fornire una raccolta uniforme di materiale audio prodotto da parlanti di italiano L2. Ad oggi sono stati registrati 105 parlanti stranieri di madrelingua: francese, romena, spagnola, inglese, tedesca, e russa. In aggiunta, il corpus è dotato di un gruppo di controllo composto da 16 parlanti italiani. Sono disponibili circa 8 ore di registrazioni, sia di parlato letto (prima e seconda lettura) che di parlato spontaneo. L'articolo evidenzia la necessità di costruire questo tipo di database, e descrive la progettazione e le caratteristiche di CorAIt.*

1 Introduction

It has become clear that accurately designed speech corpora are of essential importance for the development of efficient speech technologies. Investigating how native and foreign-accented speech differ is a necessary step in non-native speech recognition (Tomokiyo, 2001).

Currently, the number of non-native speech databases seems almost insignificant if compared to corpora of native speech.

Moreover, until recently, the majority of the research has focused on English. Therefore, some of the largest non-native speech databases are available for this language: *TED* (Lamel et al., 1994), *Duke-Arslan* (Arslan & Hansen, 1997), *ISLE* (Menzel et al., 2000), *IBM-Fisher* (Fisher et al., 2003), *ATR-Gruhn* (Gruhn et al., 2004), *CSLU* (Lander, 2007), *NATO M-ATC* (Pigeon et al., 2007), and *Speech Accent Archive* (Weinberger, 2015). Large speech corpora of foreign-accented English are owned by Speechocean, and they are specifically built for commercial purposes, especially for training and testing speech recognizers, but some of them are also available for academic research on the KilingLine Data Center platform (Speechocean, 2017).

Only in the last few years has there been an interest in other languages. Without claiming to be exhaustive, some of the largest non-native speech databases for languages other than English will be mentioned: *BAS Strange I+II* (University of Munich, 1998) for German, *WP Russian* (La Rocca & Tomei, 2003) for Russian, *Tokyo-Kikuko* (Nishina, 2004) for Japanese, *TC-STAR* (van den Heuvel et al., 2006) and *WP Spanish* (Morgan, 2006) for Spanish, *SINOD* (Žgank et al., 2006) for Slovenian, and iCALL (Chen et al., 2015) for Chinese.

However, as a result of that fact that many non-native speech databases are built for commercial purposes within private research centres, it is actually quite difficult to map all the resources of this type ever built (Cf. Gruhn et al., 2011, for an overview of the non-native speech databases available at the date their study was published).

This paper presents CorAIt, a non-native speech database for Italian. The database is part of a Ph.D. project which intends to study foreign accented Italian speech both from a computational perspective (automatic identification and classification of non-native accent) and a perceptual perspective (interpretation of quantitative and qualitative judgments delivered by expert and naïve native Italian speakers with respect to non-native pronunciations). The design and the development of this corpus were determined by several factors, which are outlined below.

2 Motivations

Currently, the automatic speech recognition systems for Italian which are integrated into generally available virtual assistant software (e.g. *Google Now*, *Google Assistant*, *Siri*, *Cortana*, etc.) perform quite well on native speech. However, despite recent advances in this field, non-native accents still represent a challenge. This may be due to fact that there is significantly less training data available for automatic speech recognition systems on non-native pronunciations. Considering that Italy is a multicultural country, with over 5 million foreign citizens, representing 8.3% of the entire population residing on its territory[1], it would be desirable to provide services to users who speak Italian with non-native accents.

Apart from acting like training sets for automatic speech recognition systems or for text-to-speech systems, non-native speech databases might be beneficial in the fields of computer assisted language learning (CALL) and mobile-assisted language learning (MALL), as well as for linguistic profiling tasks. Glottologists and scholars working on Italian as a foreign language might also benefit from the presence of these resources.

At the date this research began, there was only one audio corpus for foreign-accented Italian speech, freely available for online consultation, namely *DILS - Dialoghi in Italiano Lingua Straniera* (Savy et al., 2012), consisting of semi-spontaneous audio material obtained by means of the task-oriented dialogue elicitation technique. *DILS* contains 9 large audio samples (for a total duration of 100 minutes) uttered by 18 speakers: 12 Dutch females, 3 Spanish females and 3 Spanish males.

It is worthwhile to mention that there are several other learner corpora for Italian: *VALICO - Varietà Apprendimento Lingua Italiana Corpus Online* (Barbera & Marello, 2004), which is a collection of non-native written Italian; *LIPS - Lessico dell'italiano parlato da stranieri* (Vedovelli et al., 2006); and *Corpus Parlato di Italiano L2* (Spina et al., 2006). The last two corpora consist of transcriptions of audio samples produced by non-native speakers.

In addition to the above-mentioned corpora, there exists a database of written and spoken non-native Italian, entitled *ADIL2 - Archivio Digitale di Italiano L2* (Palermo, 2009*),* which is purchasable in the form of a DVD. However, despite the sophistications of its search tool, the accurate transcription, as well as the admirable amount of data collected, *ADIL2* presents a series of issues that cannot be ignored, such as: imbalance with respect to the speakers' mother tongues (i.e. some languages are underrepresented while others are overrepresented) and elicitation technique used for some samples (i.e. interviews repeated various times over variable time-frames to the same subjects). These aspects render *ADIL2* unsuitable for the type of research to be taken on. Therefore, it became necessary to collect a database of non-native Italian speech.

3 Data Collection

The corpus was designed, collected and developed from January 2016 through July 2017, and it was aimed at providing a uniform collection of audio material produced by adult non-native speakers of Italian residing in Bologna[2].

Initially, the intent was to collect data for 11 different mother tongues (L1s): Maghrebi Arabic, Urdu, Mandarin Chinese, Albanian, Russian, English, German, French, Romanian, Spanish, and Italian (as a control group). The first 10 groups correspond to the L1s spoken by some of the major foreign populations residing in Italy. However, recruiting speakers for all these groups

[1] The data were provided by the National demographic balance (year 2016) produced by the Italian National Institute for Statistics (ISTAT). The full report is available at: http://www.istat.it/it/files/2017/06/bilanciodemografico-2016_13giugno2017.pdf

[2] To simplify the data collection process, the author chose to recruit people that studied or worked in Bologna (her city of residence).

proved to be a challenging task. This may be result of the fact that participation was entirely voluntary and no material reward was provided to informants. Since it was not possible to recruit enough speakers of Maghrebi Arabic, Urdu, Mandarin Chinese and Albanian, these four groups were abandoned.

3.1 Speakers' recruitment

Specific criteria of quality, quantity and diversity were observed, as much as possible, for each L1 when the participants were selected (Cf. section 4.1).

All speakers were recruited locally in Bologna. Most informants were enrolled as regular or exchange students in B.A., M.A. and Ph.D. programmes at the University of Bologna and they were contacted on their personal e-mail address. The e-mail message contained a description of the research project and informed potential participants about the tasks they would have performed. Nearly one fourth of them replied positively to the call.

3.2 Experimental protocol

All informants were aware that they were recorded. They gave informed consent in writing to the use of their speech samples and their sociolinguistic data for research purposes.

In order to guarantee uniformity, the same experimental protocol was employed for all subjects. Before each recording session, speakers were asked to fill in a detailed form regarding their sociocultural and sociolinguistic background. The digital recordings were performed with a Samson METEOR MIC cardioid pickup microphone (condenser diaphragms: 25 mm) on the Praat software (Boersma & Weenink, 2017). The sampling parameters were the following: mono channel, 16-bit, 44,100 Hz, linearly encoded WAV.

Each recording session lasted around 60 minutes. The sessions were individual-based, and they were guided and monitored by the author.

3.3 Speech modalities

The speakers were asked to perform two tasks: reading an article excerpt published on the Italian newspaper *Corriere della Sera*[3]; and describing spontaneously how they spent their last holidays. That specific reading fragment was chosen because it presented various levels of complexity and it contained all Italian phonemes. The reading task was necessary for triggering difficulties that could emerge as a result of conflicting orthographic conventions between the speakers' mother tongues and Italian (Wottawa & Adda-Decker, 2016). Moreover, it could allow speakers comparisons and analyses on the same type of material.

All participants had two reading attempts and they were asked to read and speak as naturally as they could.

4 Description of CorAIt

CorAIt is a non-native speech database for Italian, which has become fully and freely available online for academic research consultation[4]. It contains 2,244 audio samples produced by 105 non-native speakers of Italian. It also includes 300 audio samples obtained from 16 native Italian speakers. In total there are almost 8 hours of speech, consisting in roughly 72,000 words.

4.1 Speakers' statistics

Originally, it was planned to recruit at least 15 speakers for each L1. This threshold was reached for French, and it was exceeded for the other six groups.

Regarding the age distribution of the informants, the range is 19-40 years, but most speakers are older than 20 and younger than 30 years (Cf. Table 1).

Mother tongue	Number of speakers	Age means (±S.D.)
Russian	17	27.06 (5.94)
English	16	23.75 (4.69)
German	17	23.53 (4.06)
French	15	23.33 (2.72)
Romanian	20	25.40 (2.22)
Spanish	20	23.65 (2.95)
Italian	16	29.00 (4.62)

Table 1: Speakers' basic statistics.

Despite the efforts to call up the same number of male and female speakers, the corpus is not perfectly gender-balanced. Apparently, it was

[3] The newspaper article is available online at: http://cinquantamila.corriere.it/storyTellerArticolo.php?storyId=0000002228555. The excerpt was included in this frame: "Don Geretti è un grande affabulatore […] Pietro terminò il suo cammino terreno e quello, tormentatissimo, verso la fede."

[4] CorAIt database is accessible for consultation (prior to registration) at: www.corait.it

easier to engage female participants. In fact, 80 females (66%) and 41 males (34%) were recorded for this project. Nonetheless, according to literature, gender has not been reported as a major source of pronunciation issues, and for this reason it was assumed that gender imbalance had minor effects on this type of studies (Gruhn et al., 2011).

For the corpus design, the age of Italian language onset was taken into consideration, and it was distributed as follows: childhood (23%), adolescence (26%), and adulthood (51%). The data are predictable, considering that most informants learnt Italian naturalistically (64%) soon after they moved to Italy. In fact, only 36% of them claimed that they used mainly scholastic methods for learning Italian, and that they had already spoken the language at their arrival in Italy.

Since most informants were exchange students, 59% of them had spent 6-12 months in Italy at the time they were recruited for this research. The remaining part had lived in Italy for 12-24 months (15%), or for more than 24 months (26%). Not surprisingly, the great majority of speakers claimed that they had been exposed only to the Bolognese variety of Italian.

Because it was almost impossible to predict the speakers' proficiency level[5] in Italian before meeting them, the balancedness is not guaranteed for all accent groups (e.g. no Romanian speaker had A2 waystage/elementary level in Italian). For the sake of brevity, at the general level, this variable is represented as follows in the database: waystage/elementary level - A2 (12%), threshold/intermediate level - B1 (28%), vantage/upper-intermediate level - B2 (28%), and advanced/proficiency levels - C1 and C2 (32%).

4.2 Speech samples

An average of 4 minutes and 25 seconds of raw audio material consisting of read and spontaneous speech were recorded for each speaker. Some speakers had to terminate the registration session earlier than planned, so in those cases it was possible to record only their first reading attempt. Regardless of that, the spontaneous speech collected (23%) is, however, inferior to the reading speech material (77%). All raw samples were segmented manually into utterances corresponding to grammatical sentences for the reading material, and to phonological sentences for the spontaneous speech. In most cases, the material was not qualitatively altered, so hesitation phenomena and disfluencies were generally left as they were.

4.3 Webapp architecture[6]

One contribution of this project is that of making the corpus available to the research community. Following the model of similar tools, a website that would host the database was created. Then, the embedded webapp could extrapolate and classify the audio files from the dataset, according to specific criteria.

For the creation of the webapp, the web framework *Django* as well as several *Python* libraries (*MySQL-python*, *django-treebeard*, *django-filer*, *html5lib*, *sorl*, *wsgi*, *polymorphic*, *classy-tags*, *audiofield*, *appconf*, etc.) were employed. That allowed the use of a powerful *ORM* system, equipped with a web interface for storing multiple data types into our *MySQL* database.

Moreover, the *Django* web framework *ORM* favoured the realization of data collection models: a model is the only final data source containing the fields and the essential behaviours of the dataset and of the reference objects.

Generally, each model is mapped to a single database table and each attribute represents a database field. The queries are performed by means of ad-hoc *APIs* for each model.

The project is hosted on a server with a *CentOS 7* operating system. CortAIt is already configured for various types of *SQL* and *NoSQL* databases (*PostgreSQL*, *MongoDB*, *Cassandra*, etc.). It also supports the execution of some cloud computing platforms, such as *Amazon Web Services* (*AWS*), which could improve its performance in case of an exponential growth of the computational complexity.

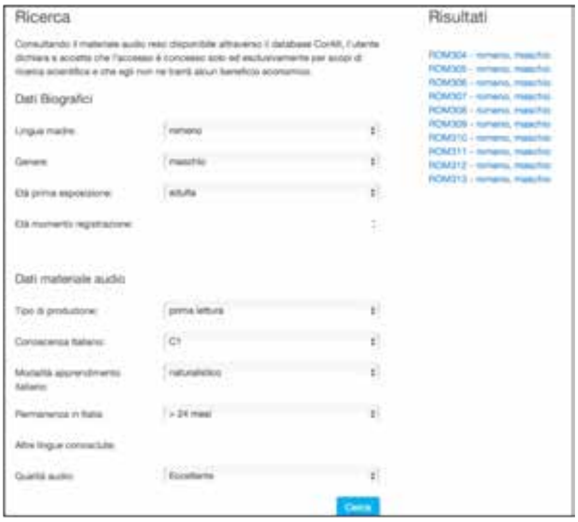
4.4 Front-end presentation

The web database is queryable from the dedicated section of CorAIt website, prior to registration and approval. Due to storage issues, and observing the design of *Speech Accent Archive* (Weinberger, 2015), the format of the audio files available on the online version of CorAIt is .mp3. Samples coded in other formats (e.g. .wav, .flac, etc.) are freely available under request.

[5] All participants self-assessed their Italian level based on the Common European Framework of Reference for Languages, available at: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

[6] The section 4.3 was written with the contribution of Antonio Maria Tenace, who provided support on the graphical implementation of the webapp and was in charge with the technical aspects of its architecture.

To enable advanced queries, various layers of metadata were added to each audio file: the speaker's mother tongue, gender, age of Italian language onset, age at the time the sample was recorded, level of Italian proficiency, Italian learning method, length of residence in Italy, proficiency in other foreign languages. Moreover, information on the type of sample and its quality was included (Cf. Figure 1).



Ricerca

Consultando il materiale audio reso disponibile attraverso il database CorAIt, l'utente dichiara e accetta che l'accesso è concesso solo ed esclusivamente per scopi di ricerca scientifica e che egli non ne trarrà alcun beneficio economico.

Dati Biografici

Lingua madre: rumeno

Genere: maschio

Età prima esposizione: adulta

Età momento registrazione:

Dati materiale audio

Tipo di produzione: prima lettura

Conoscenza italiano: C1

Modalità apprendimento italiano: naturalistico

Permanenza in Italia: > 24 mesi

Altre lingue conosciute:

Qualità audio: Eccellente

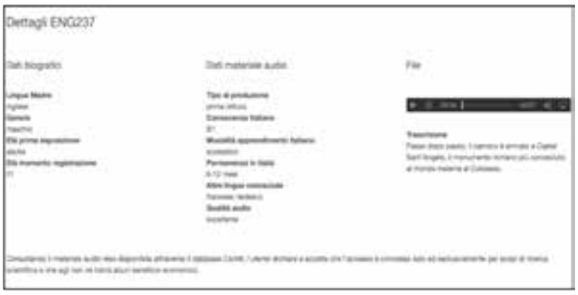
Risultati

ROM304 - rumeno, maschio
ROM305 - rumeno, maschio
ROM306 - rumeno, maschio
ROM307 - rumeno, maschio
ROM308 - rumeno, maschio
ROM309 - rumeno, maschio
ROM310 - rumeno, maschio
ROM311 - rumeno, maschio
ROM312 - rumeno, maschio
ROM313 - rumeno, maschio

Figure 1: Search tool.

The corpus has not been transcribed nor annotated yet. However, following the example of the *Speech Accent Archive* (Weinberger, 2015), the sole grammatical sentences from the reading excerpts were inserted under the samples corresponding to the reading task.

Besides the embedded audio player – which allows to listen and download the audio sample – the window where the single result is displayed provides biographical and quantitative information with respect to the speaker who uttered that speech sample, as well as qualitative information regarding the audio file (Cf. Figure 2).



Dettagli ENG237

Dati biografici

Lingua Madre
inglese
Genere
maschio
Età prima esposizione
adulta
Età momento registrazione
[illegible]

Dati materiale audio

Tipo di produzione
prima lettura
Conoscenza italiano
B1
Modalità apprendimento italiano
scolastico
Permanenza in Italia
6-12 mesi
Altre lingue conosciute
francese, tedesco
Qualità audio
eccellente

File

Trascrizione
[illegible]

Consultando il materiale audio reso disponibile attraverso il database CorAIt, l'utente dichiara e accetta che l'accesso è concesso solo ed esclusivamente per scopi di ricerca scientifica e che egli non ne trarrà alcun beneficio economico.

Figure 2: Results window.

5 Conclusion and future work

Considering that currently this non-native speech database presents some imbalance issues as regards the speakers' age of Italian language onset, their proficiency level, as well as the length of their residence in Italy, the data collection will be further extended.

In the future, the web database might also be enhanced with orthographic and phonetic transcriptions. Disfluencies (i.e. false starts, filled and silent pauses, phoneme lengthening, mispronounced words), mouth clicks, and external noise could be annotated.

6 Acknowledgements

The author gratefully acknowledges the informants who participated in this research, and A. M. Tenace, whose expertise and assistance were fundamental for the architecture and the implementation of the webapp.

Reference

L. M. Arslan & J. H. Hansen. 1997. Frequency characteristics of foreign accented speech. In *Proc. of ICASSP*, pp. 1123-1126, Munich, Germany.

E. Atagi & T. Bent. 2013. Auditory free classification of non-native speech. In *J. Phon.*, 41(6).

M. Barbera & C. Marello. 2004, VALICO (Varietà di Apprendimento della Lingua Italiana Corpus Online): una presentazione. In *ITALS 4*, Guerra Edizioni, Perugia, Italy.

P. Boersma & D. Weenink. 2017. Praat: doing phonetics by computer [Computer program]. Version 6.0.33. [http://www.praat.org]

N. F. Chen et al. 2015. iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent. In *Proc. of Interspeech*, pp. 801-805, Dresden, Germany.

C. Cieri et al. 2004. The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In *Proc. of LREC*, pp. 69-71, Lisbon, Portugal.

T. Cincarek et al. 2004. Speech Recognition for Multiple Non-Accent Groups with Speaker-Group-Dependent Acoustic Models. In *Proc. of Interspeech*, pp. 1509-1512, Jeju Island, Korea.

C. Cucchiarini & H. Van Hamme. 2013. The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People. In P. Spyns & J. Odijk (editors), *Essential Speech and Language Technology for Dutch*, pp. 43–59. Springer, Heidelberg, Germany.

J. Durand et al. (editors). 2014. *The Oxford Handbook of Corpus Phonology*. Oxford University Press, Oxford, UK.

V. Fischer et al. 2003. Recent progress in the decoding of non-native speech with multilingual acoustic models. In *Proc. of Eurospeech*, pp. 3105-3108, Geneva, Switzerland.

R. Gruhn et al. 2004. A multi-accent non-native English database. In *Proc. of Acoustical Society of Japan*, Kyoto, Japan.

R. Gruhn et al. 2011. *Statistical Pronunciation Modelling for Non-native Speech Processin*g. Springer, Heidelberg, Germany.

H. Heuvel et al. 2006. TC-STAR: New language resources for ASR and SLT purposes. In *LREC*, pp. 2570-2573, Genoa, Italy.

L. M. Tomokiyo. 2001. *Recognizing non-native speech. Characterizing and adapting to non-native usage in LVCSR*. PhD thesis. Carnegie Mellon University, Pittsburgh, USA.

L. F. Lamel et al. 1994. The translanguage English database TED. In *ICSLP*, Yokohama, Japan.

T. Lander. 2007. *CSLU: Foreign Accented English*. Release 1.2. LDC2007S08. Linguistic Data Consortium, Philadelphia, USA.

A. La Rocca and C. Tomei. 2003. *West point Russian speech corpus*. Tech. Rep., LDC, Philadelphia, Pennsylvania, USA.

Ludwig Maximilian University of Munich. 1998. Bavarian Archive for Speech Signals. [http://www.phonetik.uni-muenchen.de/Bas]

W. Menzel et al. 2000. The ISLE corpus of non-native spoken English. In *LREC*, pp. 957-963, Athens, Greece.

W. Minker et al. (editors). 2014. *Spoken dialogue systems. Technology and design*. Springer, Heidelberg, Germany.

J. Morgan. 2006. *West point heroico Spanish speech*. Tech. Rep., LDC, Philadelphia, Pennsylvania, USA.

K. Nishina. 2004. Development of Japanese speech database read by non-native speakers for constructing CALL system. In *ICA*, pp. 561-564, Kyoto, Japan.

M. Palermo. 2009. *Percorsi e strategie di apprendimento dell'italiano lingua seconda: sondaggi su ADIL2*. Guerra Edizioni, Perugia, Italy.

S. Pigeon et al. 2007. Design and characterization of the non-native military air traffic communications database. In *ICSLP*, Antwerp, Belgium.

M. Raab et al. 2007. Non-native speech databases. In *Proc. of IEEE-ASRU*, pp. 413 – 418, Kyoto, Japan.

T. Robinson et al. 1995. WSJCAM0: A British-English speech corpus for large vocabulary continuous speech recognition. In *Proc. of IEEE-ICASSP*, pp. 81-84, Detroit, USA.

R. Savy et al. 2012. *DILS - Dialoghi in Italiano Lingua Straniera*. [http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/794-corpus-dils-dialoghi-in-italiano-lingua-straniera]

Speechocean. 2017. *King-ASR-L-190: Chinese English Speech Recognition Database*. [http://kingline.speechocean.com/exchange.php?id=13873&act=view]

S. Spina et al. 2006. *Corpus Parlato di Italiano L2.* [http://elearning.unistrapg.it/osservatorio/corpus/frames-cqp.html]

M. Vedovelli. 2006, LIPS - Lessico di frequenza dell'Italiano Parlato dagli Stranieri. In C. Bardel, J. Nystedt (editors), *Progetto Dizionario Italiano-Svedese*, pp-55-58. Acta Universitatis Stockholmiensis 22, Romanica Stockholmiensia, Stockholm, Sweden.

S. Weinberger. 2015. *Speech Accent Archive*. George Mason University. [http://accent.gmu.edu]

J. Wottawa & M. Adda-Decker. 2016. French Learners Audio Corpus of German Speech (FLACGS). In *LREC*, pp. 3215-3219, Portorož, Slovenia.

A. Žgank et al. 2006. SINOD – Slovenian non-native speech database. In *LREC*, pp. 1620-1623, Genoa, Italy.

Dealing with Italian Adjectives in Noun Phrase: a study oriented to Natural Language Generation

Giorgia Conte
Dipartimento Studi Umanistici
Università di Torino
giorgiaconte.gc@gmail.com

Cristina Bosco
Dipartimento di Informatica
Università di Torino
boscodi.unito.it

Alessandro Mazzei
Dipartimento di Informatica
Università di Torino
mazzei@di.unito.it

Abstract

English. This paper describes a theoretical and empirical investigation about the position of adjectives in the Italian language. The long term goal which oriented the study is the formalization of this information into a natural language generation system. Providing that adjectives mainly occur within noun phrases, we focused on them and we collected data from corpora representing very different text genres, i.e. social media and standard ones, in order to compare the theoretical predictions with the real use of the adjective in Italian. The results obtained by confirm the previsions of the modern linguistic theories but also show the different behaviour of adjectives in the distinct analysed genres.

Italiano. *Questo lavoro presenta un'analisi teorica ed empirica sulla posizione degli aggettivi nella lingua Italiana. L'orientamento del lavoro è dato dalla necessità di formalizzare questa informazione nell'ambito di un sistema di generazione automatica della linguaggio. Poiché gli aggettivi si presentano principalmente nei sintagmi nominali, ci si è concentrati su questi, raccogliendo dati da corpora che rappresentano generi di testo diversi, ovvero social media e standard, al fine di confrontare le previsioni teoriche con l'uso reale dell'aggettivo in Italiano. I risultati ottenuti confermano le previsioni delle moderne teorie linguistiche ma mostrano anche il diverso comportamento degli aggettivi nei diversi generi analizzati.*

1 Introduction

Corpus linguistics is a methodological approach based on the extraction from a set of texts of data useful for the study of language. Even if in principle any collection of texts can be called *corpus*, the term assumes a more precise connotation in the context of modern linguistics, where a corpus is featured by sampling, representativeness, finite size, machine-readable form and a standard reference (McEnery and Wilson, 2001).

In this work we have applied a corpus-based approach and we considered two different corpora which represent two different text genres: one concerning social media language (PoSTWITA corpus) and one concerning balanced standard Italian (UD-it corpus). Indeed, while social media texts have recently gained great attention from the NLP community since they have many peculiar properties, standard texts can give a more accurate view on the status of some linguistic notions in "traditional" written text.

These above mentioned corpora allowed us an in depth investigation about the position of the adjective in the nominal phrase. Indeed, even if this grammatical category is described in several traditional Italian grammars (Renzi et al., 2001; Serianni, 2006; Patota, 2006), its theoretical status is not currently enough formalized to be used within the computational context. A more useful perspective on the behaviour of the adjective is proposed in a recent theoretical study which is focussed on the position of the adjective in Romance languages (Giusti, 2016).

This work aims at achieving two major goals. The first is to empirically confirm with the analysis of corpora the theoretical predictions given in (Giusti, 2016). The second goal is instead to provide a representation and classification of Italian adjective category that can be spent within the SimpleNLG-IT (Mazzei et al., 2016), a surface re-

alizer for Italian language.
The paper is organized as follows: in Section 2 we review the linguistic literature concerning the position of the adjective within the Italian noun phrase. In Section 3, we explain the details of our corpus linguistic investigation. In Section 4, we describe the use of the empirical data in the SimpleNLG-IT realizer. Finally, the Section 5 closes the paper with conclusions and some pointers to future work.

2 The Theoretical Status of the Adjective in the Nominal Phrase

We take into account the adjective in its primary use (Bhat, 1994), that is as modifier of a noun. In Italian, within the nominal phrase, the adjective can be positioned before or after the noun to which it refers. In accordance with the traditional grammar, e.g. (Serianni, 2006), these alternative positions are described as unmarked, when the adjective follows the noun, and marked, when it precedes the noun.

These different behaviour of the adjective also carry different semantic values: nominal phrases where the adjective precedes the noun indicate more subjectivity or more stylistic refinement if compared to the more neutral and objective expressions where the adjective follows the noun, as in the following examples (extracted from (Serianni, 2006)): *gli occhi neri* (the eyes black) and *gli alberi alti* (the trees high) vs. *i neri occhi* (the black eyes) and *gli alti alberi* (the high trees)[1]. In the left side of the versus, the adjectives *neri* (black) and *alti* (high) objectively qualify the nouns that they follow, and the information they carry is indeed verifiable by a true/false criterion; in the other side instead the same adjectives qualify the nouns but they also emphasize a desire for stylistic elaboration by those who write or speak (Serianni, 2006).

Moreover, a descriptive function is usually attributed in literature to pre-nominal adjectives, while a restrictive function is attributed to post-nominal ones, e.g. in (Serianni, 2006). This can be clearly exemplified by the difference between the following sentences: *le vecchie tubature hanno ceduto* (the old pipes has collapsed) and *le tubature vecchie hanno ceduto* (the pipes old has collapsed). In the first sentence, the pre-nominal adjective *vecchie* (old) has a descriptive function: it describes a quality of the related noun, i.e. *tubature* (pipes). Instead in the second sentence, the same adjective, in post-nominal position, has restrictive function with respect to the meaning of the related noun: it adds to the noun a distinctive qualification which identifies it as the only one with a certain quality (the *old* pipes, not the *new* ones) (Serianni, 2006). However the value of the adjective in the post-nominal position, being unmarked, may be ambiguous between these two functions, whereas an adjective in pre-nominal position can only have appositive (that is descriptive) function (Giusti, 2010).

2.1 A hierarchy of the Descriptive Adjectives

In (Giusti, 2010) a further distinction among the descriptive adjectives in sub-categories is provided. It is based on a cross-linguistically defined hierarchy where the rank that the adjective assumes is strictly related to the position that it can assume with respect to the noun. The categories are the following:

- evaluative, e.g. *bello* (beautiful)
- dimension, e.g. *alto* (high)
- age, e.g. *vecchio* (old)
- physical property, e.g. *duro* (hard)
- colour, e.g. *rosso* (red)
- relational, e.g. *nazionale* (national)

The adjectives collocated in the lower part of the hierarchy are more prone to assume post-nominal positions, where those in the higher part more frequently assume the pre-nominal ones. For instance, the relational adjectives, that are at the lower level of the hierarchy, are predominantly post-nominal. The others can be freely positioned before or after the noun, but those occupying lower positions within the hierarchy have a stronger tendency for post-nominal positions, while those in higher part of the hierarchy are more freely placed before or after noun (Giusti, 2016). For more details about the classification of the adjectives and how we applied it to those we extracted from corpora, see the following section.

3 Extracting Adjectives from Corpora

In order to validate the assumptions made in literature, and described in section 2 about the behaviour of the adjective, we selected corpora where Italian is annotated for what concerns morphology and syntax and representing also differ-

[1] The English glosses for the examples are literal and can not correspond to the correct English expressions.

ent text genres. We applied scripts in Python and SQL queries for detecting the presence of adjectives and noun phrases in both the reference corpora, but their classification is manually done, for carefully dealing with cases where ambiguity occurs.

We found a substantial help for finding a decision-making criterion for the classification of adjectives in the examples proposed in the Treccani online vocabulary. For instance, we tagged as evaluative the adjective *pericoloso* (dangerous), which is derived from the noun *pericolo* (danger), according to the vocabulary example *un viaggio pericoloso* (a dangerous journey). We tagged instead as relational the adjective *solare* (solar), like in the example *luce solare* (solar light), considering that the adjective is derived from the noun *sole* (sun), indicating an entity rather than a quality.

A particular attention must be paid to homonymous adjectives, like e.g. *reale* that may mean 'royal' or 'real'. In this case, two different entries in the vocabulary must be introduced, one for each meaning of the adjective: the first tagged as relational, for the meaning derived from the noun *re* (king), and the second tagged as evaluative, for indicating the meaning 'actually existing'.

In the rest of this section the resources we used in our investigation are described also showing the differences that make them especially interesting for validating our results in two different contexts and text domains.

The data sets we used are respectively extracted from two different corpora: PoSTWITA (Bosco et al., 2016) and UD-it[2], both tagged in accordance with the Universal Dependencies annotation scheme [3]. While the PoSTWITA corpus is only morphologically tagged and it is taken from the social network Twitter, the other resource is a treebank which includes other variety of more standard texts.

3.1 PoSTWITA

PoSTWITA characterised by short texts (140 characters maximum) and a typical social media Italian jargon that is featured by a frequent use of creative expressions and incorrect words like in the following example:

ho un disparato bisogno di soffocati di coccole. <3 ti amo piccola mia. ([I] have a desperate need to suffocate you with pampering. <3 [I] love you my baby.)

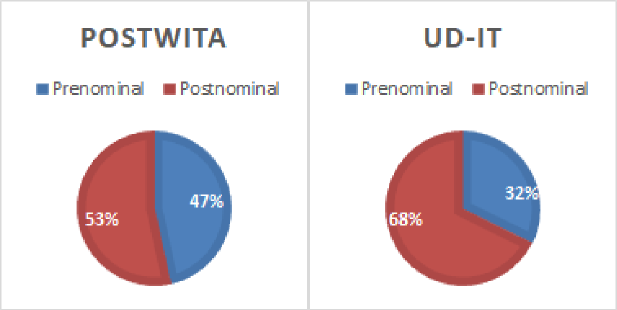


Figure 1: The percentage of pre-nominal and post-nominal adjectives in PoSTWITA and UD-it.

where two incorrect words occur: *disparato* instead of *disperato* and *soffocati* instead of *soffocarti*.

Also distinctive graphic practices due to the particular medium are symbols are very frequent in Twitter posts, like e.g. acronyms and abbreviations and elements without a clearly defined syntactic function like hashtags, mentions and emoticons (Chiusaroli, 2016), whose presence is mainly motivated by communicative goals of the authors, like the following example shows: *"@pari_biosteria Alessandro #Bergonzoni Contro lo #stigma nei confronti della malattia mentale #passaparola http://t.co/daHsNTcBmh"* (@pari_biosteria Alessandro #Bergonzoni Against the #stigma towards the disease mental #passaparola http://t.co/daHsNTcBmh)

where some hashtag is exploited as common noun (*#stigma*), other as proper noun (*#Bergonzoni*) or with a proper communicative function *#passaparola*).

Each word of PoSTWITA is associated with a tag showing its grammatical category selected within the inventory of tags proposed for the part of speech tagging within the Universal Dependency project; only a few tags extends this inventory for better describe typical social media elements, like EMO for emoticons or URL for web addresses.

Within our corpora we focused only on the words tagged as ADJ (adjectives), NOUN (common nouns) and PROPN (proper nouns), that is those involved in the noun phrase structures. Nevertheless, it must be observed that since PoSTWITA corpus is only tagged morphologically, a proper notion of noun phrase is not marked in it. In order to detect adjectives that are syntactically linked to nouns within noun phrases, we considered the

[2] `http://universaldependencies.org/it/overview/introduction.html`

[3] `www.universaldependencies.org`

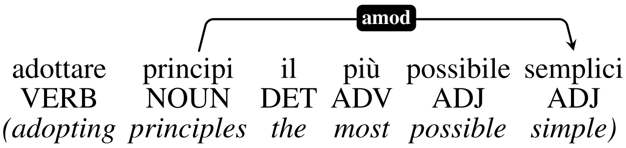
adjectives that were immediately before or after nouns or proper nouns. According to this strategy, the number of adjectives occurring in prenominal position is 1,519, while the number of those in postnominal position is 1,740.

3.2 UD-it

UD-it corpus is tagged both morphologically and syntactically. It is derived from the conversion of different resources developed by Turin and Pisa University's Computer Science Departments and Pisa CNR's Computational Linguistics Institute. This corpus is composed by legal texts (Italian Constitution and part of the Civil Code), Wikipedia and newspaper articles. We can therefore say that, unlike PoSTWITA corpus, UD-it corpus is representative of the so-called Standard Italian, that is encoded, over regional, elaborate, belonging to the upper classes, invariant and written (Berruto, 2010), like the following example shows: *"La prima attività ha lo scopo di creare e sviluppare una rete di ricognizione globale con l'intento di monitorare il rispetto dei trattati internazionali contro la proliferazione di armi di distruzione di massa e la definizione dei confini territoriali."* (The first activity has the objective of creating and developing a network of global reconnoiting with the goal of monitoring the respect of international treatises against the diffusion of the weapons of mass destruction and the definition of territorial borders.)

Providing that UD-it corpus is fully annotated according to the dependency grammar framework of the Universal Dependencies, a notion of noun phrase can be derived from its structures, even if it is not properly annotated, as usual in dependency formats. We considered in this corpus all the adjectives that are related with a noun or a proper noun with the dependency relation *amod*, that is the dependency featuring the adjectival modifiers. Taking into account this relation, we collected 4,469 adjectives occurring in pre-nominal position and 9,362 in the post-nominal one.

It must be observed that the availability of the syntactic annotation of the UD-it corpus has allowed more reliable results with respect to that obtained from PoSTWITA. Indeed we can not be sure that an adjective is related to a specific noun just because it is near that noun, providing that an adjective can refer to a noun even if distant from it, as the following example shows, where an adverbial modifier is collocated between the noun and the adjective that modifies it:



adottare	principi	il	più	possibile	semplici
VERB	NOUN	DET	ADV	ADJ	ADJ
(adopting	*principles*	*the*	*most*	*possible*	*simple)*

3.3 Discussion of Results

The pie charts (Fig. 1) show the data extraction results. The largest percentage of the post-nominal adjectives provides some hints about the markedness of the pre-nominal position for both PoSTWITA and UD-it.

For what concerns the distribution in pre- and post-nominal position of the categories of adjectives described in sec. 2.1, it is represented in the histograms as detected in Figure 3 (PoSTWITA) and Figure 2 (UD-it). We collected these data by applying to our datasets scripts in Python and SQL queries running on a database version of the resources.

The diagrams show how the adjectives in the lower portion of the hierarchy (relational, colour and physical property) are predominantly in post-nominal position within the noun phrase, whereas the adjectives in the higher portion of the hierarchy (age and dimension) are in majority in the pre-nominal one. Evaluative adjectives are the most equally distributed. These results confirm the theoretical tenets presented in the previous part of the paper and collocate the behaviour of the adjective within a context that can be used for modelling in a computational perspective this grammatical category.

4 Ordering adjectives in SimpleNLG-IT

The formalization of linguistic properties is a fundamental process both for NL processing as well as for NL generation systems. In particular, a widespread architecture for NLG assumes a specific module for the linguistic *realization*, that is essentially an algorithmic implementation of a formal grammar (Reiter and Dale, 2000). Recently, as can be read in (Mazzei et al., 2016), a common set of API for the linguistic realization has been adapted also for Italian language. A key component of SimpleNLG-IT is the reference lexicon, i.e. the computational dictionary specifying the computational properties of the words that the realizer can generate (Mazzei et al., 2016). The de-

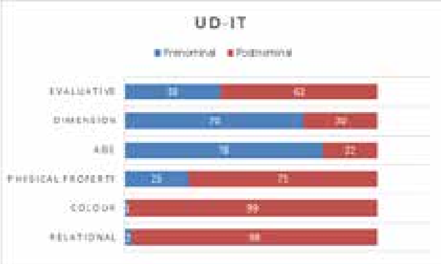


Figure 2: The distribution of the classes of the descriptive adjective in UD-it.

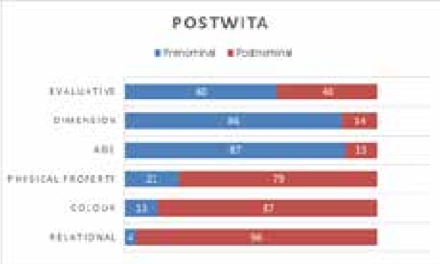


Figure 3: The distribution of the classes of the descriptive adjective in PoSTWITA.

Category	**NVdB/UD-it**
dimension	15/16
age	7/7
physical property	4/4
colour	10/11
relational	111/121
evalutative$^{\mathrm{pre}}$	33/35
evalutative$^{\mathrm{post}}$	61/68

Table 1: The adjectives occurrences in NVdB/UD-it respectively.

fault position for adjective which is assumed in SimpleNLG-IT is the post-nominal one, with the only exception of ordinals adjectives.

Nevertheless, providing that a more correct modelling of the behaviour of words has a positive impact on the human-machine interaction, in SimpleNLG-IT we devised a new version of the lexicon by following the procedure described in (Mazzei, 2016). We started from the newly released *Vocabolario di base della lingua italiana*[4] (NVdB) (Chiari and De Mauro, 2014) which represent the basic lexicon typical of a standard Italian speaker. Moreover, according to (Giusti, 2016), we classified the adjectives as: relational, colour, physical property, age, dimension, evalutative$^{\mathrm{pre}}$ and evalutative$^{\mathrm{post}}$. Indeed, following the data reported in the Figure 2, we formalized that adjective belonging to the relation, colour, physical property sets are generated in prenominal position. In contrast, adjectives belonging to age and dimension classes are generated in post-nominal position. Since evaluative adjectives do not show a clear default position, we further split the set in two different subsets that are generated in pre-/pos-tnominal position respectively. Note that not all the adjectives used for UD-it analysis belong to NVdB, e.g. *maggiore* (greater) or *agrario* (agrarian). Table 1 reports the occurrences of the adjectives in NVdB/UD-it respectively.

All the resource developed are made available on a free access repository[5].

5 Conclusion and future work

The paper presents a study about the behaviour of the adjective within the noun phrase. Providing that the qualitative description given by traditional grammars does not allow the definition of a formal model, we considered a recent study that classifies the descriptive adjectives. The long term goal which oriented this study is to contribute to the development of a natural language generation system for Italian featured by a more careful modelling of the behaviour of words within sentence structures.
Assuming a corpus-based perspective we tested on two corpora for Italian the tenets of this study. The results confirm and validate the theory thus opening the window for a definition of a formal model that can be exploited in our computational framework.

Future work is planned to extend the validation of our model on larger datasets, where a wider variety of adjectives is used and also more complex noun phrase structures are taken into account with respect to the simple <*adjective - noun*>or <*noun - adjective*>associations here considered. In particular, providing that more than one adjective can occurs within a noun phrase and can be syntactically linked to a single noun, we intend to investigate on the preference order also in these cases.

[4] `https://dizionario.internazionale.it/nuovovocabolariodibase`

[5] `https://github.com/alexmazzei/SimpleNLG-IT`

References

G. Berruto. 2010. Italiano standard. www.treccani.it.

D.N.S. Bhat. 1994. *The adjectival category. Criteria for differentiation and identification*. John Benjamins Publishing Company.

Cristina Bosco, Fabio Tamburini, Andrea Bolioli, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 Part Of Speech on Twitter for ITAlian task. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Isabella Chiari and Tullio De Mauro. 2014. The New Basic Vocabulary of Italian as a linguistic resource. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *1th Italian Conference on Computational Linguistics (CLiC-it)*, volume 1, pages 93–97. Pisa University Press, December.

F. Chiusaroli. 2016. Scritture brevi e tendenze della scrittura nella comunicazione di Twitter. In *Linguaggio e apprendimento linguistico. Metodi e strumenti tecnologici*. Officinaventuno.

G. Giusti. 2010. Il sintagma aggettivale. In Giampaolo Salvi and Lorenzo Renzi, editors, *Grammatica dell'italiano antico*. Il Mulino.

G. Giusti. 2016. The structure of the nominal group. In *The Oxford guide to the Romance Languages*. Oxford University Press.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Alessandro Mazzei. 2016. Building a computational lexicon by using SQL. In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*, volume 1749, pages 1–5. CEUR-WS.org.

T. McEnery and A. Wilson. 2001. *Corpus linguistics. An introduction.* Edimburgh University Press.

G. Patota. 2006. *Grammatica di riferimento dell'italiano contemporaneo.* Garzanti Linguistica.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

L. Renzi, G. Salvi, and A. Cardinaletti. 2001. *Grande grammatica italiana di consultazione.* Il Mulino.

L. Serianni. 2006. *Grammatica italiana. Italiano comune e lingua letteraria.* Utet università.

Variabili Rilevanti nella Rappresentazione delle Parole nel Lessico Mentale: Dati Psicolinguistici da una Banca-Dati di Nomi e Verbi Italiani

Maria De Martino, Azzurra Mancuso, Alessandro Laudanna
LaPSUS, Laboratory of Experimental Psychology, University of Salerno
Via Giovanni Paolo II, 132 Fisciano, SA, 84084, Italy
{mdemartino,amancuso,alaudanna}@unisa.it

Abstract

Italiano. Il lavoro descrive un database lessicale composto da 100 nomi e 100 verbi italiani. Per ciascun nome e verbo il database rende disponibili informazioni sulle proprietà formali, distribuzionali, morfo-fonologiche, semantiche e sintattiche e i valori relativi al tempo medio e all'accuratezza di riconoscimento. Il database è utile nelle aree di ricerca in cui sia necessario manipolare e controllare le variabili linguistiche che influenzano il processing lessicale delle parole.

Inglese. *The present paper describes a lexical database containing 100 Italian nouns and 100 Italian verbs. For each noun and verb the database provides information about formal, distributional, semantic, morpho-phonological and syntactic characteristics and average recognition times and accuracy. This set of norms is supposed to be helpful in research areas where it is necessary to manipulate and to control for linguistic variables affecting lexical processing of words.*

1. Introduzione

La ricerca sull'elaborazione lessicale ha dimostrato che numerose proprietà linguistiche delle parole incidono sul modo in cui esse vengono riconosciute e prodotte dai parlanti. D'altra parte, isolare sperimentalmente il ruolo di variabili singole è un'impresa quasi impossibile dato l'effetto di altre variabili concomitanti (Baayen, 2008). In tal senso, la disponibilità di banche-dati descrittive e comportamentali è cruciale in molte aree di ricerca sull'elaborazione di parole.

Questo lavoro ha lo scopo di presentare un database operativo e facile da interrogare composto attualmente[1] da 200 parole italiane. Un aspetto innovativo del database è rappresentato dalla ricchezza di parametri descrittivi per le singole parole e per la presenza di misure di elaborazione delle stesse (tempi medi di risposta e accuratezza in un compito di riconoscimento visivo). Tale ricchezza, soprattutto in termini di informazioni sulle caratteristiche morfologiche, semantiche e sintattiche pressoché assenti in altri strumenti di ricerca disponibili, rende il database particolarmente utile per preparare liste di stimoli adeguatamente bilanciati per esperimenti fattoriali in tutte le aree di ricerca in cui le parole possono essere usate come stimoli. Un secondo aspetto innovativo è rappresentato dal fatto che, a differenza di analoghi strumenti disponibili per l'italiano, la banca-dati qui illustrata propone modalità di analisi linguistica e di classificazione omogenee per le due principali classi di parole, i nomi e i verbi[2].

1.1 Organizzazione della Banca-Dati

La banca-dati contiene 200 entrate principali, 100 sostantivi e 100 verbi italiani, e 400 sotto-entrate: per i sostantivi vi sono 100 sotto-entrate che corrispondono alla forma singolare e 100 sotto-entrate che corrispondono al plurale del nome; per i verbi vi sono 100 sotto-entrate che corrispondono alla forma dell'infinito e 100 sotto-entrate che corrispondono alla 3^a persona dell'indicativo presente; tali forme verbali sono quelle più diffusamente impiegate negli studi psicolinguistici e neurolinguistici.

[1] La banca-dati qui descritta è parte di uno studio più ampio ancora in via di completamento che comprende complessivamente 490 entrate e 1960 sotto-entrate.

[2] La banca-dati è disponibile in formato Microsoft Excel ed è consultabile inviando un messaggio di posta elettronica a mdemartino@unisa.it.

Ciascuna voce inserita nella banca-dati è corredata da informazioni relative ad un ampio numero di variabili descritte nei paragrafi successivi che possono essere classificate in sei categorie: variabili formali e distribuzionali, variabili morfologiche e morfo-fonologiche, variabili sintattiche, variabili semantiche, giudizi su variabili soggettive, misure comportamentali.

2. Variabili Formali e Distribuzionali

Gli effetti delle variabili formali sull'accesso lessicale sono tra i più noti e consistenti nella letteratura psicolinguistica: tipicamente le parole più corte e con maggiore frequenza d'uso tendono ad essere elaborate con maggiore efficienza dal processore lessicale (Colombo, 1993), così come le parole che si conformano al pattern di accentazione prevalente nella lingua rispetto alle parole che deviano da tale pattern (Colombo, 1992). Analogamente, l'elaborazione lessicale di parole che sono altamente confondibili con altre per la presenza di vicini ortografici[3] risente della numerosità e della frequenza cumulativa del vicinato ortografico o di singoli vicini (Laudanna, 2006). Altri dati (Bracco e Laudanna, 2012) mostrano che la frequenza relativa della forma di una parola, ovvero il rapporto quantitativo tra la frequenza individuale di quella forma e la frequenza cumulativa delle altre forme del paradigma di quella parola, ha un valore predittivo sulla velocità e l'accuratezza nel riconoscimento.

La banca-dati qui descritta rende disponibili una serie di valori relativi alle principali variabili formali e distribuzionali indicate dalla ricerca sull'accesso lessicale.

Per ciascuna entrata del database (e relative sotto-entrate) è indicato il pattern di accentazione e sono disponibili informazioni sulla lunghezza espressa in termini di numero di lettere, sillabe e fonemi.

ENTRATA	SOTTO-ENTRATA	ACCENTO	LET	SIL	FON
dedica	dedica	*sdrucciola*	*6*	*3*	*6*
	dediche	*sdrucciola*	*7*	*3*	*6*
abbaiare	abbaiare	*piana*	*8*	*4*	*7*
	abbaia	*piana*	*6*	*3*	*5*

Tabella 1: Esempi di codifica del pattern di accentazione e della lunghezza.

La banca-dati contiene svariati indici sulla frequenza d'uso delle singole entrate e delle relative sotto-entrate. Sono disponibili i valori relativi alla frequenza cumulativa di tutte le forme del paradigma di nomi e verbi nello scritto per il lessico adulto (fonte: CoLFIS, Bertinetto, Burani, Laudanna, Marconi, Ratti, Rolando e Thornton, 2005), alla frequenza cumulativa di tutte le forme del paradigma di nomi e verbi nel parlato per il lessico adulto (fonte: LIP, De Mauro, Mancini, Vedovelli e Voghera, 1993) e alla frequenza cumulativa di tutte le forme del paradigma di nomi e verbi nello scritto per il lessico infantile (Marconi, Ott, Pesenti, Ratti e Tavella, 1993).

Per le sotto-entrate sono disponibili i valori della frequenza della forma nello scritto adulto, i valori della frequenza della forma nel parlato adulto ed il rapporto tra la frequenza della singola forma e la frequenza cumulativa dell'intero paradigma della parola nello scritto e nel parlato adulti.

LESSICO SCRITTO ADULTO				
SOTTO-ENTRATA	VALORE GREZZO	VALORE PER MILIONE	FQ RELATIVA GREZZA	FQ RELATIVA PER MILIONE
assaggiare	*5*	*1,6*	*0,2*	*0,07*
assaggia	*6*	*1,9*	*0,3*	*0,08*
gita	*30*	*9,4*	*0,6*	*0,18*
gite	*23*	*7,2*	*0,4*	*0,14*

Tabella 2: Esempi di codifica dei valori di frequenza nel lessico scritto adulto.

LESSICO PARLATO ADULTO				
SOTTO-ENTRATA	VALORE GREZZO	VALORE PER MILIONE	FQ RELATIVA GREZZA	FQ RELATIVA PER MILIONE
assaggiare	*4*	*8*	*0,5*	*1*
assaggia	*0*	*0*	*0*	*0*
gita	*10*	*20*	*1*	*2*
gite	0	0	0	0

Tabella 3: Esempi di codifica dei valori di frequenza nel lessico parlato adulto.

Per entrate e sotto-entrate sono disponibili misure relative al vicinato ortografico: è riportato il numero dei vicini ortografici, la frequenza del vicino ortografico con frequenza maggiore, la frequenza media e la somma delle frequenze dei vicini ortografici; tali dati sono stati ottenuti usando un algoritmo di ricerca dei vicini ortografici applicato alle occorrenze del CoLFIS[4].

Tutti i valori di frequenza sono disponibili sia come misura grezza che riportati ad 1 milione di occorrenze con l'obiettivo di renderli comparabili tra loro.

[3] I "vicini ortografici" di una parola sono le parole ottenute dalla parola data cambiando una lettera per volta per volta in una determinata posizione.

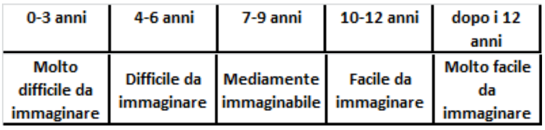
[4] L'algoritmo è disponibile al link: http://ip146172.psy.unipd.it/claudio/vicini2.php

SOTTO-ENTRATA	N COUNT		FREQUENZA MEDIA VICINATO ORTOGRAFICO	VICINO CON FREQUENZA MAGGIORE	SOMMA DELLA FREQUENZA DEI VICINI
arare	*4*	*valore grezzo*	*26*	*89*	*100*
		valore per milione	*8*	*28*	*31*
ara	*27*	*valore grezzo*	*744*	*9442*	*20046*
		valore per milione	*233*	*2959*	*6282*
capriola	*4*	*valore grezzo*	*3*	*5*	*12*
		valore per milione	*0,94*	*1,57*	*3,77*
capriole	*3*	*valore grezzo*	*3*	*5*	*10*
		valore per milione	*1,04*	*1,57*	*3,13*

Tabella 4: Esempi di codifica della numerosità e della frequenza del vicinato ortografico.

3. Giudizi su Variabili Soggettive

Non solo le proprietà linguistiche oggettive delle parole incidono significativamente sui processi di rappresentazione ed elaborazione lessicale; alcuni fattori rilevanti dipendono, piuttosto, dall'esperienza soggettiva dei parlanti. È il caso di due variabili come l'età di acquisizione delle parole, ovvero l'età alla quale sono stati appresi per la prima volta in forma scritta e/o parlata una parola e il suo significato (Carroll e White, 1973), e l'immaginabilità, la proprietà di una parola di evocare un'immagine mentale, una rappresentazione visiva o un'altra esperienza sensoriale (Paivio, Yuille e Madigan, 1968). Per il loro carattere di soggettività è molto difficile disporre di dati riguardanti queste due variabili che sono cruciali in compiti di lettura, riconoscimento o produzione.



0-3 anni	4-6 anni	7-9 anni	10-12 anni	dopo i 12 anni
Molto difficile da immaginare	Difficile da immaginare	Mediamente immaginabile	Facile da immaginare	Molto facile da immaginare

Figura1: Scale a 5 punti usate per ottenere i valori soggettivi sull'età di acquisizione e sull'immaginabilità.

In questa banca-dati sono confluiti i risultati derivanti dalla conduzione di due studi finalizzati ad ottenere valutazioni soggettive dell'immaginabilità e dell'età di acquisizione delle parole attraverso l'impiego delle scale a 5 punti riportate in Figura 1[5].

	NOMI		VERBI	
	media	dev. st.	media	dev. st.
NUMERO DI LETTERE	*7,5*	*1,4*	*8,1*	*1,5*
NUMERO DI SILLABE	*3,1*	*0,6*	*3,5*	*0,6*
NUMERO DI FONEMI	*7,1*	*1,4*	*7,5*	*1,3*
FREQUENZA CUMULATIVA* NELLO SCRITTO ADULTO	*40,1*	*52,2*	*87,6*	*147,8*
FREQUENZA CUMULATIVA* NEL PARLATO ADULTO	*29,4*	*68,2*	*98,5*	*257,4*
FREQUENZA CUMULATIVA* NELLO SCRITTO INFANTILE	*21,7*	*62,2*	*160,0*	*322,1*
NUMERO DI VICINI ORTOGRAFICI	*2,6*	*2,9*	*2,2*	*2,4*
FREQUENZA MEDIA* VICINI ORTOGRAFICI	*8*	*12,4*	*4,8*	*7,6*
ETÀ DI ACQUISIZIONE	*7,1*	*2,3*	*6,1*	*2,2*
IMMAGINABILITÀ	*3,6*	*0,6*	*4,0*	*0,4*

**calcolata su un milione di occorrenze*

Tabella 5: Medie e deviazioni standard per le principali variabili formali, distribuzionali e soggettive dei nomi e verbi contenuti nella banca-dati.

4. Variabili Morfologiche e Morfo-Fonologiche

Il ruolo della variabili distribuzionali incide sulla rappresentazione delle parole nel lessico mentale non solo in termini assoluti: alcune proprietà distribuzionali possono prevedere la tendenza delle parole ad essere elaborate attraverso i morfemi costituenti durante l'accesso lessicale. In tal senso, hanno un ruolo variabili come la numerosità, la regolarità e la produttività dei paradigmi flessivi delle parole (Colombo, Laudanna, De Martino e Brivio, 2004), la trasparenza semantica o la frequenza dei costituenti morfemici delle parole, la confondibilità di affissi morfologici con sequenze di segmenti ricorrenti (Taft e Forster, 1975; Laudanna e Burani, 1995) e, infine, le implicazioni morfologiche di variabili grammaticali come il genere dei nomi (De Martino, Bracco, Postiglione e Laudanna, 2017).

Pertanto, con lo scopo di dare indicazioni sulle principali variabili morfologiche e morfo-fonologiche, nomi e verbi della banca-dati sono stati sottoposti a una serie di operazioni di codifica. Per ciascuna entrata è indicato se si tratti di una forma derivata da un nome (*zampata, pugnalare*), da un verbo (*passeggiata*, *scavalcare*), da

[5] A ciascuno studio hanno preso parte 55 studenti universitari.

un aggettivo (*carezza*, *aggiustare*), da un avverbio (*attraversare*), se si tratti di una forma composta (*benedire*, *parapiglia*) o parasintetica (*arricciare*).

1^A CONIUGAZIONE	2^a CONIUGAZIONE		3^a CONIUGAZIONE
	RIZOATONI	RIZOTONICI	
73	*14*	*1*	*12*

Tabella 6: Distribuzione delle 100 entrate-verbo in base alla coniugazione di appartenenza.

Ciascuna entrata è corredata dall'informazione sulla classe flessiva di appartenenza; per i verbi è indicata l'appartenenza alla 1^a, 2^a (con distinzione tra verbi rizoatoni (*cadere*), o rizotonici (*accendere*)) o 3^a coniugazione ed è segnalata la presenza di irregolarità all'interno del paradigma della parola (allomorfi o di variazioni fonotattiche della radice).

ENTRATA	SOTTO-ENTRATA	PRESENZA DI IRREGOLARITÀ NEL PARADIGMA	REGOLARITÀ DELLA FORMA
svenire	*svenire*	*sì*	*reg*
	sviene	*sì*	*irr*
rispondere	*rispondere*	*sì*	*reg*
	risponde	*sì*	*reg*

Tabella 7: Esempi di codifica della presenza di irregolarità nel paradigma dei verbi della banca-dati.

Per i nomi è indicato il genere grammaticale (*capriola*, femminile; *furto*, maschile) e il tipo di alternanza della vocale finale tra singolare e plurale (*capriola/capriole,* a_e; *furto/furti,* o_i, *analisi,* invariabile, ecc.).

GENERE	ALTERNANZA DELLA VOCALE FINALE TRA FORMA SINGOLARE E FORMA PLURALE				
		a_e	o_i	e_i	invariabili
FEMMINILE	*71*	*67*	*0*	*3*	*1*
MASCHILE	*29*	*0*	*28*	*1*	*0*

Tabella 8: Distribuzione delle 100 entrate-nome in base al genere grammaticale e all'alternanza della vocale finale tra le forme del singolare e del plurale.

5. Variabili Semantiche

Aspetti cruciali per la rappresentazione lessicale delle parole sono le caratteristiche del significato: parole che veicolano significati multipli o parole polisemiche tendono ad essere elaborate in maniera diversa da parole dal significato univoco e spesso è stato osservato che l'effetto dell'ambiguità e della polisemia sono modulati dalla frequenza d'uso dei vari sensi o significati (Klepousniotou e Baum, 2007; Mancuso, Tagliaferri e Laudanna, 2016; Rodd, Gaskell e Marslen-Wilson, 2004).

È stato anche evidenziato che nella rappresentazione lessicale hanno un ruolo importante alcuni aspetti della rappresentazione concettuale delle parole come quelli che riguardano l'uso di uno strumento o di una parte del corpo per eseguire un'azione denotata da un nome o un verbo (Hauk, Johnsrude e Pulvermuller, 2004; Jonkers e Bastiaanse, 2007).

Seguendo le indicazioni della letteratura, per ciascuna entrata è indicata la presenza di eventuali forme di ambiguità lessicale come l'omonimia, ovvero l'esistenza di più entrate nel dizionario (Sabatini e Coletti, 2008) corrispondenti a significati multipli e tra loro non connessi veicolati dalla stessa forma, l'ambiguità grammaticale, ovvero l'esistenza di più entrate nel dizionario in base a differenze di classe grammaticale con sovrapposizione di significato e, infine, la polisemia, ovvero l'esistenza di più sensi semanticamente e/o etimologicamente connessi e ricondotti nel dizionario alla stessa voce.

Nel database, per ciascun senso veicolato è riportato un esempio di frase.

SOTTO-ENTRATA	OMONIMIA	AMBIGUITÀ GRAMMATICALE	POLISEMIA	ESEMPIO
usura	+	+	+ *3 sensi*	*L'usura è un reato gravissimo* *L'olio evita l'usura dell'ingranaggio* *L'attrito usura le ruote*
critica	-	+	+ *4 sensi*	*Ho letto la tua critica del progetto* *Aldo è esperto di critica storica* *Sul film la critica è divisa* *Il chairman spesso critica tutti gli interventi*
benedire	-	-	+ *3 sensi*	*Il papa benedice i fedeli* *Che Dio ti benedica* *Benedico il giorno in cui presi quella decisione*

Tabella 9: Esempi di codifica di casi di omonimia, ambiguità grammaticale e polisemia tratti dalla banca dati.

Per i verbi, inoltre, è indicato per ognuno dei possibili significati o sensi se esso si riferisca ad azioni per le quali è necessario l'uso di uno strumento o di una parte del corpo e se il soggetto può essere umano o non umano. Infine, sia per i nomi sia per i verbi è indicato l'eventuale impiego all'interno di espressioni polirematiche (*analisi* in "*in ultima analisi",* loc. avv.; *vendere* in "*vendere cara la pelle*", loc.v. De Mauro, 2014).

	SENSO	USO DI UNO STRUMENTO	PARTE DEL CORPO COINVOLTA NELL'AZIONE	ANIMATEZZA DEL SOGGETTO
bollire	*l'acqua bolle a 100 gradi*	-	-	*non umano*
	ho bollito il riso	+	*braccia*	*umano*
	Luigi bolle di rabbia	-	-	*umano*
	la teiera bolle sul fuoco	-	-	*non umano*
	le patate stanno bollendo	-	-	*non umano*

Tabella 10: Esempi della codifica di informazioni semantico-concettuali dei verbi presenti nella banca-dati.

6. Variabili Sintattiche

Un numero crescente di studi sta mettendo in evidenza che l'elaborazione lessicale di nomi e verbi è sensibile alla manipolazione di variabili sintattiche come la struttura argomentale (Collina, Marangolo, e Tabossi, 2001; Thompson, Lange, Schneider, e Shapiro, 1997), il tipo di sotto-categorizzazione e il numero di ruoli tematici (De Bleser e Kauschke, 2003). In base alle indicazioni di questi studi, per ciascun verbo della banca-dati e per ciascun senso o significato ammesso dal verbo stesso è specificato il comportamento sintattico in termini di transitività o intransitività; per l'uso intransitivo dei verbi è segnalata la possibilità di avere un complemento oggetto interno (*dormire*). Per ciascun uso del verbo è indicato anche il numero minimo di argomenti ammessi e la struttura di sotto-categorizzazione.

LEMMA	ESEMPIO	USO SINTATTICO	NUMERO DI ARGOMENTI	STRUTTURA DI SOTTO-CATEGORIZZAZIONE
tagliare	*questo coltello taglia bene*	*intransitivo*	*0*	*sogg-v*
	(Fig) tagliare per i prati	*intransitivo*	*1*	*sogg-v-prep.arg*
	tagliare una torta	*transitivo*	*1*	*sogg-v-arg.*
	mi sono tagliato	*riflessivo*	*0*	*sogg-v*
	mi taglio i capelli	*riflessivo*	*1*	*sogg-v-arg*

Tabella 11: Esempi della codifica della struttura argomentale e della struttura di sotto-categorizzazione di verbi.

Infine, è presente l'informazione relativa alla possibilità di usare il verbo in forma riflessiva ed è riportato il tipo di ausiliare ammesso per ciascun significato possibile (*essere* per *svenire*; *avere* per *abbaiare*; entrambi per *imbiancare* a seconda del significato del verbo*: Aldo ha imbiancato le pareti*; *Aldo è imbiancato precocemente*).

Per ciascun senso possibile di un verbo o nome è indicato il numero di argomenti che ne completano il significato e che devono essere obbligatoriamente espressi al fine di usare il verbo o il nome in frasi grammaticalmente corrette e non semanticamente incomplete (*appendere*, 2 argomenti: *Ho appeso i vestiti nell'armadio*; *crescita*, 1 argomento: *La crescita del bambino è stata rapidissima intorno ai tre anni*).

7. Misure Comportamentali

Per ciascuna entrata della banca-dati sono disponibili misure comportamentali ricavate dalla somministrazione di un esperimento di decisione lessicale visiva[6] a 110 parlanti italiani. I 100 nomi e i 100 verbi[7] della banca-dati sono stati presentati all'interno di una lista più ampia in cui sono state impiegate altre parole-filler (140 nomi e 140 verbi). La lista complessiva di 480 nomi e verbi è stata suddivisa in due sotto-liste composte da 280 parole (140 nomi e 140 verbi) e 280 non-parole (140 pseudo-nomi e 140 pseudo-verbi). A ciascun partecipante all'esperimento è stata somministrata una lista da 560 stimoli (280 parole e 280 non-parole).

	NOMI		VERBI	
	media	dev. st.	media	dev. st.
TEMPI DI RISPOSTA (MILLISECONDI)	539,8	48,7	524,7	44,2
NUMERO DI ERRORI	5,3	7,4	3,0	4,9

Tabella 12: Tempi medi di riconoscimento e numero di errori.

I tempi medi di riconoscimento e l'accuratezza ottenuti con questo esperimento, al pari delle altre informazioni contenute nella banca-dati, possono essere usati per selezionare nomi e verbi perfettamente bilanciati per tutti i parametri psicolinguistici rilevanti nell'accesso lessicale in

[6] In questo esperimento le parole venivano presentate al centro dello schermo di un computer per un tempo limite di un secondo. I partecipanti dovevano decidere se esse fossero parole reali dell'italiano.

[7] In questo esperimento i verbi sono stati presentati nella forma infinita e i nomi nella forma singolare.

studi finalizzati a confrontare le due classi di parole.

Riferimenti Bibliografici

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., Thornton, A. M. (2005). *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS).* http://linguistica.sns.it/CoLFIS/Home.htm

Bracco G., Laudanna, A. (2012). Meccanismi di competizione tra forme verbali nell'accesso al lessico mentale. In: P. M. Bertinetto, V. Bambini, I. Ricci e Collaboratori (a cura di). *Linguaggio e cervello / Semantica*, Atti del XLII Convegno della Società di Linguistica Italiana (Pisa, Scuola Normale Superiore, 25-27 settembre 2008). Roma: Bulzoni. Volume 2.

Carroll, J. B., White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology*, 25(1), 85-95.

Collina, S., Marangolo, P., Tabossi, P. (2001). The role of argument structure in the production of nouns and verbs. *Neuropsychologia*, 39(11), 1125-1137.

Colombo, L. (1992). Lexical stress effect and its interaction with frequency in word pronunciation. *Journal of Experimental Psychology: Human Perception and Performance*, 18(4), 987-1003

Colombo, L. (1993). Locus o loci dell'effetto frequenza. In: A. Laudanna, C. Burani (a cura di). *Il lessico: processi e rappresentazioni.* Roma: La Nuova Italia Scientifica.

Colombo, L., Laudanna, A., De Martino, M., Brivio, C. (2004). Regularity and/or consistency in the production of the past participle? *Brain and language*, 90(1), 128-142.

De Bleser, R., Kauschke, C. (2003). Acquisition and loss of nouns and verbs: Parallel or divergent patterns? *Journal of Neurolinguistics*, 16, 213–229.

De Martino, M., Bracco, G., Postiglione, F., Laudanna, A. (2017). The influence of grammatical gender and suffix transparency in processing Italian written nouns. *The Mental Lexicon*, 12(1), 107-128.

De Mauro, T. (2014). https://dizionario.internazionale.it/

De Mauro, T., Mancini, F., Vedovelli, M., Voghera, M. (1993*). Lessico di frequenza dell'italiano parlato.* Milano: Etas libri.

Hauk, O., Johnsrude, I., Pulvermuller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron,* 41, 301–307.

Jonkers, R., Bastiaanse, R. (2007). Action naming in anomic aphasic speakers: Effects of instrumentality and name relation. *Brain and Language*, 102, 262–272.

Klepousniotou, E., Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics,* 20(1), 1–24.

Laudanna, A. (2006). Ortografia. A. Laudanna, M. Voghera, M. (a cura di), *Il linguaggio. Strutture linguistiche e processi cognitivi*, Bari, Laterza.

Laudanna, A., Burani, C. (1995). Distributional properties of derivational affixes: Implications for processing. In L.B. Feldman (a cura di). *Morphological aspects of language processing: Cross-Linguistic Perspectives*. Hillsdale: Lawrence Erlbaum Associates.

Mancuso, A., Tagliaferri, R., Laudanna, A. (2016). Parole ambigue nel lessico mentale: un modello computazionale per spiegare gli effetti di omonimia e polisemia in riconoscimento. In: M. Cruciani, O. Gigliotta, D. Marocco, O. Miglino, S. Moretti, M. Ponticorvo, F. Rubinacci (a cura di). *Apprendimento, cognizione e tecnologia.* Napoli: Università degli Studi di Napoli Federico II.

Marconi, L., Ott, M., Pesenti, E., Ratti, D., Tavella, M. (1993*). Lessico Elementare. Dati statistici sull'italiano letto e scritto dai bambini delle elementari.* Bologna: Zanichelli.

Paivio, A., Yuille, J. C., Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1), 1-25.

Pinker, S. Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6 (11),* 456-463.

Rodd, J., Gaskell, G., Marslen-Wilson, W. (2004) Modeling the effects of semantic ambiguity in word recognition. *Cognitive Science,* 28, 89–104.

Sabatini, F., Coletti, V. (2008). DISC: Dizionario Italiano Sabatini Coletti. Firenze: Edizione Giunti. http://dizionari.corriere.it/dizionario_italiano/

Taft, M., Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, 14(6), 638-647.

Thompson, C. K., Lange, K. L., Schneider, S. L., Shapiro, L. P. (1997). Agrammatic and non-brain-damaged subjects' verb and verb argument structure production. *Aphasiology, 11*, 473-490.

Tagging Semantic Types for Verb Argument Positions

Francesca Della Moretta
University of Pavia / Pavia, Italy
francesca.dellamoretta01
@universitadipavia.it

Anna Feltracco
Fondazione Bruno Kessler / Trento, Italy
University of Pavia / Pavia, Italy
University of Bergamo / Bergamo, Italy
feltracco@fbk.eu

Elisabetta Jezek
University of Pavia / Pavia, Italy
jezek@unipv.it

Bernardo Magnini
Fondazione Bruno Kessler / Trento, Italy
magnini@fbk.eu

Abstract

English. Verb argument positions can be described by the semantic types that characterise the words filling that position. We investigate a number of linguistic issues underlying the tagging of an Italian corpus with the semantic types provided by the T-PAS (Typed Predicate Argument Structure) resource. We report both quantitative data about the tagging and a qualitative analysis of cases of disagreement between two annotators.

Italiano. *Le posizioni argomentali di un verbo possono essere descritte dai tipi semantici che caratterizzano le parole che riempiono quella posizione. Nel contributo affrontiamo alcune problematiche linguistiche sottostanti l'annotazione di un corpus italiano con i tipi semantici usati nella risorsa T-PAS (Typed Predicate Argument Structure). Riportiamo sia dati quantitativi relativi all'annotazione, sia una analisi qualitativa dei casi di disaccordo tra due annotatori.*

1 Introduction

Words that fill a certain verb argument position are characterised for their semantic properties. For instance, the fillers of the object position of the verb "eat" are typically required to share the fact that they are edible objects, like "meat" and "bread". There has been a vast literature in lexical semantics addressing, under different perspectives, this issue, including the notion of selectional preferences (Resnik, 1997) (McCarthy and Carroll, 2003), the notion of prototypical categories (Rosch, 1973), and the notion of lexical sets (Hanks and Jezek, 2008) (Jezek and Hanks, 2010). However, despite the large theoretical interest, there is still a limited amount of empirical evidences (e.g. annotated corpora) that can be used to support linguistic theories. Particularly, for the Italian language, there has been no systematic attempt to annotate a corpus with semantic tagging of verb argument positions

In this paper we assume a corpus-based perspective, and we focus on manually tagging verb argument positions in a corpus with their corresponding semantic classes, selected from those used in the T-PAS resource (Jezek et al., 2014). We make use of an explicit set of semantic categories (i.e., an ontology of Semantic Types), hierarchically organised (e.g. inanimate subsumes food): we are interested in a qualitative analysis, a rather different perspective with respect to recent works that exploit distributional properties of words filling argument positions (Ponti et al., 2016; Ponti et al., 2017). We run a pilot annotation on a corpus of sentences. We aim at investigating how human annotators assign semantic types to argument fillers, and to what extent they agree or disagree.

A mid term goal of this work is the extension of the T-PAS resource with a corpus of annotated sentences aligned with the T-PASs of the verbs (see section 2). This would have a twofold impact: it would allow a corpus based linguistic investigation, and it would provide a unique dataset for training semantic parsers for Italian.

The paper is structured as follows. Section 2 introduces T-PAS and the ontology of semantic types used in the resource. Section 3 describes the annotation task and the guidelines for annotators. Section 4 presents the annotated corpus and the data of the inter-annotator agreement. Finally, Section 5 discusses the most interesting phenomena that emerged during the annotation exercise.

2 Overview of the T-PAS resource

The T-PAS resource is an inventory of 4241 Typed Predicate Argument Structures (T-PASs) - for example [[Human]] **partecipa a** 'takes part in' [[Event]] - for 1000 average polysemy Italian verbs, acquired from the ItWaC corpus (Baroni and Kilgarriff, 2006) by manual clustering of distributional information about Italian verbs (Jezek et al., 2014), following the Corpus Patterns Analysis (CPA) procedure (Hanks, 2004) (Hanks and Pustejovsky, 2005) which consists in recognising the relevant structures of a verb and identifying the Semantic Types (STs) for their argument slots by generalizing over the lexical sets observed in a sample of 250 concordances. The current list of about 230 semantic types used in the resource (e.g. human, event, location, artifact - henceforth, STs) is corpus derived, that is, STs are the result of manual generalization over the lexical sets found in the argument positions in the concordances, for example in the [[Event]] argument position of *partecipare* we find *gara*, *riunione*, *selezione*, and so forth. Besides the T-PASs and the hierarchically organized list of STs, the resource contains a corpus of sentences that instantiate the different T-PASs for each verb. Each sentence is therefore currently tagged with the number of the T-PAS it instantiates; the tag is located on the verb. No further information is present in the instance except for the T-PAS number.

3 Annotating Semantic Types

The main goal of the annotation effort reported in this paper is to enrich the annotation already present in the examples associated with each T-PAS. Specifically, given a T-PAS of a verb and an example from the corpus, we annotate the lexical items (in the example) generalised by the STs (in the T-PAS).

For instance, Example (1) shows the T-PAS#1 of the verb *vendere* (Eng. 'to sell'), and a sentence associated to it. The task consists in annotating *prodotti tipici* (Eng. 'traditional products') as a lexical item for [[Inanimate]]-obj.

(1) [[Human | Business Enterprise]] **vendere** [[Inanimate | Animal]]

"[..] il nome di un'associazione brasiliana che **vendeva** anche prodotti tipici" [1]

We annotate the content word(s) that is the head-noun both in case of the noun-phrases (NP) (e.g. *give a cake*) and in case of prepositional-phrases (PP) (e.g. *give a cake to his little son*). In the case the head-noun is a quantifier, the quantifier is not tagged but the quantified element is (e.g. *to give a piece of cake*).

Notice that more than one token can be annotated, e.g. in the case of multiword expressions such as *prodotti tipici* in Example (1), and more than one item can be tagged for the same argument position, e.g. in case of coordination, such in *[..] che **vendeva** anche prodotti tipici e cartoline*" [2].

In the case an argument is not present in the sentence (for instance, when the subject of the verb is unexpressed), we do not signal this lack.

On the other hand, the annotation accounts for the following cases.

Semantic mismatches. Lexical items are annotated according to the T-PAS; however, the annotator can use a different ST, if she/he thinks the one specified in the T-PAS does not apply. For instance, Example (2) reports another instance of T-PAS#1 of *vendere* in which *lavoro* has been annotated as [[Activity]], a ST not selected by the T-PAS#1 of *vendere* in object position (see the T-PAS in Example (1)).

(2) "il lavoro come qualsiasi altra cosa può essere acquistato e **venduto**."[3]

Syntactic mismatches. We account for cases in which the syntactic role of the lexical items does not match with the one proposed in the T-PAS, e.g. in cases of passive forms of verbs, where the subject and prepositional phrase introduced by *da* correspond respectively to the object and the subject of the active construction. In Example (2), *lavoro* is the syntactic subject of the passive clause, and it is generalized by [[Activity]]) in the object position of the T-PAS. In such cases we annotate both the ST of the lexical item and its grammatical relation using the one in the T-PAS.

Pronouns. In case the argument of the verb is realised as a pronoun, we tag the pronoun without assigning a ST. The pronoun is then linked to the noun(s) it refers to, and this noun is actually

[1] Eng. '[..] the name of that Brazilian association that **was selling** traditional products'

[2] Eng. '[..] that **was selling** traditional products and postcards'

[3] Eng. 'jobs can be **sold** and bought just like anything.'

tagged with the ST label. In case the pronoun is agglutinated to the verb (i.e. it is found in the same token of the verb, e.g. *venderla*, Eng. 'to sell it'), the part of the token corresponding to the pronoun is specified and, as just specified, the noun is annotated with the ST.

Impersonal constructions. In case of impersonal constructions with an indefinite pronoun, the pronoun is annotated and the ST it refers to is specified: e.g. *In Germania [..] si* ***vende*** *a 10 euro al chilo* [4], *si* is annotated with [[Human]].

We annotated the examples in T-PAS using CAT (Content Annotation Tool)[5], a general-purpose text annotation tool (Bartalesi Lenzi et al., 2012).

4 Results of the Pilot Annotation

The pilot annotation consisted in a selection of 3554 sentences extracted from the current version of T-PAS[6] associated to 25 Italian verbs, selected with different levels of polysemy (from a minimum of 2 to a maximum of 10 T-PASs), and argument structure. The average polysemy of the 25 verbs (i.e. number of senses divided by the number of verbs) is 4.08, and for each T-PAS (sense) we have an average of 34.84 annotated sentences.

The annotation was carried out by a master student in linguistics, who was trained on the T-PAS resource, but had no previous experience in annotation. The annotator was able to tag the 3554 sentences in one month.

Table 1 shows the main data of the pilot annotation. Overall, we annotated 5342 argument positions expressed in the 3554 sentences, with an average of 1.5 argument per sentence. Out of the 230 Semantic Types available in the T-PAS ontology, 99 have been selected during the annotation, which means that we used about 40% of the STs contained in the hierarchy.

Data	Total
# Verbs	25
# T-PASs	102
# Examples	3554
# Examples per T-PAS	34.84
# Semantic Types used	99

Table 1: Pilot annotation results.

4.1 Inter-annotator Agreement

In order to assess the reliability of the annotated data, we run an Inter-Annotator Agreement (IAA) test.[7] We asked a second annotator to annotate a sample of 11 T-PASs associated to 3 different verbs (i.e., *pulire*, *vendere* and *sbottonare*). These verbs were chosen because they correspond to about 10% of the annotated sentences. Moreover, we selected them because they present a low or middle degree of polysemy with respect of the group of 25 verbs initially annotated. The second annotator was provided with the task guidelines and a training session was done to solve potential uncertainties in annotation. The second annotator was trained on a selection of corpus instances derived from verb lemmas, which are not included in the evaluation we report here.

Table 2 shows the results of the IAA for each T-PAS. We measured both the agreement on argument annotation, calculated with the Dice's coefficient (Rijsbergen, 1979), and the agreement on ST annotation, calculated as the accuracy (Manning et al., 2008) among the two annotators. As reported in the last row of Table 2, the average agreement is 0.87 for argument annotation, and 0.83 for ST annotation.

T-PAS	Argument *Dice's value*	ST *Accuracy*
Pulire, T-PAS#1	0.83	0.74
Pulire, T-PAS#2	1	1
Sbottonare, T-PAS#1	0.94	0.89
Sbottonare, T-PAS#2	0.95	0.98
Sbottonare, T-PAS#3	1	1
Sbottonare, T-PAS#4	0.88	0.90
Vendere, T-PAS#1	0.87	0.81
Vendere, T-PAS#2	0.33	0.5
Vendere, T-PAS#3	0.8	1
Vendere, T-PAS#4	1	1
Vendere, T-PAS#5	1	1
Overall average	0.87	0.83

Table 2: Inter Annotator Agreement.

A special case is *vendere T-PAS#2*, which shows the lowest score for both argument and STs annotation. The annotation task allowed annotators to discard sentences which according to their opinion did not fit the sense of the T-PAS taken into consideration. *Vendere T-PAS#2* has only a few corpus instances, which were mostly discarded or

[4] Eng. 'In Germany, they sell it at 10 euro per kilo'.

[5] `https://dh.fbk.eu/resources/cat-content-annotation-tool`

[6] http://tpas.fbk.eu

[7] Cinková et al. (2012) held an IAA on pattern-identification using the CPA procedure in 30 English verbs.

tagged differently by the two annotators, causing low agreement in the results for this T-PAS.

5 Discussion

This Section discusses the most interesting phenomena that emerged during the annotation exercise, particularly in light of the Inter-annotator Agreement.

5.1 Discussion: Argument Tagging

In this paragraph, we focus on the disagreements we found in argument tagging. The annotation task was difficult because the annotators had to identify the semantic structure of the verbs, using syntactic criteria to distinguish whether a lexical element was an argument or not.

Annotating pronouns was also a very demanding process since it implies the identification of co-reference chains. Differences in argument annotation between the two annotators, that impact the arguments Dice score, lie mainly in the annotation of pronouns and in the identification of co-referents. One annotator usually tends to annotate all the pronouns contained in an utterance whereas the other tags only the pronoun which is an argument of the verb taken into consideration. In addition, one usually does not identify co-referents which are lexically realised at great distance of words from the tagged verb, whereas the other sometimes annotates co-referents even if the argument has already been identified. There are also differences concerning the extension of annotation e.g. one interpreted *prodotti tipici* as multiword expression and the other did not. Overall, we obtained good agreement results, although some disagreements still remain even if we tried to reduce potential differences in annotation treating as many cases as possible in the guidelines.

5.2 Discussion: Semantic Type Tagging

The main goal of this section is to analyse the results of IAA on ST selection. Annotators used approximately 40 STs even though their expected number (according to the T-PAS resource) was 11. Table 3 represents the ST usage in the IAA experiment for each T-PAS.

Annotators used approximately the expected number of semantic types with some T-PASs, while with others they used many more. To a higher number of STs employed corresponds a lower ST accuracy score (see Table 1), more specifically this correlation is shown by *pulire* T-PAS#1, *sbottonare* T-PAS#1,#4, *vendere* T-PAS#1. There are a number of reasons that justify this STs usage. In some cases one annotator tends to tag the entity denoted by single lexical items instead of the generalisations made by the T-PASs. This causes a sentence specific annotation that employs STs that are end nodes in the hierarchy, which do not correspond to the ones in the reference T-PAS. As future work, we plan to develop a methodology to normalize the STs to the appropriate level of abstraction.

T-PAS	ST Expected *according to the T-PAS*	ST used *A+B*
Pulire, T-PAS#1	4	23
Pulire, T-PAS#2	3	4
Sbottonare, T-PAS#1	2	6
Sbottonare, T-PAS#2	2	4
Sbottonare, T-PAS#3	1	1
Sbottonare, T-PAS#4	1	4
Vendere, T-PAS#1	4	23
Vendere, T-PAS#2	2	3
Vendere, T-PAS#3	3	3
Vendere, T-PAS#4	1	1
Vendere, T-PAS#5	1	1

Table 3: Expected and used STs in the IAA test.

There are also linguistic reasons that intervene in the assignment of different STs to the same lexical element. Annotators captured repeatedly the phenomenon known as *inherent polysemy* by tagging the same lexical elements in two totally different ways. An inherent polysemous noun denotes, depending on the context, a single aspect of an entity which is inherently complex, i.e. that can be described simultaneously by more than one ST (see (Jezek, 2016) and references therein). An example is provided by the nouns that denote countries that in our annotation exercise have been tagged as [[Business Enterprise]], [[Institution]] or [[Area]], pointing out their complex nature of territorial, politic and economic entity. In some cases annotators have privileged different semantic components in the ST annotation process. This is due to the context in which the words are embedded, that determines certain interpretations instead of others. However, sometimes the compositionality principle does not strictly define the meaning of an utterance. Hence some lexical items remain underspecified so that they can receive more than one ST at once.

For instance in example (3) one annotator tagged *lente* as [[Artifact]] highlighting its nature of manufactured object, whereas the other has annotated the lexical item as [[Physical Object Part]] focusing on its nature of constituent element of a bigger object.

(3) "Giles **pulisce** una lente dei suoi occhiali."[8]

Moreover, there are differences is ST assignment caused by regular polysemy (Apresjan, 1974), systematic alternation of meaning that apply to classes of words (Jezek, 2016). IAA results reveal regular polysemy patterns for nouns.

6 Conclusions

We performed a pilot experiment to tag the arguments of verbs, as recorded in the T-PAS resource, with their associated semantic type. We obtained good result in the annotation. By analyzing the cases of inter annotator disagreement, we were able to identify phenomena which lie at the core of such disagreements, such as the presence of inherent polysemous words. Ongoing work includes spelling out the rules for polysemous words tagging more clearly in the guidelines.

References

Iurii Derenikovich Apresjan. 1974. Regular polysemy. *Linguistics*, 32.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90. Association for Computational Linguistics.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. Cat: the celct annotation tool. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12)*, pages 333–338.

Silvie Cinková, Martin Holub, Adam Rambousek, and Lenka Smejkalová. 2012. A database of semantic clusters of verb usages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, pages 3176–3183.

Patrick Hanks and Elisabetta Jezek. 2008. Shimmering lexical sets. In *Proceedings of the XIII EURALEX International Congress*, pages 391–402.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue française de linguistique appliquée*, 10(2):63–82.

Patrick Hanks. 2004. Corpus pattern analysis. In *Proceedings of the Eleventh EURALEX International Congress.*

Elisabetta Jezek and Patrick Hanks. 2010. What lexical sets tell us about conceptual categories. *Lexis*, 4(7):22.

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS: a resource of corpus-derived types predicate-argument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).*

Elisabetta Jezek. 2016. *The lexicon: an introduction*. Oxford University Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Edoardo Maria Ponti, Elisabetta Jezek, and Bernardo Magnini. 2016. Grounding the lexical sets of causative-inchoative verbs with word embedding. In *Proceedings of the Second Italian Conference on Computational Linguistic (CLiC-it 2016).*

Edoardo Maria Ponti, Elisabetta Jezek, and Bernardo Magnini. 2017. Distributed representations of lexical sets and prototypes in causal alternation verbs. *Italian Journal of Computational Linguistics*, to appear.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57.

CJ van Rijsbergen. 1979. Information retrieval. 1979.

Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology*, 4(3):328–350.

[8] Eng.'Giles **cleans** a lens of his glasses'

Linguistic Features and Newsworthiness: An Analysis of News style

Maria Pia di Buono, Jan Šnajder

University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{mariapia.dibuono,jan.snajder}@fer.hr

Abstract

English. In this paper, we present a preliminary study on the style of headlines in order to evaluate the correlation between linguistic features and newsworthiness. Our hypothesis is that each particular linguistic form or stylistic variation can be motivated by the purpose of encoding a certain newsworthiness value. To discover the correlations between newsworthiness and linguistic features, we perform an analysis on the basis of characteristics considered indicative of a shared communicative function and of discriminating factors for headlines.

Italiano. *Questo contributo descrive uno studio preliminare sullo stile dei titoli nelle notizie, al fine di valutare la correlazione tra gli aspetti linguistici e il valore delle notizie. La nostra ipotesi è che ogni particolare forma linguistica o variazione stilistica possa essere motivata dall'obiettivo di codificare un certo valore di notiziabilità. Al fine di analizzare la correlazione tra il valore delle notizie e gli aspetti linguistici, effettuiamo un'analisi sulla base delle caratteristiche considerate indicative di una funzione comunicativa condivisa e di fattori discriminanti per i titoli.*

1 Introduction

Newsworthiness refers to a set of criteria by means of which quantity and type of events are selected in order to produce news (Wolf and de Figueiredo, 1987). That is to say, 'news is not simply that which happens, but that which can be regarded and presented as newsworthy' (Fowler, 2013). Galtung and Ruge (1965) identify a list of factors that an event should satisfy to become news; in other words, the likelihood of an event being considered newsworthy increases with the number of factors it complies with.

The newsworthiness factors reflect a set of values and provide a certain representation of the world (Fowler, 2013). This representation and the corresponding values are constructed and encoded in the language used in the news. For this reason, each particular linguistic form or stylistic variation can be motivated by the purpose of representing a certain value. According to Labov's axiom (1972), style ranges along a single dimension, namely the attention paid to speech. Bell (1984) refutes this axiom, stating that style can be considered also as a response to other factors. These factors constitute a new dimension of stylistic variation, that, in headlines, might be related to the necessity of reflecting newsworthy factors, and meeting two needs: attracting users attention and summarizing contents (Ifantidou, 2009).

This paper aims to provide a preliminary analysis of the linguistic features in news headlines and how these relate to specific newsworthiness categories. The analysis rests on the hypothesis that each particular linguistic form or stylistic variation can be motivated by the purpose of encoding a certain newsworthy value. The remainder of the paper is structured as follows. In Section 2, we describe the related work on stylistic analysis of news and headlines. In Section 3, we describe the data set and the classification scheme we use. In Section 4 we introduce our methodology together with the analysis we perform, while in Section 5 we discuss the results. Section 6 concludes the paper.

2 Related Work

Several works, based on sociolinguistic and discourse analysis frameworks, have investigated stylistic features and linguistic variations in both newspapers and headlines, on the basis of different parameters and aspects (Develotte and Rechniewski, 2001; Pajunen, 2008). The large amount of existing contributions to the field is justified by the social implications of news media communication and its language.

A considerable amount of research has analyzed the language of news media from a broader prospective (Bell, 1991; Matheson, 2000; Cotter, 2010; Conboy, 2013; Fowler, 2013; Van Dijk, 2013). Generally speaking, these works emphasize the influence of news language on our perception of the world, due to the fact that news media operate a selection of events and narrative, and use the language to project those.

Another strand of research focuses on specific linguistic aspects in journalistic style. For instance, Tannenbaum and Brewer (1965) analyze the syntactic structure across different news content areas, while Schneider (2000) analyzes the textual structures in British headlines, revising the traditional distinction among verbal and nominal headlines.

3 Data Description

In our work, we adopt the data set proposed for SemEval-2007 task 14 (Strapparava and Mihalcea, 2007), which is a corpus formed by 1250 headlines, extracted from major newspapers and news web sites such as New York Times, CNN, BBC News, and Google News search engine. Originally, SemEval-2007 task 14 data set has been developed for emotion classification and annotated with emotion labels. Relevant for the purpose of the present work is the annotation of this dataset by di Buono et al. (2017), who provided additional newsworthiness labels (“news values”), using the scheme proposed by Harcup and O’Neill (2016). Harcup and O’Neill proposed a set of 15 values, corresponding to a set of requirements that news stories have to satisfy to be selected for publishing. They claimed that these criteria are related also to practical considerations, e.g., the availability of resources and time, and to a mix of other influences, e.g., who is selecting news, for whom, in what medium and by what means (and available resources), that can cause fluctuations within the suggested hierarchy. Di Buono et al. report that two out of 15 news value labels (Audio-visuals, News organization’s agenda) were difficult to annotate out of context even for trained annotators, while two (Exclusivity, Relevance) were not well-represented in the data. Their final dataset thus contains 11 labels.

Table 1 lists the news value labels, their counts in the data set, and the inter-annotator agreement measured in terms of (adjudicated) kappa-score, as reported by Di Buono et al.

News value	Count	IAA
Bad news	85	0.74
Celebrity	82	0.76
Conflict	86	0.56
Drama	178	0.66
Entertainment	351	0.84
Follow-up	29	0.45
Good news	65	0.56
Magnitude	45	0.37
Shareability	130	0.34
Surprise	43	0.41
Power elite	166	0.72

Table 1: News values labels, their counts, and the inter-annotator agreement in terms of kappa-score.

4 Linguistic Features

Our methodology to define the stylistic variations related to newsworthiness categories relies on a descriptive analysis of different features, i.e., syntactic, lexical and compositional features.

We extracted these using Coh-Metrix,[1] a computational tool that provides a wide range of language and discourse metrics (Graesser et al., 2004; McNamara et al., 2014). Coh-Metrix has been developed on the basis of cognitive models in discourse psychology to detect both coherence and cohesion in texts. According to Louwerse (2004), “coherence refers to the representational relationships of a text in the mind of a reader whereas cohesion refers to the textual indications that coherent texts are built upon.” Coh-Metrix describes coherence and cohesion by means of more than one hundred linguistic features, based on a multilevel framework, i.e., words, syntax, the situation model, the discourse genre, and rhetorical structure (Dowell et al., 2016).

We ran Coh-Metrix analysis on headlines from our dataset, grouped according to the 11 newsworthiness labels. We then analyzed these results manually and decided to adopt a subset of Coh-Metrix indices, which, according to our initial hypothesis, we consider to be discriminating factors for newsworthiness, i.e., *text easibility principal component* and *word information* indices. Being representative of linguistic characteristics and syntax context, such features are suitable to represent stylistic variations and, therefore, the underlied news value.

[1] http://cohmetrix.com

News value	PCSYNz	PCCNCz
Bad news	2.274	3.669
Celebrity	2.239	1.933
Conflict	1.834	1.992
Drama	2.502	2.194
Entertainment	1.923	1.788
Follow-up	1.27	2.646
Good news	1.741	3.122
Magnitude	2.585	2.873
Shareability	2.057	1.678
Surprise	1.451	3.253
Power elite	2.671	0.896

Table 2: Z-scores for PC Syntactic simplicity (PCSYNz) and PC Word concreteness (PCCNCz).

5 Analysis and Results

In our preliminary analysis, we consider two main types of linguistic features: *text easability* and *word information scores*.

5.1 Text Easability Features

Coh-Metrix text easibility indices ("Text easability principal component scores") are designed to measure text ease that goes beyond traditional readability metrics. We focused specifically on two indices related to the syntactic simplicity (PCSYNz) and word concreteness (PCCNCz) (Table 2).

The syntactic simplicity is evaluated on the basis of the number of words and the complexity of syntactic structures of sentences. As far as the syntactic simplicity is concerned, the variability among the categories is not so high, nevertheless, we can distinguish two groups. The first group, with a higher PCSYNz, consists of headlines labeled with the 'Power elite', 'Bad news', 'Shareability', 'Drama', 'Magnitude', and 'Celebrity' news values. Higher scores here indicate that the sentence presents more words and uses complex syntatic structures, as exemplifed by the following headlines from this group:

(1a) *China says rich countries should take lead on global warming* (Power elite)

(1b) *Iraqi suicide attack kills two US troops as militants fight purge* (Bad news)

(1c) *Second opinion: girl or boy? as fertility technology advances, so does an ethical debate* (Shareability)

(1d) *Damaged Japanese whaling ship may resume hunting off Antarctica* (Drama)

(1e) *Ready to eat chicken breasts recalled due to suspected listeria* (Magnitude)

(1f) *Jackass' star marries childhood friend The secrets people reveal* (Celebrity)

The second group consists of headlines labled with 'Entertainment', 'Surprise', 'Follow up', 'Good news', and 'Conflict', which received lower PCSYNz scores, and are thus of less syntactic complexity. Examples of headlines form this group are as follows:

(2a) *Action games improve eyesight* (Entertainment)

(2b) *Breast cancer drug promises hope* (Good news)

(2c) *Merkel: Stop Iran* (Conflict)

The second index, word concreteness, differentiates three groups of headlines: (i) 'Power elite', 'Entertainment', 'Shareability', 'Celebrity', and 'Conflict', all with a low z-score; (ii) 'Follow up', 'Drama' and 'Magnitude', with a medium z-score; and (iii) 'Bad news', 'Surprise' and 'Good news' with a high z-score. The following headlines exemplify each of the three groups:

(1a) *Action intensity boosts vision* (Shareability)

(2a) *Ex-suspect slams anti-terror laws* (Drama)

(3a) *Ancient coin shows Cleopatra was no beauty* (Surprise)

The word concreteness index measures the concreteness level of content words. Thus, news values with lower scores are characterized by a higer number of abstract words and, for this reason, may be less easy to understand without an appropriate context. Our analysis thus suggests that 'Bad news', 'Surprise', and 'Good news' headlines are typically refering to more concrete events and entities than the other categories of news values.

5.2 Word Information

This Coh-Metrix index refers to information about syntactic categories and function words, evaluated in the sentence context. To visualize the relations among newsworthiness and word information, we performed a hierarchical cluster analysis. We first represent each headline as a vector of ten word incidence scores (the number of words of a specific part-speech per 1000 words): incidence scores for nouns, verbs, adjectives, adverbs, personal pronouns, pronouns in first, second, and third person, separately for singular and plural. We then use hierarchical agglomerative clustering with complete linkage and one minus Pearsons correlation coefficient as the distance measure to obtain the clusters.

Fig. 1 shows the resulting dendrogram. We can identify three groups of news values on the basis of their syntactic structures.

The first group consists of only news values that can be defined positive contents/sentiments, namely 'Good news', 'Entertainment', and 'Shareability'. This group is characterized by a quite high incidence of adjective, low incidence of first person singular and third person plural pronouns. Furthermore, this group presents the highest incidence of second person pronouns. As in the samples below:

(1a) *Feeding your brain: new benefits found in chocolate* (Good news)

(1b) *Free Will: Now you have it, now you don't* (Entertainment)

(1c) *Nap your way to a successful career* (Shareability)

The second group consists of 'Celebrity', 'Power elite', and 'Drama'. This group presents low incidence of adjective and adverbs. The most incident pronouns are the first person plural and the third person singular.

(2a) *Beyonce new SI bikini cover girl* (Celebrity)

(2b) *Bush vows cooperation on health care* (Power elite)

(2c) *Collision on icy road kills 7* (Drama)

The third group consists of two subsets, the first one formed by 'Surprise' and 'Magnitude', and the second subset formed by 'Follow up', 'Bad news', and 'Conflict'. 'Surprise' and 'Magnitude' form a different subset due to the presence of the highest score within all categories for the adjective incidence and a low incidence of pronouns. For instance:

(3a) *In the world of life-saving drugs, a growing epidemic of deadly fakes* (Surprise)

(3b) *Flu Vaccine Appears Safe for Young Children* (Magnitude)

The second subset is formed by negative contents/sentiment, characterized by the lowest incidence of adverbs and pronouns:

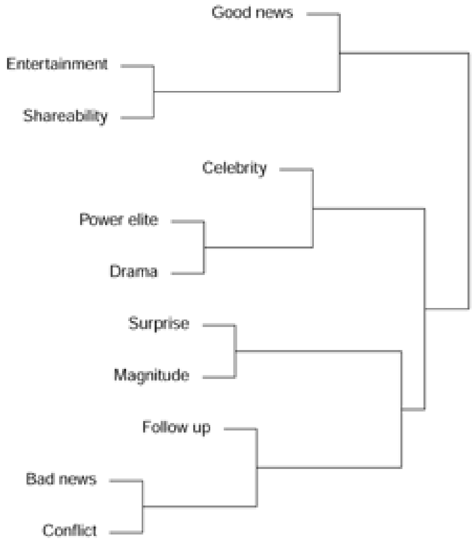


Figure 1: Dendrogram of the 11 newsworthiness categories based on the headline word information features.

(3c) *Eight years for Damilola killers* (Follow up)

(3d) *Bomb kills 18 on military bus in Iran* (Bad news)

(3e) *Venezuela, Iran fight U.S. dominance* (Conflict).

6 Conclusions and Future work

We described a preliminary study for on style of headlines in order to evaluate the correlation among syntactic features and newsworthiness. Our hypothesis is that each particular linguistic form or stylistic variation can be motivated by the purpose of encoding a certain newsworthy value. We performed a linguistic analysis to discover the correlations among newsworthiness and some stylistic features, on the basis of characteristics considered indicative of a shared communicative function and discriminating factors for headlines.

This preliminary analysis opens up a number of interesting research directions. One is the study of other stylistic variations of headlines, besides the ones examined in this paper. Another research direction is the comparison between style in headlines and full-text stories. It would also be interesting to analyze how communicative functions in headlines correlate with the events described in the pertaining text. We intend to pursue some of this work in the near future.

Acknowledgments

This work has been funded by the Unity Through Knowledge Fund of the Croatian Science Foundation, under the grant 19/15: EVEnt Retrieval Based on semantically Enriched Structures for Interactive user Tasks (EVERBEST).

References

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(02):145–204.

Allan Bell. 1991. *The Language of News Media*. Language in society. Blackwell.

Martin Conboy. 2013. *The language of the news*. Routledge.

C. Cotter. 2010. *News Talk: Investigating the Language of Journalism*. Cambridge University Press.

Christine Develotte and Elizabeth Rechniewski. 2001. Discourse analysis of newspaper headlines: a methodological framework for research into national representations. *Web Journal of French Media Studies*, 4(1).

Maria Pia di Buono, Jan Šnajder, Bojana Dalbelo Bašić, Goran Glavaš, Martin Tutek, and Natasa Milic-Frayling. 2017. Predicting news values from headline text and emotions. In *Proceedings of Natural Language Processing Meets Journalism Workshop (EMNLP 2017)*, page to appear.

Nia M. Dowell, Arthur C. Graesser, and Zhiqiang Cai. 2016. Language and discourse analysis with cohmetrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3):72–95.

Roger Fowler. 2013. *Language in the News: Discourse and Ideology in the Press*. Routledge.

Johan Galtung and Mari Holmboe Ruge. 1965. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of peace research*, 2(1):64–90.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Tony Harcup and Deirdre O'Neill. 2016. What is news? news values revisited (again). *Journalism Studies*, pages 1–19.

Elly Ifantidou. 2009. Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *Journal of Pragmatics*, 41(4):699–720.

William Labov. 1972. *Sociolinguistic patterns*. Number 4. University of Pennsylvania Press.

Max M Louwerse. 2004. Un modelo conciso de cohesión en el texto y coherencia en la comprensión. *Revista signos*, 37(56):41–58.

Donald Matheson. 2000. The birth of news discourse: Changes in news language in British newspapers, 1880-1930. *Media, Culture & Society*, 22(5):557–573.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Juhani Pajunen. 2008. Linguistic analysis of newspaper discourse in theory and practice.

Kristina Schneider. 2000. The emergence and development of headlines in British newspapers. *English Media Texts, Past and Present: Language and Textual Structure*, 80:45.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Percy H. Tannenbaum and Richard K. Brewer. 1965. Consistency of syntactic structure as a factor in journalistic style. *Journalism Quarterly*, 42(2):273–275.

Teun A Van Dijk. 2013. *News as discourse*. Routledge.

Mauro Wolf and Maria Jorge Vilar de Figueiredo. 1987. *Teorias da comunicação*. Presença.

Can Monolingual Embeddings Improve Neural Machine Translation?

Mattia A. Di Gangi
University of Trento, Trento, Italy
Fondazione Bruno Kessler
via Sommarive, 18, Trento, Italy
digangi@fbk.eu

Federico Marcello
Fondazione Bruno Kessler
via Sommarive, 18, Trento, Italy
federico@fbk.eu

Abstract

English. Neural machine translation (NMT) recently redefined the state of the art in machine translation, by introducing deep learning architecture that can be trained end-to-end. One limitation of NMT is the difficulty to learn representations of rare words. The most common solution is to segment words into subwords, in order to allow for shared representations of infrequent words. In this paper we present ways to directly feed a NMT network with external word embeddings trained on monolingual source data, thus enabling a virtually infinite source vocabulary. Our preliminary results show that while our methods do not seem effective under large-data training conditions (WMT En-De), they instead show great potential for the typical low-resourced data scenario (IWSLT En-Fr). By leveraging external embeddings learned on Web crawled English texts, we were able to improve a word-level En-Fr baseline trained on 200,000 sentence pairs by up to 4 BLEU points.

Italiano. *La traduzione automatica con reti neurali (neural machine translation, NMT) ha ridefinito recentemente lo stato dell'arte nella traduzione automatica introducendo un'architettura di deep learning che può essere addestrata interamente, dall'input all'output. Una limitazione della NMT è comunque la difficoltà di apprendere rappresentazioni di parole poco frequenti. La soluzione più adottata consiste nel segmentare le parole in sottoparole, in modo da consentire rappresentazioni condivise per parole poco frequenti. In questo lavoro presentiamo dei metodi per fornire ad una rete word embedding esterni addestrati su testi nella lingua sorgente, consentendo quindi un vocabolario virtualmente illimitato sulla lingua di input. I nostri risultati preliminari mostrano che i nostri metodi, pur non sembrando efficaci sotto condizioni di addestramento con molti dati (WMT En-De), risultano invece promettenti per scenari di addestramento con poche risorse (IWSLT En-Fr). Sfruttando* word embedding *appresi da testi inglesi estratti dal Web, siamo riusciti a migliorare un sistema NMT basato a parole e addestrato su 200.000 coppie di frasi fino a 4 punti BLEU.*

1 Introduction

The latest developments of machine translation have been led by the neural approach (Sutskever et al., 2014; Bahdanau et al., 2014), a deep-learning based technique that has shown to outperform the previous methods in all the recent evaluation campaigns (Bojar et al., 2016; Cettolo et al., 2016).

NMT mainly relies on parallel data, which are expensive to produce as they involve human translation. Recently, *back-translation* (Sennrich et al., 2015a) has been proposed to leverage target language data. This consists in enriching the training data with synthetic translations produced with a reverse MT system (Bertoldi and Federico, 2009). Unfortunately, this method introduces noise and seems really effective only when the synthetic parallel sentences are only a fraction of the true ones. Hence, this approach does not allow to leverage huge quantities of monolingual data.

One consequence of the scarcity of parallel data is the occurrence of out-of-vocabulary (OOV) and rare words. In fact, being NMT a statistical approach, it cannot learn meaningful representations

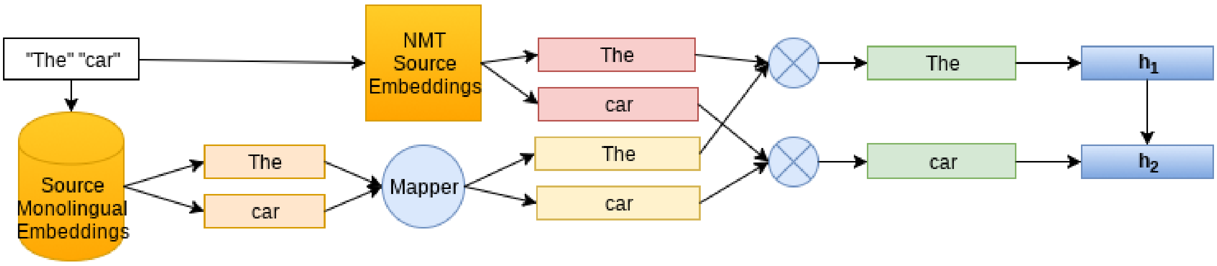


Figure 1: Merging external embeddings with the normal NMT embeddings in the encoder side. The tokens ”The” and ”car” are used to extract the two kinds of embeddings that are merged before being used as input for the encoder RNN.

for rare words and no representation at all for OOV words. The solution up to this moment is to segment words into sub-words (Sennrich et al., 2015b; Wu et al., 2016) in order to have a better representation of rare and OOV words, as parts of their representation will be ideally shared with other words. The drawback of this approach is that it generates longer input sequences, thus exacerbates the handling of long-term dependencies (Bentivogli et al., 2016). In this paper, we propose to keep the source input at a word level while alleviating the problem of rare and OOV words. We do it by integrating the usual word indexes with word embeddings that have been pre-trained on huge monolingual data. The intuition is that the network should learn to use the provided representations, which should be possibly more reliable for the rare words. This should be true particularly for the low-resource settings, where parameter transfer has shown to be an effective approach (Zoph et al., 2016). Because of the softmax layer, the same idea cannot be applied straightforwardly to the target side, hence we continue to use sub-words there. We show that the network is capable to learn how to translate from the input embeddings while replacing the source embedding layer with a much smaller feed-forward layer. Our results show that this method seems effective in a small training data setting, while it does not seem to help under large training data conditions. In the following section we briefly describe the state-of-the-art NMT architecture. Then, we introduce our modification to enable the use of external word embeddings. In Section 4, we introduce the experimental setup and show our results, while in Section 5 we discuss our solution. Finally, in Section 6 we presents our conclusions and the future work.

2 State of the art

Neural machine translation is based on the encoder-decoder-attention architecture (Bahdanau et al., 2014) which jointly learns the translation and alignment models with a sequence-to-sequence process. A sequence of source words $f_1, f_2, \ldots, f_m$ is mapped to sequence of embedding vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}$, via a look-up table $X \in R^{|V| \times d}$, where $|V|$ is the vocabulary size and d is the dimensionality of the embedding vectors. Hence, the memory occupied by the vocabulary is linear in both the vocabulary size and the embeddings size.
The embedding sequence is then processed by a bi-directional RNN (Schuster and Paliwal, 1997):

$$\overrightarrow{\mathbf{h}}_j = g(\mathbf{x}_j, \overrightarrow{\mathbf{h}}_{j-1}), \ \ j = 1, ..m$$

$$\overleftarrow{\mathbf{h}}_j = g(\mathbf{x}_j, \overleftarrow{\mathbf{h}}_{j+1}), \ \ j = m, .., 1$$

where g is the LSTM (Hochreiter and Schmidhuber, 1997) or the GRU (Cho et al., 2014) function, and the two directions are merged with functions like the vector concatenation or the point-wise sum. The sequence of vectors produced by the bidirectional RNN is the encoded representation of the source sentence.
The decoder takes as input the encoder outputs (or states) and produces a sequence of target words $e_1, e_2, \ldots, e_l$. The decoder works by progressively predicting the probability of the next target word e_i given the previously generated target words and the source context vector $\mathbf{c_i}$. At each step, the decoder computes a word embeddings $\mathbf{y}_{i-1}$ of the previous target word, applies one or more recurrent layers, an attention model function and a softmax layer. The recurrent layers produce an hidden state $\mathbf{s}_i$

$$\mathbf{s}_i = g(\mathbf{y}_{i-1}, \mathbf{s}_{i-1})$$

where, g can be computed with one or more LSTM or GRU layers. The output of the RNN is then used by the attention model (Luong et al., 2015a) to weight the source vectors according to their similarity with it.

$$\alpha_{ij} = \frac{\exp(score(\mathbf{s}_i, \mathbf{h}_j))}{\sum_{k=1}^{m} \exp(score(\mathbf{s}_i, \mathbf{h}_k))}$$

The weights are used to compute a weighted average of the encoder outputs, which represents the source context

$$\mathbf{c}_i = \sum_{j=1}^{m} \alpha_{ij} \mathbf{h}_j$$

The source context vector is then combined with the output of the last RNN layer in a new vector $\mathbf{z}_i$ that is passed as input to the softmax layer to compute the probability for each word in the vocabulary to be the next word, such that:

$$p(e \mid e_{i-1}, c_i) \propto \exp(\mathbf{o}^\top \mathbf{z}_i)$$

where $\mathbf{z_i}$ is a column of $\mathbf{Z}$, a matrix with the same size of the target-side embedding matrix. Let Θ be the set of all the network parameters, then the objective of the training is to find parameter values maximizing the likelihood of the training set S, i.e.:

$$\sum_{(\mathbf{f},\mathbf{e}) \in S} \sum_{i=1}^{|e|} \log p(e_i | e_{<i}, \mathbf{c}_i; \Theta)$$

In order to achieve open-vocabulary translation with a limited vocabulary size, the words are segmented into sub-words, and the words with shared sub-words share part of their representation. The most common segmenting approach was introduced by Sennrich et al. (2015b) and exploits only statistical information, but there are promising research lines trying to use linguistically motivated segmentations (Ataman et al., 2017)

3 Using external word embeddings

The method we propose is based on the training of word embeddings from source-language monolingual data. We use these embeddings as an input to the network, and we remove the source-side embedding matrix. As the external embeddings have been learned for a task that is not machine translation, we introduce a feed-forward layer to map the embeddings into a new space that is more useful for the translation task:

$$\tilde{\mathbf{x}_j} = \tanh(\bar{\mathbf{x}}_j^\top \mathbf{W} + \mathbf{b}) \text{ for j = } 1, \ldots, m$$

where $\bar{\mathbf{x}}_j$ is the external embedding for the word j and the vectors $\tilde{\mathbf{x}_i}$ are used merged with the internal embeddings.
In this work we experimented three different settings: (1) *only external*, (2) *mix sum*, (3) *mix gate*. While *only external* is the setting we have just described above, the other two settings combine the external embeddings with the internal NMT embeddings. The *mix sum* setting inserts a vector sum between the embeddings and the RNN which simply sums the internal embedding for the word f_j and the mapped external embedding for the same word:

$$\hat{\mathbf{x}}_j = \mathbf{x}_j + \tilde{\mathbf{x}}_j$$

In the *mix gate* setting, we let the network learn parameters to combine the internal and the external embeddings. A gate is a function that produces a vector of the same dimensionality of the input, with all elements between 0 and 1 to represent the proportion of the corresponding input element that is propagated to the following layer:

$$\mathbf{z}_j = \sigma([\mathbf{x}_j; \tilde{\mathbf{x}}_j]^\top \mathbf{W_z} + \mathbf{b_z})$$

where $\mathbf{z}_j$ is the output of the gate and σ is the sigmoid function. The new vector is produced by combining linear transformations of the inputs with the gate $\mathbf{z}_j$:

$$\hat{\mathbf{x}}_j = \tanh(\mathbf{z}_j \odot ff_1(\mathbf{x}_j) + (1 - \mathbf{z}_j) \odot ff_2(\tilde{\mathbf{x}}_j))$$

Where ff is a feed-forward layer. In this setting the network has more parameters to learn for combining the internal and external embeddings in an effective way.

4 Experimental setup

Model	TED-14
Baseline	25.37
Only External Crawl	26.13
Mix Sum Crawl	**29.45**
Mix Gate Crawl	27.10

Table 1: Small data condition: BLEU score on IWSLT TED Talk Task En-Fr.

We performed our experiments on two tasks representing two different training conditions:

Model	NEWS-15	NEWS-16
Baseline	**16.67**	**20.07**
Only External Crawl	12.73	15.58
Mix Sum Crawl	15.59	18.72
Mix Gate Crawl	16.20	19.44
Only External news	13.36	16.40
Mix sum news	16.01	19.15
Mix gate news	16.45	19.35

Table 2: Large data condition: BLEU scores on WMT News Task En-De.

large data and *small data*. The first task is the 2017 WMT News translation task, from English to German, which provides a substantial amount of parallel data. For this experiment, we use all the available training data, about 5 million sentence pairs[1], newstest2013 and 2014 as a validation set and newstest2015 (NEWS-15) and newstest 2016 (NEWS-16) as test sets. The second task in the 2016 IWSLT TED Talk translation task, from English to French, for which we only deployed a small in-domain data set consisting of 200,000 sentence pairs, dev and test sets from 2010 to 2013 as a validation sets and the test set 2014 as test set (TED-14) [2].

We used two sets of pre-trained English word-embeddings. The first is the Common Crawl set available from the GloVe website[3], which contains $1.9M$ word embeddings (dim=300) trained with Glove (Pennington et al., 2014). The second set was instead created by us with *fast-Text* (Bojanowski et al., 2016) from the newscrawl 2015 and 2016 corpora (also available from the WMT 2017 website), which can be considered in-domain for the wmt task. We selected only words appearing at least 5 times in the corpus, and did not use any character n-gram information. This process produced embedding vectors (dim=500) of about $640K$ words in the news domain.

For all the experiments we used an NMT with 500 dimensions in the embeddings and in the hidden sizes of RNN. With the WMT dataset we used vocabularies of size $40,000$ in both sides. They are words in the source side and sub-words in the target side. For IWSLT we used $80,000$ words vocabularies, which cover more than 99% of the training set vocabulary. For the training we applied Adam (Kingma and Ba, 2014) with initial learning rate 0.0003 until convergence. As a codebase we used Nematus (Sennrich et al., 2017) for all of our experiments. The reported BLEU scores (Papineni et al., 2002) are computed with multi-blue.pl from the Moses suite on detokenized texts. The results are presented in Tables 2 and 1.

Table 3: Out-of-vocabulary words in internal and external vocabularies

	TED	News15	News16
Int	289	556	738
Ext Crawl	1581	4460	6532
Ext News	-	487	394
Both Crawl	176	477	625
Both News	-	352	285

	NEWS15	NEWS16	TED
Baseline	14	15	337
Only Ext. Crawl	10	8	687
Mix Sum Crawl	64	77	672
Mix Gate Crawl	102	337	689
Only Ext. News	10	7	-
Mix Sum News	132	335	-
Mix Gate News	85	117	-

Table 4: Numbers of generated unknown words in the translations.

5 Results and Discussion

Results show that our approach is greatly beneficial in our small data condition (table 1), improving up to 4 bleu scores with the simple strategy of summing the external and internal word embeddings. For the large-data condition (table 2) the picture is instead very different, as none of the settings using external embeddings reaches the results of the baseline.

In order to verify our hypothesis that external embeddings help to extend the vocabulary, we firstly counted the number of OOV words with respect to the internal and external vocabularies for each test set, and also the number of words that are unknown in both of them. The results listed in table 3 show that in the case of TED, the number of OOVs in both vocabularies is 39% smaller than in the internal vocabulary, but at the same time in the external vocabulary it is more than 5 times larger. In all the experiments, the embeddings trained on Gigacrawl have many more OOVs than the inter-

[1] http://www.statmt.org/wmt17/translation-task.html

[2] https://wit3.fbk.eu/mt.php?release=2016-01

[3] https://nlp.stanford.edu/projects/glove/

src	so I was trained to become a gymnast for two years in Hunan , China in the 1970s .
ref	J’ai été entraînée pour devenir gymnaste pendant 2 ans , dans la province d’ Hunan en Chine dans les années 1970 .
baseline	J’ai été formée pour devenir gymnaste , pendant deux ans au Texas, en Chine dans les années 1970 .
mix-gate	J’ai donc été formé pour devenir une gymnaste pendant deux ans en UNK, en Chine dans les années 70.
src	Egyptologists have always known the site of Itjtawy was located somewhere near the pyramids of the two kings [...] .
ref	les égyptologues avaient toujours présumé qu’ Itjtawy se trouvait quelque part entre les pyramides des deux rois [...] .
baseline	Nous avons toujours connu le site de Londres , situé quelque part prés des pyramides des deux rois [...]
mix-gate	Et on sait toujours que le site de UNK était situé quelque part près des pyramides des deux rois [...].

Table 5: Example translations with words that are out of one of the two vocabularies. In the first sentence “China” is not in the external vocabulary, but it is still trained properly. In the second sentence “Egyptologists” is not in the internal vocabulary. It cannot be translated at all, but the network finds a way to come around the problem.

nal counterpart, and the difference is particularly large in newstest15 and 16. This can be a reason for degradation of representations, unless the network learns to correct the noise coming from the external side.

To have a glimpse of the degradation, we also counted the number of generated unknown words for each test set. The results are listed in table 4. What we can observe is a slightly reduced number of unknown tokens in newstest when using only the external embeddings, but in a setting where the target side uses subwords. In all the other cases, the number of unknown words during translations increases dramatically. The increase is from 5 to 22 times in WMT and about 2 times in TED. Now we want to understand if this is due to a corrupted representation of words, which mixes good embeddings with the external embedding for the unknown token, or the reason is to find somewhere else. This is particularly true because of the contemporary improvement in BLEU score.

To verify the correction capabilities of the network, we check some translations where one word is missing in one of the two vocabularies. Two example translations are shown in table 5. In the first example, the word “China” exists only in the internal vocabulary, but it’s correctly translated also by the mix-gate system. Furthermore, the baseline translates the OOV word “Hunan” with “Texas”, while our system translates it with an unknown token. The second behavior is surely one of the main reasons of the increased number of generated unknown words using external embeddings, and it is also preferable as there are methods for replacing the unknown tokens in a postprocessing step. (Luong et al., 2015b).

In the second example, “Egyptologists” is OOV for the internal vocabulary. Lacking the subject, the baseline resorts to the first person plural, and it also adds a subordinate sentence that change s the meaning with respect to the source. Moreover, again an unknown word for a location is translated with another word that is related with the source only because it is another location (in this case the system translates with “Londres”, which is the French word for “London”). By contrast, in absence of more information about the subject, the mix-gate uses the impersonal form and the grammar of its translation is better in general.

In the large-data setting, the best system using external embeddings is the mix-gate with data from the news domain. From table 3, we can relate the improvement also to the reduced number of external OOV words, but the improvement is so small that we suppose that using better corpora is not a path to follow. Moreover, our results lower than the baseline are an empirical proof that pre-trained embeddings are not useful when there are large parallel data available.

6 Conclusions

In this paper we propose three methods to extend the input word embeddings to an NMT network in order to leverage a word representation coming from a big monolingual corpus. Our results show that this approach greatly improves over an NMT baseline in a low-resource scenario, while it is not helpful for better-resourced tasks.

Using monolingual data for improving NMT is a problem also in the latter case, thus our future work will focus on how to integrate models larger than word embeddings, and trained on monolingual data, to improve word and sentence representations.

Acknowledgments

This work has been partially supported by the EC-funded projects ModernMT (H2020 grant agreement no. 645487) and QT21 (H2020 grant agreement no. 645452). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs used for this research.

References

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the fourth workshop on statistical machine translation*, pages 182–189. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation (WMT)*, volume 2, pages 131–198.

M. Cettolo, J. Niehues, S. Stker, L. Bentivogli, and M. Federico. 2016. The IWSLT 2016 evaluation campaign. In *Proceedings of the 13th Workshop on Spoken Language Translation, Seattle, pp. 14, WA*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015a. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Minh-thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine translation. In *In ACL*. Citeseer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

Distributed Processes for Spoken Questions and Commands Understanding

Dario Di Mauro
University of Naples “Federico II”

Antonio Origlia
University of Padua

Francesco Cutugno
University of Naples “Federico II”

{dario.dimauro,cutugno}@unina.it antonio.origlia@dei.unipd.it

Abstract

Commercial products labelled as smart devices usually recur to a centralised system that processes all the requests. A distributed model, where nodes independently interact with the environment, may provide a widespread support for both users and other devices. In the latter setup, each entity has a partial awareness about defines the requests accepted by the network, and this aspect complicates the task. This paper improves an existing distributed model, called PHASER, by proposing linguistic analysis techniques to manage non-matching requests. NLP methods produce a confidence; any PHASER node forwards non-matching requests to close peers. PHASER exploits the confidence to rank the adjacent peers and deliver the question to the best node. Partial Matching and a *Bag-of-Words* models will be compared with the currently adopted full matching. The Bag-of-Words approach offered the best results in terms of both quality and required time.

I prodotti commerciali, etichettati come dispositivi intelligenti, di solito usano un sistema centralizzato per processare le richieste. Un modello distribuito, dove i nodi interagiscono indipendentemente con l’ambiente, può fornire un supporto più ampio per gli utenti. Nel secondo setup, ogni entità è parzialmente a conoscenza delle richieste accettate da ogni nodo del network. Questo lavoro si propone di migliorare un modello distribuito esistente, chiamato PHASER, ricorrendo a tecniche di analisi linguistiche per gestire le richieste non accettate localmente; ogni nodo PHASER inoltra queste richieste ai nodi adiacenti. Metodi di NPL producono una confidence*; PHASER la sfrutta per ordinare i nodi vicini e inoltrare la richiesta al migliore. Modelli basati su* partial matching *e* bag-of-words *saranno confrontati con il sistema attualmente adottato, basato su* full matching*. Dal confronto,* bag-of-words *ha riportato i risultati migliori sia di qualità che di tempo necessario per la risposta.*

1 Introduction

This paper introduces a new distributed approach to question answering and command execution in different Intelligent Environments (henceforth IE). This idea is rarely encountered in literature (with a few exceptions by Surdeanu et al. (2002)). IE is a new discipline including Domotic, Internet of Things, Cultural Heritage Technological Innovation and other similar issues. In our approach, devices in the environments constitute nodes of a network and this network provides services to an interacting user. Reasons of interests for NLP studies in this kind of application lie in the idea that user requests are delivered using Natural Language (mainly speech) in the simplest case, or multimodally by integrating speech with gestures and interaction with physical controls.

Nowadays, smart devices commonly propose interaction through natural language to the users. As the services offer becomes wider, a network of specialised applications managing very specific domains is masked behind a single named character (Alexa, Siri, Cortana). While this is common for single devices hosting multiple applications, which inform the operating system about their capabilities through dedicated languages (e.g. SRGS[1], Hunt and McGlashan (2004)), the

[1] https://www.w3.org/TR/speech-grammar/ retrieved on October 2017

model is being transferred to networks of different devices. Optimising the communication between these devices is a critical issue to reduce response times, balance communication and improve quality of service. In this paper, we will concentrate on showing how the *confidence* metric, commonly available in NLP techniques, can be exploited to support rapid adaptation of the network to request dispatching, avoiding broadcasting.

We will mainly describe the simplest case of speech interaction and understanding; for a view on the complex multimodal approach, please refer to Valentino et al. (2017). Given these premises, nodes in the given network are able to respond to simple questions or commands uttered by the user. In a first approach, utterances should conserve a coherence with the "nature" of the node, i.e. if I am "talking" to the kitchen or to the microwave oven, I should make requests strictly inherent with the device functions. On the contrary, we wish to expand the "intelligence" of the environment giving to the user the possibility to make any kind of requirements to any node in the network. In this view, each node is able to classify the string deriving by the speech utterance assigning it to one of the many classes of relevant action the environment can realise, even if the node itself is not able to execute that action. The introduction of a distributed knowledge base and of network information spreading techniques concur to the realisation of an environment extremely reactive, scalable and easily configurable for different domains. The system is reactive as the network connections are strongly optimised: redundant and rarely used paths are pruned. Mechanisms for knowledge distribution are optimised in order to deliver the proper answer minimising network reaction times. The system is easily configurable to different domains as this kind of networks just require a formal description of the semantics of each node, of the action classes they are able to process, and of the most probable connection among nodes that *a-priori* the environment designer implements. In order to realise this system, many NLP software modules are needed, and among these: an automatic spoken dialogue manager, a Spoken Language Understanding system, an ontology modelling the environment and the devices. An extended description of each part can be found in Di Mauro et al. (2017); this paper focuses on a linguistic analysis to improve the navigation of the request through the network.

In Section 2 we present related works; Section 3 recalls the model of our system. In Section 4 we discuss a network of interactive entities, highlighting differences about the current version and the contribution of this paper. Experiments and Discussion are presented in Sections 5 and 6. Section 7 concludes the paper.

2 Related works

Our idea is to provide a distributed network of entities, where each node interacts with the user through multimodal interaction. Knowledge is local to the node and limited to the provided services. If the node is not able to produce the expected output for a request, it sends the message to others in the network, without a prior determined target node. The intelligence perceived by the users is built upon a collection of partial nodes' intelligence. This system, called PHASER, has been firstly introduced by Di Mauro et al. (2017).

Distributed approaches for Human-Computer Interaction have been widely discussed in literature. Multi-Agent Systems have been applied to smart environments by Li et al. (2016), Pajares Ferrando and Onaindia (2013) ; their work is based on the discovery of semantic resources and orchestration, with negotiation between user and devices. Valero et al. (2016) proposed a system with multiple users, various roles and access policies.

The goal of this work is to provide a strategy to better rank close nodes according to the exposed information about the accepted inputs. By considering the Navigation problem from a Question/Answering (Q/A) point of view, PHASER could be theoretically compared with distributed Q/A systems (Surdeanu et al., 2002). Q/A systems do not collect entire documents, but they extract just short and relevant information to produce an answer. Since the documents are not all physically stored on the same server, a distributed Q/A system deals with parallel tasks and load balancing. Even if some similarities with PHASER can be considered, the main difference is that a node ends its own work as it delivers the message.

Baeza-Yates et al. (1999) stated that the *Ranking problem* is fundamental in Information Retrieval. It can be solved with machine learning as summarised by Liu and others (2009). However, adopted processes usually manage many documents; this is not a realistic case of PHASER, where the rank is provided relying on little information.

3 Model

In IE the term *Intelligent* usually refers to Artificial Intelligence (AI) applied to environments, where technology offers more than static rooms as introduced by Augusto et al. (2013). In this Section we propose our method, a Pervasive Human-centred Architecture for Smart Environmental Responsiveness (PHASER): it is a distributed solution for an IE which provides a ubiquitous environment. The global intelligence is built upon single entities that show responsive behaviours and collaborate with each other to better support the user. PHASER has been firstly presented by Di Mauro et al. (2017). This Section recalls general aspects of our model: the description of what each node represents and how it constitutes a network with similar entities.

3.1 A Smart Entity

In our concept, PHASER gives a role to each entity who interacts with the others. Possible entities are *objects* and *people* interacting with those objects. We make use of an abstract concept of *node* to include the needs of both the entities. In real scenarios, objects are AI-powered devices, considered an important step on an evolutionary process that is affecting modern communication devices (Atzori et al., 2014; Lòpez et al., 2012). People, instead, are represented with their personal smartphone which acts as an interface.

Each object interacts with other connected entities, providing services and responding to questions. A graph results and objects individuate its nodes. For this reason, we will refer to objects as nodes as well.

3.2 Model of PHASER

A single node represents an entity in the environment. We formally define a node as a tuple:

$$N(\iota, Cnf_\iota, Close_\iota, Discovered_\iota, oBC_\iota)$$

where ι is a unique identifier of the node in the environment and $Close$ is a set of related nodes in the environment. $Discovered$ collects nodes connected after unforeseen interactions. $Close$ and $Discovered$ contain identifiers of the remote nodes. ι establishes connections and interacts with nodes in both the sets. Each node specifies a configuration Cnf, which determines ι's role in the environment. The configuration comprises inputs, outputs and how it reacts to network events. In details:

$$Cnf_\iota = (name_\iota, type_\iota, class_\iota, env_\iota, I_\iota, O_\iota, P_\iota)$$

where *type*, *class* and *env* classify ι according to an ontology, while *name* labels it. I and O represent inputs and outputs respectively; they divide data into channels as in Equation 1 for multimodal interaction, where c_x is a channel code and $RG_{c_x} = \{r_{i_1}, r_{i_2}, \ldots, r_{i_{c_x}}\}$ is a set of regular expressions. If N_{i_ι} and N_{o_ι} are the number of input and output channels, we define I_ι and O_ι in Equation 2.

$$Ch_j = (c_j, RG_{c_j}) \qquad (1)$$

$$I_\iota/O_\iota = \bigcup_{1 \leq x \leq N_{i_\iota/o_\iota}} \{Ch_x\} \qquad (2)$$

Input and Output compose the Business Card (BC) and it represents what a node may accept; each object exposes its own BC to the connected nodes; received BCs will be stored in oBC. The approach discussed in this work ranks $Close$ and $Discovered$ peers by obtaining confidences from their BC. PHASER nodes compose the network. There is not a hierarchic organisation, so all the nodes are at the same level. The network does not need a specific topology, but we assume that an expert of the considered domain designs it.

The presented formalism defines a PHASER node which establishes a connection towards other similar entities. This is the core part of our system: a distributed model where single peers interact with people - through I/O modules - and with others. Input and Output modules are intentionally generic because each node can have a customised the interaction. This approach aims at supporting Natural User Interfaces (Wigdor and Wixon, 2011).

The discussed formulae are the core part useful to understand the introduced improvements. A detailed description of the PHASER model has been provided by Di Mauro et al. (2017).

4 Navigation Problem

In PHASER, each node is expected to have a knowledge, circumscribed to its own domain: a fridge should understand questions about food or

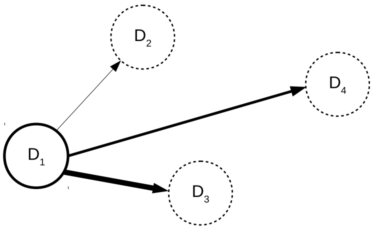


Figure 1: The thickness of each arc is proportional to the probability of $D_{2,3,4}$ to accept the received request

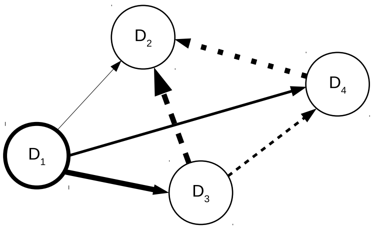


Figure 2: A command X from D_1 reaches D_2 via D_3 or D_4. Arcs' thickness is the match percentage of X; the dashed line concerns the path's length

ingredients; commands about lighting and heating are *out-of-context*. However, the environment could contain other devices to manage those requests. In such a configuration, a network of PHASER devices is populated of entities with partial knowledge, where nodes have a common strategy to propagate *out-of-context* requests.

Each node is able to interact with users. It means that each of those entities may manage commands for any other node in the network. They could handle non-matching requests in two ways: *(i)* broadcast them to all connected nodes or *(ii)* individuate the most probable nodes. The second approach is preferred in very active networks - i.e. smart museums - or where a large percentage of nodes processes unknown requests; moreover, the first approach easily overloads the network. This *"Navigation problem"* aims at forwarding the request to the best candidate; the model follows a *depth-first-search* by iterating on a sorted list of close nodes. The ranking is obtained with a greedy approach. The navigation continues until a node finds a response or all the sub-network is explored. Figure 1 reports a graphical representation of the sorting phase that is performed in the navigation presented in Figure 2. D_1 chooses the next node in the interaction by sorting the adjacent vertices.

This work analyses how the discussed probability is obtained, presenting the current technique, and investigating improvements that support partial matching to tailor the propagation on the network. The outcome is a different rank of the list of close nodes explored in the navigation of the graph.

4.1 Perfect matching

A network of PHASER nodes starts from a topology designed by experts. During the interaction, the network adapts connections to maximise the local utilities on each arc. As it is set up, each pair of connected nodes shares a business card. It comprises a set of active channels and, for each channel, a set of regular expressions (regex) for the accepted inputs.

The *"Navigation problem"* is solved by sorting the adjacent nodes according to their matching with the exposed regular expressions. Nodes with higher value of matching will be firstly called in forwarding. Inputs can be on multiple channels, so the matching complies with the structure. In this version, M calculates the value of matching as defined in Equation 3:

$$M(R, n) = \sum_{0 \leq i < |R|} m(R_i, n)/|R| \tag{3}$$

$$m(R_x, n) = \begin{cases} 1 & \text{if } R_x \text{ is valid for } n \\ 0 & otherwise \end{cases} \tag{4}$$

where n is the considered device, R is the request, divided into $|R|$ channels. The expression "R_x is valid for n" of the Equation 4 means that exists a regular expression of n that matches with R_x. The higher M is, the higher is the probability that n accepts R. $M(R, n) = 1$ means a perfect match.

4.2 Imperfect matching

The currently adopted approach is based on full matching where the outcome of each $m(x, n)$ is dichotomous. The calculated value M is then normalised to the size of R - involved channels in R -. This approach highly depends on the accuracy of the design of the set of regular expressions. Moreover, generic regexs - i.e. ".*" - accepts everything. This case undesired, as if a node accepts this input it will attract many requests with the consequence of not being able to process all of them; this would create a *black hole*, that uselessly overloads the network.

Alternative approaches perform linguistic analyses of the received question. The investigated solution is based on partial matching; it provides a *confidence* of the input, used to refine the ranking of the adjacent nodes. The improvement still must prefer a perfect matching, but it does not completely exclude the opposite case. Then, we propose a revised version of the formulae seen in Section 4.1 by introducing $m_{l_x}^{v}$ as the confidence of v on channel x and adapting M as follows:

$$M(R,n) = \prod_{0<x\leq|R|} max\left\{\left(m_{i_1}^{R_x}, \ldots, m_{i_n}^{R_x}\right)\right\} \tag{5}$$

The function in Equation 5 supports multiple channels and a set of possible grammars for each of them, but $m_{l_x}^{v}$ is now the probability that the token v from the request is accepted on an input l_x. This probability can be calculated with two strategies: regex-based and *bag-of-words* (BoW). The former approach calculates the longest substring that matches on each provided regex on the proper channels; this obtained length is then normalised on the total length of the request. The *bag-of-words* method, instead, splits both the request and the stored accepted inputs in two bags of words - B_{req} and B_{input} respectively - and calculates how many words of the request match on the total set. This value is then normalised on $|B_{input}|$. Both the strategies are locally performed by nodes on received questions that must be forwarded. No global dictionaries are saved in order to maintain a scalable distributed system where each node has partial knowledge about the others.

Since any NLP approach provides a confidence of the evaluated input request, other strategies have been considered. However, these approaches present drawbacks that will be discussed in Section 6.

5 Experiments and Results

This Section reports experiments conducted to compare the three discussed approaches in PHASER: perfect matching, partial matching, and BoW. Full and partial matching methods rely on regular expressions and assess how much the request matches the provided regexs. The system has been tested by simulating a smart house with 5 networked PHASER nodes. The considered nodes are TV, Microwave Oven (M), Fridge (F), Kettle (K), Alarm clock (A).

We considered a star-like network with TV in the middle. We tested two kinds of configurations for input representation. A request is delivered to the TV, which forwards the request to the node with the highest confidence; this is obtained with the different approaches. The network is design to let *Kettle* being the winner.

Table 1 collects data where inputs are represented with a BoW style; in Table 2, instead, inputs are represented as regular expressions. Each node used OpenDial by Lison and Kennington (2016) to manage a dialogue.

command	*perfect*	*partial*	*BoW*
prepare a tea	K (1.0)	K (1.0)	K (1.0)
warm	**A (0.1)**	M (0.44)	M (0.5)
warm water	**A (0.1)**	**A (0.1)**	K (0.667)
wake me	**A (0.1)**	A (0.438)	A (0.5)

Table 1: Winner device and confidence for each request. Each node had a bag-of-word style inputs. Bold cells refer to unsuccessful evaluations

command	*perfect*	*partial*	*BoW*
prepare a tea	K (1.0)	K (1.0)	**K (0.1)**
warm	**A (0.1)**	M (0.44)	K (0.33)
warm water	**A (0.1)**	A (0.9)	K (0.667)
wake me	**A (0.1)**	A (0.778)	A (0.4)

Table 2: Winner device and confidence for each request. Each node had a regex style inputs. Bold cells refer to unsuccessful evaluations

6 Discussion

The presented process operates in a context where the current node n is not able to understand the request r and it prefers to share it with the network, refraining from broadcasting. The node n gathers a confidence on r to sort the adjacent nodes, preferring nodes with higher values. A sequence results, where the first node is the best candidate to accept the request.

Results show that a *full matching* is not always a good choice. It requires a precise design of each regex, exposing the structure of accepted inputs; moreover, this strategy does not always discriminate different nodes and fails in many cases. *Partial matching* provides finer values and nodes are better sorted. However, this approach easily creates *black holes*, nodes that attract many inputs because of a wrong design. The BoW model gave the best results with two benefits: *(i)* the network is easier to design; *(ii)* each node could share unstructured data, improving local security.

Other strategies have been investigated. We considered more refined systems based on SRGS; however, this method has been excluded for many reasons: *(i)* SRGS requires a complete grammar from adjacent nodes and this may generate security issues because they expose a detailed structure of accepted inputs; *(ii)* grammar-based methods introduce overheads compared with the adopted approaches, due to the engine needed to recognise the request on the model represented by the grammar.

7 Conclusions

This paper presented PHASER, a distributed model for Human-Computer Interaction in Intelligent Environments. This work aims at improving the *Navigation Problem*, where a node forwards a received command if it is not able to understand or process it. Since the node operates with partial knowledge about both the request and the environment, it tries to analyse the input and choose the best adjacent node.

The most crucial part is not a refined linguistic analysis of each request, but a quick confidence on how much each adjacent node could be a good candidate to understand that request. This requirement is motivated by two reasons: *(i)* this process is part of a longer step where a user is waiting for a response; *(ii)* all the evaluations rely on information each node shares with others. In order to deeply understand the command, the node should expose sensible data and it is not always desired in a distributed context.

The work focused on three strategies: perfect and partial matches with regular expressions and a bag-of-words model. This last approach has given the best results with positive aspects mainly related to easy network design and security of each node. The investigated methods are just used to rank close peers on as an *out-of-context* request reaches the current node. It operates without understanding the request, so finer considerations are not possible. The considered approaches do not limit PHASER nodes in adopting more refined techniques in assessing and categorising an input request.

References

L. Atzori, A. Iera, and G. Morabito. 2014. From “smart objects” to “social objects”: The next evolutionary step of the internet of things. *IEEE Communications Magazine*, 52(1):97–105.

J. C. Augusto, V. Callaghan, D. Cook, A. Kameas, and I. Satoh. 2013. Intelligent environments: a manifesto. *Human-Centric Computing and Information Sciences*, 3(1):1–18.

R. Baeza-Yates, B. Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

D. Di Mauro, J. C. Augusto, A. Origlia, and F. Cutugno. 2017. A framework for distributed interaction in intelligent environments. In *European Conference on Ambient Intelligence*, pages 136–151.

A. Hunt and S. McGlashan. 2004. Speech recognition grammar specification version 1.0. *W3C Recommendation*.

W. Li, T. Logenthiran, W. L. Woo, V. T. Phan, and D. Srinivasan. 2016. Implementation of demand side management of a smart home using multi-agent system. In *IEEE Congress on Evolutionary Computation*, pages 2028–2035.

P. Lison and C. Kennington. 2016. Opendial: A toolkit for developing spoken dialogue systems with probabilistic rules. *ACL 2016*, page 67.

T. Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3:225–331.

T. S. Lòpez, D. Ranasinghe, M. Harrison, and D. McFarlane. 2012. Adding sense to the internet of things: An architecture framework for smart object systems. *Personal and Ubiquitous Computing*, 16(3):291–308.

S. Pajares Ferrando and E. Onaindia. 2013. Context-aware multi-agent planning in intelligent environments. *Information Sciences*, 227:22 – 42.

M. Surdeanu, D.I. Moldovan, and S.M. Harabagiu. 2002. Performance analysis of a distributed question/answering system. *IEEE Transactions on Parallel and Distributed Systems*, 13(6):579–596.

M. Valentino, A. Origlia, and F. Cutugno. 2017. Multimodal speech and gestures fusion for small groups. In *Workshop on Designing, Implementing and Evaluating Mid-Air Gestures and Speech-Based Interaction*. in press.

S. Valero, E. del Val, J. Alemany, and V. Botti. 2016. Enhancing smart-home environments using magentix2. *Journal of Applied Logic*. in press.

D. Wigdor and D. Wixon. 2011. *Brave NUI world: designing natural user interfaces for touch and gesture*. Elsevier.

A Reproducible Approach with R Markdown to Automatic Classification of Medical Certificates in French

Giorgio Maria Di Nunzio
Dept. of Information Engineering
University of Padua
giorgiomaria.dinunzio@unipd.it

Federica Beghini,
Federica Vezzani, Geneviève Henrot
Dept. of Linguistic and Literary Study
University of Padua
fede.beghini92@gmail.com
federica.vezzani@phd.unipd.it
genevieve.henrot@unipd.it

Abstract

English. In this paper, we report the ongoing developments of our first participation to the Cross-Language Evaluation Forum (CLEF) eHealth Task 1: "Multilingual Information Extraction - ICD10 coding" (Névéol et al., 2017). The task consists in labelling death certificates, in French with international standard codes. In particular, we wanted to accomplish the goal of the 'Replication track' of this Task which promotes the sharing of tools and the dissemination of solid, reproducible results.

Italiano. *In questo articolo presentiamo gli sviluppi del lavoro iniziato con la partecipazione al Laboratorio Cross-Language Evaluation Forum (CLEF) eHealth denominato: "Multilingual Information Extraction - ICD10 coding" (Névéol et al., 2017) che ha come obiettivo quello di classificare certificati di morte in lingua francese con dei codici standard internazionali. In particolare, abbiamo come obiettivo quello proposto dalla 'Replication track' di questo Task, che promuove la condivisione di strumenti e la diffusione di risultati riproducibili.*

1 Introduction

When researchers use 'traditional' methods of scientific publication to describe computational research, we, as readers and researchers, may incur into the so-called 'reproducible research' problem (Schwab et al., 2000). For example, a traditional conference paper usually specifies the relevant computations of the main approach, because the limitations of a paper medium prohibit a complete documentation, which would ideally include experimental data, parameter values, and the source code of the program. Those readers who wish to use the same approach of the paper, hence reproduce the results, must reimplement the whole process, which sometimes may be an unfeasible task. The extreme of reproducibility is 'replicability', i.e. a perfect replica of a scientific experiment. The discussion of the difference between replicability and reproducibility is beyond the scope of this paper (Drummond, 2009), and we will just point out that, in general, even in the most accurate replica of an experiment will be done by a different person, in a different lab, using different equipment. Researchers of different areas have identifyied the necessity for reproducibility, or reproducible research, as an attainable minimum standard for assessing the value of scientific claims (Peng, 2011). As Roger Peng suggests, "one aim of the reproducibility standard is to fill the gap in the scientific evidence-generating process between full replication of a study and no replication. Between these two extreme end points, there is a spectrum of possibilities, and a study may be more or less reproducible than another depending on what data and code are made available".

Reproducibility matters because the lack of reproducibility in science causes significant issues for science itself, for other researchers in the community, and for public policy. For example, *Nature* published a special issue about "Challenges in Irreproducible Research"[1] where the examined cases showed that there is

> [. . .] a growing alarm about results that cannot be reproduced. Explanations include increased levels of scrutiny, complexity of experiments and statistics, and pressures on researchers. Journals, scientists, institutions and funders all

[1] https://goo.gl/5SxYQJ

have a part in tackling reproducibility.

Among many other problems, the article showed that most of the drug validation studies (43 out of 67 studies) failed to reproduce. Another important case concerned *Science*, where the Editor-in-Chief retracted in 2015 a study of how canvassers can sway people's opinions about gay marriage because: " (i) Survey incentives were misrepresented [. . .], (ii) The statement on sponsorship was false. [. . .]" [2] There are also cases of papers retracted by authors themselves because "After carefully reexamining the data presented in the article, they identified that data of two different hospitals got terribly mixed. The published results cannot be reproduced in accordance with scientific and clinical correctness." as declared in the note of retraction of the paper "Low Dose Lidocaine for Refractory Seizures in Preterm Neonates" (Chakrabarti et al., 2013).

1.1 Reproducible Research in IR and NLP

The problem of reproducibility in Information Retrieval (IR) has been addressed by many researchers in the field in the last years (Ferro et al., 2016b; Ferro, 2017; Neveol et al., 2016). Despite the fact that IR has traditionally been very rigorous about experimental evaluation (the Text REtrieval Conference TREC celebrated the 25th edition in 2016[3]), many researchers raised some concerns about reproducibility in IR, which are related to system experiments (or runs); in fact, even if a researcher uses the same datasets and the same open source software, there are many parameters and variables hidden in the vode that make the full reproducibility of the runs very difficult. For this reason, there are important initiatives in the main IR conferences that support this kind of activity, see for example the open source information retrieval reproducibility challenge at SIGIR[4] or the Reproducibility track at ECIR (Ferro et al., 2016a)), as well as some Labs at the Cross-Language Evaluation Forum (CLEF) that explicitly have a task on reproducibility, such as CLEF eHealth[5].

The Natural Language Processing (NLP) community has witnessed the same problem. In 2016, the workshop "Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language" at the Language Resources and Evaluation Conference (LREC) encouraged the discussion and the advancement on the reproducibility of research results and the citation of resources, and its impact on research integrity in the research area of language processing tools and resources. The workshop gathered authors interested in discussing the challenges, the risk factors, the procedures that should be adopted including the new risks raised by the replication articles themselves and their own integrity, in view of the preservation of the reputation of colleagues.

1.2 Contribution

In this paper, we report the current developments of our first participation to the CLEF eHealth Lab (Goeuriot et al., 2017), in particular to Task 1: "Multilingual Information Extraction - ICD10 coding" (Névéol et al., 2017). The task consists in labelling death certificates with standard codes, the International Classification Diseases codes (ICD10). In particular, we wanted to accomplish the goal of the 'Replication track' of this task which promotes the sharing of tools and the dissemination of solid, reproducible results (Di Nunzio et al., 2017). Participants of this track had to submit their systems used to produce the experiments, or a remote access to the system, along with instructions on how to install and operate the system. The replication track involved analysts that attempted to replicate a team's results by running the system supplied on the test data sets, using the team's instructions.

Therefore, our main objective was to build a modular system that can be easily enhanced in order to make use of the cleaned training data available and to build a reproducible set of experiments of a system that i) converts raw data containing death certificates into a cleaned dataset, ii) implements a set of semi-manual rules to split sentences and translate medical acronyms, and iii) implements a lexicon based classification approach with the aim of building a sufficiently strong baseline (our initial objective was to achieve a classifier performance close to 50%). For this purpose, we devised a pipeline for processing each death certificate and producing a 'normalized' version of the text that will be presented in the following sections.

[2] https://goo.gl/NWA5gK

[3] http://trec.nist.gov

[4] https://goo.gl/CePVzY

[5] https://goo.gl/WgkqnZ

2 R for Reproducible Research

A Tutorial given during the UseR! 2017 conference entitled "Data Carpentry: Open and Reproducible Research with R"[6] presented an overview of the problems related to (the lack of) reproducible research and the possible solutions in particular when programming with the R Language. In the field of Data Science, the R Markdown framework[7] is considered one of the possible solutions to document the results of an experiment and, at the same time, reproduce each step of the experiment itself. Following the indications given by (Gandrud, 2015) and the suggestions discussed by (Cohen et al., 2016), we developed the experimental framework in R and publish the source code on Github[8] in order to allow other participants to reproduce our results. In particular, in this paper we will focus on the classification of death certificates in French, a part of the work that was partially presented as non-official experiments in the original paper (Di Nunzio et al., 2017).

2.1 Dataset

The CèpiDc corpus was provided by the French institute for health and medical research (INSERM) for the task of ICD10 coding in CLEF eHealth 2017 (Task 1). It consists of free text death certificates collected from physicians and hospitals in France over the period of 2006-2014 (Névéol et al., 2017). Indeed, death certificates are standardized documents filled by physicians to report the death of a patient, but the content of each document contains heterogeneous and noisy data that participants had to deal with (Kelly et al., 2016). For example, some certificates contain non-diacritized text, or a mix of cases and diacritized text, acronyms and/or abbreviations, and so on. In Table 1, we show an example of a death certificate of the training set (the English version) split in three lines, Table 1a, and its correct classification with the ICD10 codes, Table 1b. In this case, the last line of the death certificate should be classified with two ICD10 codes (I64 related to acute cerebral issues, and G20 related to Parkinson's disease). In Table 1c, we show an example of a French death certificate aligned with the cause of death and the 'standard' clean text. In both cases, there are issues related with misspellings: the word 'atrial' has been written as 'atrail', as well as many diacritics missing in the French raw text (hemorragie instead of hémorragie).

2.2 Pipeline for Data Cleaning

In order to process the raw death certificate and produce a clean dataset, we implemented the following pipeline for data ingestion: read a line of a death certificate, split the line according to a list of expressions (i.e. "dans un contexte de", suite à un[e]", etc.); remove extra white space (leading, trailing, internal); transform letters to lower case; remove diacritics (optional); remove punctuation; expand acronyms (if any); correct common patterns (if any).

The removal of diacritics was surprisingly effective for the French dataset, as discussed in the preliminary experiments (Di Nunzio et al., 2017). For this reason, in this paper we will only show experiments containing this modification. Acronym expansion was also a crucial step to normalize data and make the death certificate clearer and more coherent with the ICD10 codes. For the expansion of French acronyms, we used the Wikipedia page "Liste d'abréviations en médecine"[9] that contains 1,059 options for acronym expansion. After a manual cleaning of the broken/missing/duplicated entries, we produced a table of 1,179 expanded acronyms.

In this paper, we use a simple semi-automatic step to correct misspellings based on the dictionary of ICD10 codes that was not present in the original experiment. In particular, after cleaning the data and expanding the acronyms, we computed the generalized Levenshtein distance[10] between each token of the death certificate and each token of the dictionary. At the end of this process, we found 4,142 tokens having no match (distance greater than zero) with the ICD10 vocabulary. The terms having more than 10 occurrences in the certificates were hard-coded in the source code, while all the others were automatically substituted on-the-fly.

The vocabulary has 6,295 unique entries, and there are 91,953 lines of 31,682 death certificates to classify.

[6] https://goo.gl/soe9i6

[7] http://rmarkdown.rstudio.com

[8] https://goo.gl/coCyAe

[9] https://goo.gl/t41LXn

[10] Given a strings s and t, the Levenshtein distance is the minimal possibly weighted number of insertions, deletions and substitutions needed to transform s into t (so that the transformation exactly matches t).

DocID	YearCoded	LineID	RawText
1	2015	1	PNUEMONIA
1	2015	2	ATRAIL FIBRILLATION
1	2015	6	CVA PARKINSONS DISEASE

(a) Example of death certificate.

DocID	YearCoded	LineID	Rank	ICD10
1	2015	1	1	J189
1	2015	2	1	I48
1	2015	6	1	I64
1	2015	6	2	G20

(b) Example of ICD10 codes for death certificate.

DocID	YearCoded	LineID	RawText	CauseRank	StandardText	ICD10
11	2007	1	hemorragie digestive	1-1	hémorragie digestive	K922
11	2007	2	gastrite	2-1	gastrite	K297
11	2007	5	Pneumopathie , ethylisme chronique , stéatose hépatique	6-1	pneumopathie	J189
11	2007	5	Pneumopathie , ethylisme chronique , stéatose hépatique	6-3	stéatose hépatique	K760
11	2007	5	Pneumopathie , ethylisme chronique , stéatose hépatique	6-2	éthylisme chronique	F102

(c) Example of ICD10 codes for death certificate.

Table 1: Example of death certificate (left) and its correct classification (right) in English Table 1a and 1b. Example of French aligned data in Table 1c.

Table 2: Example of out of vocabulary terms at Levenshtein distance 1.

token	dictionary
alcolique	alcoolique
alcoolo	alcool
artheriopathie	arteriopathie

2.3 Classification rule

The classification of each line of a death certificate uses the approach, proposed by (Eisenstein, 2017), which is performed in the following way: for each line, the score s_i of each entry i of the ICD10 dictionary is computed according to the following sum

$$s_i = \sum_{t_j} w_j \tag{1}$$

which the sum of the weights w_j of each term t_j using binary weighting (one if term present, zero if absent). In those cases where two or more classes have the same score, the first class in the list is assigned by default.

3 Experiments and Results

For the experiments of this paper, we used the 'raw' dataset, that is the portion of dataset where a file records the native text entered in the death certificates (referred to as 'raw causes' thereafter). System performance was assessed by means of a script provided by the organizers of the Lab; the script computes micro-Precision (the fraction of correct instances among the retrieved instances), micro-Recall (the fraction of relevant instances that have been retrieved over total relevant instances), and micro-F1 measure (the harmonic mean between micro-Precision and micro-Recall). As requested by the task, these measures were computed for all causes (FR-ALL) in the datasets and for external causes (FR-EXT), where the evaluation is limited to ICD codes addressing a particular type of deaths, called external causes or violent deaths (see the Task overview for more information (Névéol et al., 2017)).

In Table3, we compare the preliminary results of the non-official French experiments submitted in (Di Nunzio et al., 2017) with our ongoing work on cleaning data that makes use of the semi-automatic approach to correct misspellings and different strategies to split the sentences of the death certificate. In particular, we kept the best performing experiment for all causes named **Unipd-run7** which uses binary weights, automatic creation of expanded acronyms and transliteration (removal) of diacritics. The results show the performances on all causes (FR-ALL) as well as the external causes (FR-EXT).

In the new experiment, we tried to vary the approach of splitting the sentences of a death certificate by: non-splitting the sentence (no-split), using only punctuation characters to split like commas, semi-colon, etc. (simplesplit), and using the same strategy of the original experiment (allsplit). We also tried to use the semi-automatic checkspelling (exp) that uses a mix of manual checking for the most common misspelled words (a misspell that occurs more than 10 times in the dataset) and an automatic substitution for all the remaining misspelled words (partialexp).

The experimental results showed that in all cases we could achieve our initial goal that was a classification performance around 0.50 for the F1 measure; moreover, our approach performed bet-

Table 3: Comparison of results with the best performing unofficial French runs and different approaches to certificate segmentation and semi-automatic spell-checking. The average and median performances of all the experiments of the participants of CLEF eHealth Task 1 are reported at the bottom of the table.

	FR-ALL			FR-EXT		
	Precision	Recall	F1	Precall	Recall	F1
Unipd-run7	0.630	0.468	0.537	0.362	0.251	0.296
Unipd-exp-nosplit	0.645	0.400	0.494	0.438	0.220	0.293
Unipd-exp-simplesplit	0.644	0.456	0.534	0.421	0.233	0.300
Unipd-exp-allsplit	0.645	0.483	0.552	0.393	0.253	0.307
Unipd-partialexp-allsplit	0.646	0.484	0.554	0.409	0.255	0.314
average	*0.475*	*0.358*	*0.406*	*0.367*	*0.247*	*0.292*
median	*0.541*	*0.414*	*0.508*	*0.443*	*0.283*	*0.377*

ter than the average and the median score of all the experiments that were submitted to the CLEF eHealth Task 1. This was a bit of a surprise considering that our classification approach does not use any machine learning approach, but it just cleans the data and assigns the most frequent ICD10 code. This is an encouraging result that sets a solid basis of cleaned data on which we can apply more sophisticated NLP techniques, like those used by the best systems like LIMSI (see (Zweigenbaum and Lavergne, 2017)) which relied upon dictionary projection and supervised multi-class, single-label text classification using dictionaries and token bigram features (Névéol et al., 2017).

4 Final remarks and Future Work

The aim of this work was to continue the work on the reproducible research approach that can be used as a baseline for further experiments. The performance of the system that uses a semi-manual spell-checking approach improved the baseline set by the original paper. The documentation produced for the reproducibility approach helped us to spot bugs during the implementation phase and we strongly believe that this type of actions should be supported more and more because, as reported by the analysis who tested the systems at CLEF eHealth “[...] still experienced varying degrees of difficulty to install and run the systems. [...] Analysts also report that additional information on system requirements, installation procedure and practical use would be useful for all the systems submitted, although documentation was overall more abundant and detailed compared to last year’s experiments. [...] The results of the experiments suggest that replication is achievable. However, it continues to be more of a challenge than one would hope.”

References

Raktima Chakrabarti, Hans-Georg Topf, and Michael Schroth. 2013. Retraction note to: Low dose lidocaine for refractory seizures in preterm neonates. *The Indian Journal of Pediatrics*, 80(6):529–529, Jun.

Kevin B Cohen, Jingbo Xia, Christophe Roeder, and Lawrence Hunter. 2016. Reproducibility in natural language processing: A case study of two r libraries for mining pubmed/medline. In *In LREC 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 6 – 12. European Language Resources Association (ELRA).

Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. 2017. A Reproducible Approach with R Markdown to Automatic Classification of Medical Certificates in French. In *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, Dublin, Ireland, September 11-14, 2017.*, CEUR Workshop Proceedings. 1866.

C. Drummond. 2009. Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML.*

Jacob Eisenstein. 2017. Unsupervised learning for lexicon-based classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3188–3194.

Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors. 2016a. *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR*

2016, Padua, Italy, March 20-23, 2016. Proceedings, volume 9626 of *Lecture Notes in Computer Science*. Springer.

Nicola Ferro, Norbert Fuhr, Kalervo Jarvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. 2016b. Increasing reproducibility in ir: Findings from the dagstuhl seminar on ”reproducibility of data-oriented experiments in e-science”. *SIGIR Forum*, 50(1):68–82. http://sigir.org/files/forum/2016J/p068.pdf.

Nicola Ferro. 2017. Reproducibility challenges in information retrieval evaluation. *J. Data and Information Quality*, 8(2):8:1–8:4, January.

Christopher Gandrud. 2015. *Reproducible Research with R and R Studio*. Chapman and Hall/CRC, second ed. edition.

Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon, editors. 2017. *CLEF 2017 eHealth Evaluation Lab Overview. CLEF 2017 - 8th Conference and Labs of the Evaluation Forum*, Lecture Notes in Computer Science. Springer.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, João R. M. Palotti, and Guido Zuccon. 2016. Overview of the CLEF ehealth evaluation lab 2016. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, pages 255–266.

Aurelie Neveol, Kevin Cohen, Cyril Grouin, and Aude Robert. 2016. Replicability of research in biomedical natural language processing: a pilot evaluation for a coding task. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 78–84, Auxtin, TX, November. Association for Computational Linguistics.

Aurélie Névéol, Robert N. Anderson, K. Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire Rey, Aude Robert, Claire Rondet, and Pierre Zweigenbaum. 2017. Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes*, CEUR Workshop Proceedings. CEUR-WS.org.

Roger D. Peng. 2011. Reproducible research in computational science. *Science*, 334(6060):1226–1227.

M. Schwab, N. Karrenbach, and J. Claerbout. 2000. Making scientific computations reproducible. *Computing in Science Engineering*, 2(6):61–67, Nov.

Pierre Zweigenbaum and Thomas Lavergne. 2017. Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*

Contrast-Ita Bank: A corpus for Italian Annotated with Discourse Contrast Relations

Anna Feltracco
Fondazione Bruno Kessler
University of Pavia, Italy
University of Bergamo, Italy
feltracco@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler
Trento, Italy
magnini@fbk.eu

Elisabetta Jezek
University of Pavia
Pavia, Italy
jezek@unipv.it

Abstract

English. We present Contrast-Ita Bank, a corpus annotated with discourse contrast relations in Italian. We annotate both explicit and implicit contrast relations, following the schema proposed in the Penn Discourse Treebank. We provide and discuss quantitative data about the new resource.

Italiano. *Presentiamo Contrast-Ita Bank, un corpus annotato con relazioni di contrasto in italiano. Abbiamo annotato sia relazioni esplicite che implicite, adottando lo schema proposto nel Penn Discourse Treebank. Portiamo e discutiamo dati quantitativi sulla nuova risorsa.*

1 Introduction

A relevant task in Natural Language Processing is the automatic identification of semantic relations between portions of text, such as textual entailment, text similarity, and temporal relation. In this contribution we focus on discourse contrast.

By *discourse relation* we mean a relation between two parts of a coherent sequence of sentences, propositions or speeches (i.e. discourse). We consider as discourse *contrast*: i) cases in which one of the two parts (henceforth *arguments*) is similar to the other in many aspects but different in one aspect for which they are compared, as in example (1), where both situations refer to a change in the price, but with different values; ii) cases in which one argument is denying an expectation that is triggered from the other argument, as in (2), where 'going to the beach' denies the expectation that, since it is raining, one would stay home. *Contrast* in text can be conveyed explicitly, by mean of a lexical element (connective), as by *while* in (1) and *although* in (2), or implicitly as in (3).

(1) The price of x increased of 5%, while the price of y decreased of 2.3.%

(2) Although it was raining, we went to the beach.

(3) Mary passed the exam. John failed it.

We present Contrast-Ita Bank [1], a corpus of Italian documents annotated with contrast, a very frequent relation in discourse. We aim to understand how frequent the contrast relation is in discourse, when it is expressed explicitly and implicitly, and which are the connectives that convey contrast. The final result of the annotation represents a first step toward a corpus of discourse relations for Italian, compatible with the Penn Discourse Treebank (PDTB) project (Prasad et al., 2007), the largest and the most used corpus annotated with discourse relations in the NLP field. A number of annotated corpora similar to the PDTB have been realised since its creation, for instance, the Prague Discourse TreeBank (Bejček et al., 2013)), the Chinese Discourse TreeBank (Zhou and Xue, 2015)), the Leeds Arabic Discourse TreeBank (Al-Saif and Markert, 2010)).[2] For Italian, a similar attempt was proposed by Tonelli et al. (2010), which uses the PDTB scheme for the annotation of the LUNA conversational spoken dialogue corpus. The authors annotated 60 real dialogues in the domain of software/hardware troubleshooting. Another project for Italian inspired by the PDTB is proposed by Pareti and Prodanof (2010) and it is focused on the relation of *attribution*, i.e "the relation of ownership between abstract objects and individuals or agents" (Prasad et al., 2007, p. 40).

Resources manually annotated with discourse relation have been used for instance for develop-

[1] https://hlt-nlp.fbk.eu/technologies/contrast-ita-bank

[2] Prasad et al. (2014) propose an overview of projects also mentioning resources for French, Turkish and Hindi.

ing methods and tools for the automatic identification and disambiguation of explicit marked or implicitly conveyed discourse relations[3], for the identification of the spans of text that are linked by relations (discourse segmentation), for the automatic creation of a summary of a written text (text summarization) (Marcu, 1998), and for machine translation (Meyer and Webber, 2013).

The paper is structured as follows: Section 2 introduces the contrast relation; Section 3 describes the annotation guidelines; Section 4 presents the content of the resource and Section 5 discusses the inter annotator agreement.

2 The Contrast Relation

Discourse contrast has been described in various theories and annotation schema. In the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), *contrast* is defined as the relation between two spans of texts such that the situations presented in the two spans are: "(i) comprehended as the same in many respects, (ii) comprehended as differing in a few respects, and (iii) compared with respect to one or more of these differences" (Mann and Thompson, 1988). In the framework of RST, Carlson and Marcu (2001) propose a discourse relations corpus; in their schema, *contrast* is part of a broader class of relations called *Contrast*, together with *concession*, described as "characterised by a violated expectation"(Carlson and Marcu, 2001).

In the Segment Discourse Representation Theory framework, Asher and Lascarides (1993; 2003) define *contrast* as a relation that involves constituents that are structurally similar but semantically dissimilar. According to them, this relation includes cases of *violation of expectation* in which what can be inferred from one of the constituents of a relation is denied in the second constituent (Asher and Lascarides, 2003, p. 167).

The Penn Discourse Treebank schema (Prasad et al., 2007) proposes different *senses* of the connectives that provide a semantic description of the discourse relation they convey. These senses are annotated as *sense tags*. The sense tag CONTRAST applies to cases in which the two arguments of a relation "share a predicate or a property and a difference is highlighted with respect to the values assigned to the shared property"; the sense tag CONCESSION is used for cases in which "the highlighted differences are related to expectations raised by one argument which are then denied by the other" (Prasad et al., 2007).[4]

We consider as *contrast* both what has been called *formal contrast* (Asher, 1993) and CONTRAST (Prasad et al., 2007) on the one hand (see Example (1) and (3)), and *violation of expectation* (Asher, 1993) or CONCESSION (Carlson and Marcu, 2001; Prasad et al., 2007) on the other hand (as in Example (2)).

3 Adopting the PDTB Schema

The Contrast-Ita Bank guidelines follow the PDTB 2.0 Annotation Manual (Prasad et al., 2007) and the recent proposal by Webber et al. (2016).

Following the PDTB 2.0, we annotate *explicit relations* (see Examples (1) and (2) above) by identifying the discourse connectives that trigger the relations and the respective arguments. We also annotate cases in which the relation is not marked by a connective and can be inferred between adjacent sentences. These cases include *implicit* relations, i.e. the relation is not lexically marked, as in Example (3), and *alternatively lexicalized* (*altlex*) relations, i.e. the relation is inferred by mean of another expression that is not a connective. By definition, these are cases where a discourse relation is inferred between adjacent sentences in absence of a connective, but where providing a suggestion of connective leads to redundancy in the expression of the relation (Prasad et al., 2007). For instance, in 'She prepared a cake. The reason: it was his birthday.'[5], a cause relation is conveyed through 'The reason:'; this relation is a case of *Altlex*, since 'The reason:' is not a connective, and providing a suggestion of connective (e.g. *because*) will lead to redundancy. Differently from the PDTB 2.0, we annotate *implicit* relations also among comma separated clauses and *altlex* among non adjacent sentences.

Specifically, our task involves: i) the annotation of the arguments of the relation (named *Arg1* and *Arg2*, being *Arg2*, the argument in the clause that is syntactically bound to the connective, and *Arg1*, the other one); ii) the annotation of the connectives that convey *contrast* in the case of *explicit* relations, of the first token of *Arg2* in the case of

[3] The task of identifying discourse relations in the form of a discourse connective taking two arguments is also called *shallow discourse parsing* and constituted a shared task of the CONLL conference in 2015 and 2016 (Xue et al., 2015).

[4] In the PDTB3.0 hierarchy (Webber et al., 2016), the two sense types belong to the class COMPARISON.

[5] See a similar example in (Prasad et al., 2007, p.7).

implicit relations, and of the expression that make us inferring the relation in the case of *altlex* relations; iii) the tagging of the sense of the relation. An example from the PDTB2.0 Manual (Prasad et al., 2007) is provided in (4), in which the connective appears underlined, *Arg1* is in italics, and *Arg2* is in bold.

(4) *Most bond prices fell on concerns about this week's new supply and disappointment that stock prices didn't stage a sharp decline*. **Junk bond prices moved higher**, <u>however</u>. (sense tag: *Contrast*)

Connectives. We followed the PDTB also for the definition of connectives that convey an *explicit* relation. They belong to three syntactic classes: (i) subordinating conjunctions (e.g. *when, because*); (ii) coordinating conjunctions (e.g. *and, or, but*); (iii) discourse adverbials, including both adverbs (e.g. *however, instead*), and prepositional phrases (e.g. *on the other hand, as a result*).

Arguments. According to the PDTB, relations are annotated when they are connecting "two *abstract objects* such as events, states, and propositions (Asher, 1993)" (Prasad et al., 2007), that are realised mostly as clauses, nominalisations, or anaphoric expressions. We follow the same guidelines, including conjoined VPs, as proposed by Webber et al. (2016).[6] We also adopt the Minimality Principle, according to which "only as many clauses and/or sentences should be included in an argument selection as are minimally required and sufficient for the interpretation of the relation"(Prasad et al., 2007). This means that there is no constrain on the length of an argument or that more than a sentence can be annotated (i.e. punctuation is generally not a limiting constrain).

Senses of relations. We consider a broad semantic definition of *contrast*, corresponding to the PDTB sense tags CONTRAST and CONCESSION. Specifically, we follow the PDTB 3.0 schema (Webber et al., 2016) in which CONCESSION has two subtypes, depending on which argument creates the expectation and which one denies it: if Arg2 creates an expectation that Arg1 denies, the proper tag is CONCESSION_Arg1.as.denier; conversely, when Arg1 creates an expectation that Arg2 denies, the tag that needs to be used is CONCESSION_Arg2.as.denier. In line with the PDTB2.0 we allow the annotation of more than one sense for a connective and, thus, the possibility of marking e.g. both CONTRAST and CONCESSION_Arg1.as.denier. Table 1 summarises the definition of the tags.

Relation and Definition in the PDTB
CONTRAST → the two Args share a predicate or a property and the difference between the two situations (in the Args) is highlighted with respect to the values assigned to the property.
CONCESSION → expectations raised by one argument which are then denied by the other. - *Arg1.as.denier* if Arg1 denies expectation - *Arg2.as.denier* if Arg2 denies expectation

Table 1: CONTRAST and CONCESSION in the PDTB 3.0 (Webber et al., 2016).

4 Contrast-Ita Bank

Contrast-Ita Bank is based on a corpus of 169 news stories selected from Ita-TimeBank (Caselli et al., 2011), for a total of 65,053 tokens (average length = about 385 tokens per document).[7] For the annotation we used the CAT tool (Bartalesi Lenzi et al., 2012). The annotation was carried by one expert annotator in about two weeks.

We annotated *explicit*, *implicit* and *altlex* relations of contrast for a total of 372 relations (average 2.16 per document). Table 2 reports the data of the annotation. *Explicit* relations are the most common and correspond to 91% of all the relations. We register a maximum number of 15 explicit relations in one document and an average of 2 relations per document. *Implicit* relations are less frequent and occur 15 times inter-sentencially and 9 times infra-sentencially, for a total of 24 annotations. This is different from the PDTB2.0, in which the ratio between *explicit* and *implicit* for what concerns CONTRAST and COMPARISON, and their subtypes, is about 0.45, while in Contrast-Ita Bank is ten time less. This might be due to the fact that in Contrast-Ita Bank annotators were asked to mark contrast, and it is possible that they simply fail to capture implicit relations, while in the PDTB2.0 annotators were asked to mark also cases where no relation can be inferred between adjacent sentences, thus analysing in detail if a relation appears between every pair of sentences. *Altlex* relations are rarer: in Contrast-Ita

[6] This change includes avoiding the annotation the span of text that can be referred to both arguments in case of inter-sentencial VP conjoined arguments (e.g. in 'Mary *likes fruits* <u>but</u> **hates peaches**, 'Mary has not been annotated).

[7] The same corpus is annotated with factuality information in Fact-Ita Bank (Minard et al., 2014) and partially annotated with negation in Fact-Ita Bank-Negation (Altuna et al., 2017).

	Explicit	Implicit	AltLex	Total
CONTRAST	87	12	3	102
CONC.Arg1-denier	21	0	1	22
CONC.Arg2-denier	201	8	3	212
Double relations	32	4	0	36
Total	341	24	7	372
Density	0.0052	0.0003	0.0001	0.0056

Table 2: Contrast relations in Contrast-Ita Bank.

Bank there are 7 cases.[8] In these cases relations are *alternatively lexicalized* by: ‘anche al netto di’, ‘Certo’, ‘Il punto è che’, ‘Non’, ‘Peccato che’ ‘quella sì’, ‘Macchè’; none of these expressions is a connective.

Table 2 also shows that the per token density of *contrast* in the corpus is 0.0056, similar to the PDTB (i.e. 0.0072).[9]

The most frequent sense tag is CONCESSION. Arg2-as-denier (i.e. when Arg2 denies an expectation that rises from Arg1), which covers about 56% of the cases. CONTRAST covers almost a quarter of the cases and the two relations have been annotated together 32 times (out of the total 36 cases of double annotation). CONCESSION.Arg1-as-denier is far less frequent both as single type as with other relations, and has been annotated less than 10% of the cases. This subtype is associated to a limited set of connectives: despite the list of connectives in Contrast-Ita Bank consists of 19 connectives (see Table 3), 7 of them (e.g. *nonostante*) signal CONCESSION.Arg1-as-denier all the times.

Not surprisingly, *ma* accounts for almost half of the cases (the equivalent *but* is also the most used for these senses in the PDTB 2.0), and *invece*, *mentre*, *però* for about a 10%. Table 3 shows that, as it happens for content words, the most frequent connectives are the most polysemous ones.

5 Inter Annotator Agreement

We computed the agreement (IAA) between two annotators on 18 documents (10.6% of the whole corpus), which followed the same written guidelines. Data are reported in Table 4.

connective	**#**	**%**	**% for CONTRAST**	**% for CONC.Arg1-denier**	**% for CONC.Arg2-denier**	**% for Double Relation**
ma	164	48.09	4.3		87.2	8.5
invece	41	12.02	78		9.75	12.25
mentre	36	10.56	88.9		2.8	8.3
però	35	10.26	2.9		85.7	11.4
nonostante	11	3.23		100		
anche se	10	2.93			90	10
e	8	2.35	75			25
se	8	2.35	75			25
eppure	7	2.05		100		
comunque	4	1.17		100		
pur	4	1.17		100		
tuttavia	4	1.17			100	
a dispetto di	2	0.59		100		
seppure	2	0.59		100		
al contrario	1	0.29	100			
al contrario di	1	0.29	100			
da una parte.. dall’altra	1	0.29	100			
in verità	1	0.29				100
in realtà	1	0.29		100		

Table 3: Contrast connectives in Contrast-Ita Bank along with: total number, percentage over the total cases, percentage of cases per sense tags.

First we measured the agreement on recognising *explicit*, *implicit* or *altlex* contrast relations (*relation identification*), considering the text span marked by the annotators to signal a relation (e.g. agreement if both marked *ma* or if one marked *se* and the other *anche se* to signal the presence of a contrast relation). We calculated the final score adopting the Dice’s coefficient (Rijsbergen, 1979).[10] The result is that annotators agree in 37 cases (Dice 0.68). We consider this result reasonable given the difficulty of the task which has not to be underestimated. To identify contrast relation in a document means to distinguish cases in which a lexical element is playing the role of connective of contrast or it is not, and also to identify *implicit relations* that by definition are not marked in the text. In order to understand the motivations of these discrepancies, we have adopted a reconciliation strategy among annotators in which they were asked to motivate their choices with the possibility of revising them. After the reconciliation dis-

[8] This is also the rarest type in the PDTB 2.0, among the three considered here.

[9] It is possible that contrast is more frequent in corpora of other domains, such as in documents reporting debates in which people contrast their opinions. However, with the idea of maximising the compatibility with the PDTB, we annotated contrast on a corpus of news.

[10] The Dice’s coefficient measures how similar two sets are by dividing the number of shared elements of the two sets by the total number of elements they are composed by. This produces a value from 1, if both sets share all elements, to 0, if they have no element in common.

cussion 16 cases were reconciliated and the Dice value increased to 0.84.

In other cases disagreement remained. These mainly include cases in which both annotators recognized a discourse relation but one interpreted the relation to be of contrast, while the other did not. In many cases, these relations are conveyed by the coordinating conjunction ‘e’. We report an example in which one annotator recognized a contrast; while the other considered the arguments as non-contrasting parts of a description.

(5) [..] *sono portatori sani di Talassemia Mayor* e **il loro bambino, Luca, cinque anni, è talassemico**.[11] [doc:5402]
CONTRAST vs NON-MARKED

Agreement on *connectives identification* is calculated considering if both annotators agree on recognising the same explicit relation and the same exact span of text to be a connective (thus excluding cases of *altlex* and *implicit*). In these terms, cases of agreement for *connectives identification* are a subset of cases of agreement already captured by the *relation identification*. The resulting agreement is 0.68 (Dice’s coefficient).

For the 37 cases of agreement on *relation identification*, we calculated the IAA on the span of *arguments* in two ways. In the *exact match* mode, we have agreement if the two annotators consider the exact span of text as Arg1 or Arg2 for the same relation; in the *relaxed match* mode, we consider agreement if the text span identified by the annotators matches at least for its 50%. Agreement in the *exact match* for Arg1 is 0.51 and for Arg2 is 0.70; in the *relaxed match* mode is 0.89 for Arg1 and 0.91 for Arg2. We expected the *exact match* agreement difficult to reach. In fact, as described in Section 3, we adopt the Minimality Principle for the annotation of the arguments. The selection of the arguments span thus relies significantly on the interpretation of the annotators and cases in which there is no exact match can be frequent.

Agreement in identifying CONTRAST and CONCESSION (*sense type*) is calculated counting 1 point if annotators agree to assign (or not) the same tag(s), 0.5 if one chooses a tag and the other both, 0 for total disagreement. IAA is obtained summing the points for each annotation and dividing by the total of 37 relations that both annotators identified. Agreement for *sense type* is 0.73, showing that recognising the type of contrast can be a controversial decision among annotators. However, we believe that this result is fair, considering that the annotation regards non mutually exclusive types of the same class.

# of relations by annotators: A= 57; B= 51; A ∩ B= 37	
IAA on:	
relation identification	0.68
relation identification - post reconciliation	0.84
connectives identification - explicit	0.68
arguments span - exact match (Arg1; Arg2)	0.51; 0.70
arguments span - relaxed match (Arg1; Arg2)	0.89; 0.91
sense type: CONTRAST - CONCESSION	0.73
sense subtype: Arg1.as.denier - Arg2.as.denier	0.9

Table 4: InterAnnotator Agreement.

Finally, when there is agreement on CONCESSION, we applied the same formula to calculate IAA between *CONCESSION subtypes: Arg1.as.denier - Arg2.as.denier*: agreement is 0.9. Specifically, annotators agree in 10 cases to mark CONCESSION but in one case they disagree over the direction of the relation.[12]

Overall, the IAA highlights that the main difficulties of annotating *contrast* concern: the *relation identification*, especially for *implicit* and *altlex* relations; the extent of the *arguments*: the two annotators frequently do not mark exactly the same tokens but it is very likely that their annotations match at least for their 50%; *sense type*: one annotator tends to annotate also the CONCESSION_Arg2.as.denier when marking CONTRAST, while the other annotator does not.

6 Conclusion and Further Work

We presented Contrast-Ita Bank, a corpus annotated with discourse contrast relations in Italian. Following the PDTB annotation schema, we annotated *explicit*, *implicit* and *altelex* relations of contrast. We also present the list of connectives that convey contrast in the corpus. The new resource can be integrated with LICO, the Lexicon of Italian Connectives (Feltracco et al., 2016), validating the list of connectives and adding examples from corpus to the connectives. Contrast-Ita Bank

[11] Eng.:[..] *they are carrier of Talassemia Mayor* and **their son, Luca, five years old, is thalassaemic**.

[12] For the argument identification in the PDTB 2.0, Prasad et al. (2008) report an agreement of 90.2% for explicit relation and 85.1% for implicit (we do not calculate the value considering this granularity); when relaxing the match to partial overlap, the two values increase to 94.5% and to 85.1%. Additionally, authors report an agreement of 94% for sense class, of 84% for sense type, and of 80% for the subtype level.

is distributed under a CC-BY-NC 4.0 licence.

References

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10).*

Begoña Altuna, Manuela Speranza, and Anne-Lyse Minard. 2017. The Scope and Focus of Negation: A Complete Annotation Framework for Italian. *Semantics Beyond Events and Roles (SemBEaR) 2017*, page 34.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. Cat: the celct annotation tool. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12)*, pages 333–338.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague Dependency Treebank 3.0. `http://ufal.mff.cuni.cz/pdt3.0/`.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151.

Anna Feltracco, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. 2016. Lico: A lexicon of italian connectives. *Proceedings of the Second Italian Conference on Computational Linguistic (CLiC-it 2016)*, page 141.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 1998. *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. thesis, University of Toronto.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the 1st DiscoMT Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, number EPFL-CONF-192528.

Anne-Lyse Minard, Alessandro Marchetti, and Manuela Speranza. 2014. Event factuality in italian: Annotation of news stories from the ita-timebank. In *Proceedings of the First Italian Conference on Computational Linguistic (CLiC-it 2014).*

Silvia Pareti and Irina Prodanof. 2010. Annotating attribution relations: Towards an italian discourse treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The Penn Discourse Treebank 2.0 Annotation Manual.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse Tree-Bank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*.

Cornelis van Rijsbergen. 1979. *Information retrieval*. Butterworth, London.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind K Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A Discourse-Annotated Corpus of Conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *CoNLL Shared Task*, pages 1–16.

Yuping Zhou and Nianwen Xue. 2015. The chinese discourse treebank: A chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

Ironic Gestures and Tones in Twitter

Simona Frenda
Computer Science Department - University of Turin, Italy
GruppoMeta - Pisa, Italy
simona.frenda@gmail.com

Abstract

English. Automatic irony detection is a young field of research related to Sentiment Analysis. When dealing with social media data, the shortness of text and the extraction of the statement from his context usually makes it hard to understand irony even for humans but especially for machines. In this paper we propose an analysis of the role that textual information plays in the perception and construction of irony in short texts like tweets. We will focus on the impact of conventional expedients of digital writing, which seem to represent a substitution of typical gestures and tones of oral communication, in figurative interpretation of messages in Italian language. Elaborated computational model has been exploited in the development of an irony detection system, which has been evaluated in the Sentipolc's shared task at EVALITA 2016.

Italiano. *Il riconoscimento automatico dell'ironia è un ambito di ricerca giovane, rilevante per la Sentiment Analisys. Quando si tratta di* social media data*, la brevità del testo e la sua estrazione dal contesto rendono difficile la comprensione dell'ironia anche per l'uomo e in particolare per le macchine. In questo lavoro, si propone un'analisi sul ruolo che l'informazione testuale gioca nella percezione e realizzazione dell'ironia nei* tweet*. Ci si focalizzerà sull'impatto di elementi convenzionali della scrittura digitale, che sembrano rappresentare una sostituzione dei tipici gesti e toni della comunicazione orale, nell'interpretazione figurativa dei messaggi in italiano. Il modello computazionale elaborato è stato usato in un sistema di* irony detection*, valutato a Sentipolc, Evalita 2016.*

1 Introduction

The growing scientific interest on natural language understanding has been supported in the last decade by a great amount of user-generated texts available on the Web. People usually use social media platforms, such as Facebook and Twitter, to express their opinions on different topics, which can be exploited, for example, by companies for marketing researches. This is one of the motivations which prompted actual research in this direction on automatic analysis of short-texts. Social micro-texts are great examples of rhetorical production due to their shortness, which supports the creativity of linguistic expressions (Ghosh et al., 2015). In fact 140 characters of tweets encourage users to use some creative devices in order to communicate briefly their opinions or their feelings about events, products, services or other individuals. Among creative devices, irony and sarcasm hinder correct sentiment analysis of texts and, therefore, correct opinion mining. Indeed, irony is a figurative language device used to convey the opposite of literal meaning: *contrarium quod dicitur intelligendum est* (Quintiliano, Institutio Oratoria, 9, 22-44). In order to express an ironic utterance in short text, users prefer to use conventional expedients in digital writing or particular linguistic constructs which seem to represent a substitution of typical gestures and tones of oral communication. These reveal themselves as good clues for Irony Detection as demonstrated by results obtained with our system participating in SENTIPOLC's at EVALITA 2016 (Frenda, 2016), where we ranked third on twelve participants. In this paper we present linguistic analysis on ironic tweets extracted from corpora used in SENTIPOLC and computational model elaborated in Master's thesis upon which our rule-based system is based.

2 Related Work

Automatically understanding texts that are susceptible to different interpretations from their literal meaning is a hard task that presents challenging aspects even for humans. Nevertheless, automatic irony detection is becoming one of the biggest challenges of Natural Language Processing (NLP), especially to correctly determine the polarity of texts. Indeed, in the last years several studies arose with the aim of detecting irony and sarcasm by extricating their multiple aspects and exploiting various computational models in different languages: as regards English the research by Utsumi (1996) was one of the first approaches; Veale and Hao (2009) focused on figurative comparisons ("as X as Y"); Reyes et al. (2013) exploited features ranging from textual to stylistic dimensions, and Barbieri and Saggion (2014) considered lexical and semantic features of the words in tweets. Relative to French, Karoui et al. (2015) focused on the presence of negation markers and the implicit and explicit opposition in ironic tweets. Finally, multilingual perspective is proposed by Karoui et al. (2017), which examine the impact of pragmatic phenomena in the interpretation of irony in English, French and Italian tweets. The main work inspiring our researches here is Carvalho et al. (2009) which distinguished eight oral and gestural "clues" for irony detection in Portuguese online newspaper comments. Their attention focused in particular on positive comments: positive sentences are more subjected to irony and it is more difficult to recognize their true polarity. Many of these clues have been used in our analysis on ironic Italian tweets to observe how these textual features are distributed in negative and positive sentences to bring out possible incongruities between literal and real meaning.

3 Methodology

The irony detection task is a very recent challenge in NLP community and in 2014 and 2016 EVALITA, an evaluation campaign of NLP and speech tools for Italian, proposed a battery of tasks related to Sentiment Analysis in tweets, including *Irony detection*. The task of automatic irony detection is treated as a problem of classification of texts in ironic and non ironic ones, and the main approaches used by previous works are based on the development of supervised machine-learning or rule-based systems.

We developed a rule-based system, implemented in Perl, which, analysing a corpus of Italian tweets, identifies possible ironic clues and distinguishes ironic and non ironic texts. This system is based on computational model that is the result of linguistic research carried out during Master's thesis redaction. The scope of this analysis is to understand the impact of conventional elements of web writing and syntactic constructions on automatic process of recognition of ironic short-texts.

We tested our computational model with good results participating in SENTIPOLC's task at EVALITA in 2016.

3.1 Corpora of tweets

Tweet corpora used in our works have been provided by organizers of SENTIPOLC task in EVALITA 2014 and 2016: SENTIPOLC 2014 corpus includes 4513 tweets in the training set and 1935 in the test set, and SENTIPOLC 2016 includes 7410 in the training set and 2000 in the test set. The former has been used for linguistic analysis in Master's thesis and the latter to participate at evaluation campaign. These corpora have been annotated manually and according to a multi-layered annotation scheme where tweets are labelled according to different dimensions: subjectivity, overall and literal polarity (positive/neutral/negative/mixed), irony. These corpora contain a collection of both political and generic tweets, and also a collection of socio-political tweets (concerning topic *la buona scuola*).

3.2 Resources and Data Processing

Considering various textual elements of digital writing which make up tweets, that are essential to linguistic analysis of this kind of text, we developed a lexicon of interjections [1] annotated according polarity, a list of emoticons extracted from Wikipedia and annotated as EMOPOS (*=)* , *:D*), EMONEG (as *:(* , *:'(*) and EMOIRO (*^L ^* , *:P*), and a list of ironic hashtags extracted from ironic tweets in corpora analysed[2].

In order to clean up the texts and avoid hampering syntactic analysis and ironic clues retrieval we replaced emoticons with appropriated labels

[1] Extracted from Morph-it! (Zanchetta and Baroni, 2005) and Treccani (`http://www.treccani.it`).

[2] For more details about resources see (Frenda, 2016).

and removed characters of url from text. Cleaned texts have been processed by TreeTagger (Schmid, 1994) for obtaining POS-tagged and lemmatized corpora, using Italian tagset by Baroni.

4 Irony Detection Model

People in social network use a new kind of language between speech and writing: oral elements are included in writing by means of graphic characters, punctuation and so on. Users express their emotions and opinions with informal language especially in the social network, using interjections or expressing tones with exclamatory expressions. Considering the shortness of text users tend to use conventional marks, like hashtags, to provide additional information (context, emotion, and so on).

In our work we exploit these textual patterns, many of whom are extracted from Carvalho et al. (2009) and adapted to Italian language. Indeed, their results demonstrated that more productive patterns in ironic texts are the ones related to orality and gestures. We considered also regional expressions and other forms of exclamation specifically of Italian language. In Italian texts, like in Portuguese, these linguistic elements, which seem to reproduce oral communication, are the most productive as demonstrated in Figure 1 and 2. In these figures we can observe the impact of our computational model in corpora analysed.

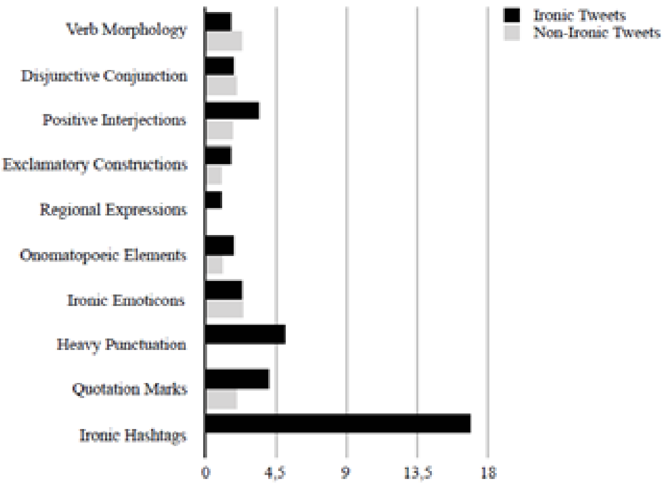


Figure 1: Ironic clues in SENTIPOLC 2014 corpus (in percentage)

Although in ironic tweets most of the frequencies of these patterns are promising for irony recognition task, these corpora contain an imbalanced data distribution (564 ironic tweets on 4513 in SENTIPOLC 2014 and 865 ironic tweets on 7410 in SENTIPOLC 2016) that hinder the possible generalization of model.

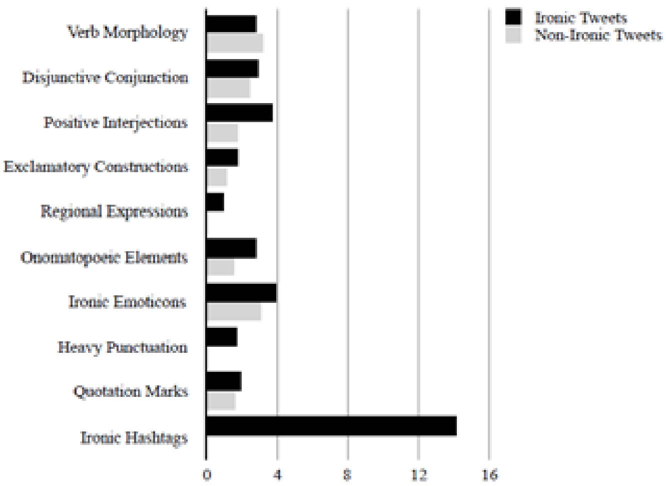


Figure 2: Ironic clues in SENTIPOLC 2016 corpus (in percentage)

Below, we summarily describe linguistic features considered in our model and their frequencies in positive and negative sentences (Figure 3 and 4), observing specifically in texts how user express ironic utterance:

- Verb morphology: the use of pronoun *tu* and, in a pro-drop language like Italian, morphological inflection of the verb *essere* for second singular person allows to express a certain proximity also artificial or false if interlocutor is a well-known person.

- Disjunctive conjunctions (*o*, *oppure*) sometimes introduce strange combinations that surprise the readers and encourages an ironic interpretation.

- Positive interjections and exclamatory expressions, like expressions with an emphasised use of pronoun or adjective *che* (like *Che sorpresa!*, *Che bella giornata!*), represent a simple way for users to communicate emotions, feelings, mental states or reactions to specific situations, reversing also the literal meaning of statement.

- Regional expressions, like exclamatory expressions and interjections, are a way for users to express immediately and informally their moods or opinions, especially in ironic perspective. In corpora analysed, it is prevalent the use of expressions of dialect from central Italy, such as: *annamo bene*, *ce vuole* or *ce sta*.

- Onomatopoeic expressions for laughter are used by users like markers to suggest an ironic interpretation of text.
- Ironic emoticons: emoticons allow to express briefly the user's moods (happiness, sad, laughter, ect) or to communicate to the reader ironic or humorous intention, for instance, with wink (*;)*).
- Heavy punctuation is used to set a tone in writing, in particular in short texts, where the verbal components are essential to express concisely the feelings.
- Quotation marks, also imitated in gestures of speaking, are used to quote what has been said by others or to emphasize the content suggesting a possible additional interpretation of text.
- Ironic hashtags: the hashtag complies with necessity of simplification and containment (Chiusaroli, 2014) and plays a special role since it is employed by Twitter's users as digital extralinguistic equivalent of non-verbal expressions (Liebrecht et al., 2013), sometimes affecting also the sentiment of tweets (Maynard and Greenwood, 2014).

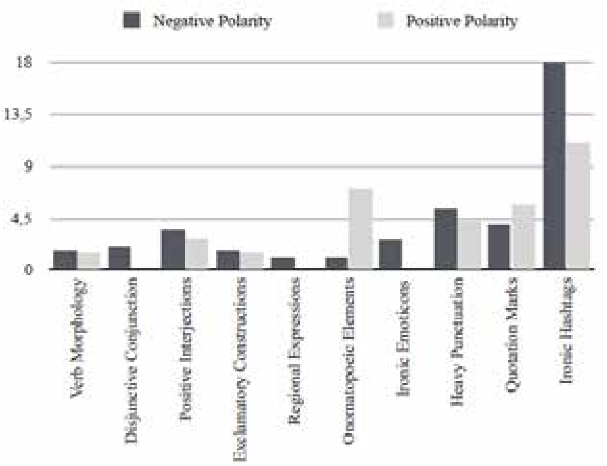


Figure 3: Distribution in positive and negative sentences in SENTIPOLC 2014 corpus (in percentage)

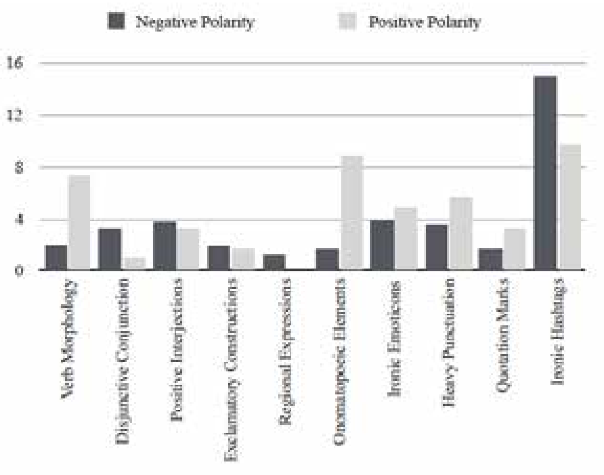


Figure 4: Distribution in positive and negative sentences in SENTIPOLC 2016 corpus (in percentage)

5 Discussion and Conclusions

Although limited amount of Italian ironic examples, this analysis and the results of developed computational system (Frenda, 2016) show that people tend to use textual and conventional expedients of oral communication to express irony in informal context as social networks. We can observe this in Figure 1 and 2, where some linguistic constructions expressing tone and accent of user-speaker, like regional expressions and heavy punctuation, are used mainly in ironic tweets. With respect to ironic hashtags we can observe that same hashtags are mentioned in different ironic tweets in both corpora, revealing their important role of established conventional elements in communication in social networks. Finally, in Figure 3 and 4 we can observe that there are cases of incongruity between literal and real meaning, for example there are sentences with negative polarity that contain positive interjections or exclamatory constructions used, indeed, in ironic manner. It is interesting to underline that most of ironic tweets are negative in both corpora: 493 negative ironic tweets on 564 ironic tweets in SENTIPOLC 2014 corpus and 742 on 865 in SENTIPOLC 2016 corpus.

In this scenario where automatic irony detection is still challenging for Italian, pragmatic analysis of ironic texts allows to take a closer look at how people use the language and his expedients to express irony.

References

Francesco Barbieri and Horacio Saggion. 2014. Modelling Irony in Twitter, Features Analysis and Evaluation. *Language Resources and Evaluation conference, LREC*. Reykjavik, Iceland.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTIment POLarity Classification Task. *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*. Academia University Press.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, Marco Mazzoleni. 2004. Introducing the la Repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004*. Lisbon: ELDA. 1771-1774.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Baltimore, Maryland, USA. 42–49.

Paula Carvalho, Luís Sarmento, Mário J. Silva and Eugénio De Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.*. ACM. 53–56.

Francesca Chiusaroli. 2014. Sintassi e semantica dell'hashtag: studio preliminare di una forma di Scritture Brevi. *Proceedings of the Fourth International Workshop EVALITA 2014*. Pisa University Press.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. CoNLL '10. Stroudsburg, PA, USA. Association for Computational Linguistics. 107–116.

Pierluigi Di Gennaro, Arianna Rossi and Fabio Tamburini. 2014. The FICLIT+CS@UniBO System at the EVALITA 2014 Sentiment Polarity Classification Task. *Proceedings of the Fourth International Workshop EVALITA 2014*. Pisa University Press.

Simona Frenda. 2016. Computational rule–based model for Irony Detection in Italian Tweets. *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*. Academia University Press.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, John Barnden. 2015. SemEval-2015 Task 11:Sentiment Analysis of Figurative Language in Twitter *in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. pages 470–478, Denver, Colorado, June 4-5, 2015.

Roberto González-Ibáñez, Smaranda Muresan and Nina Wacholder. 2011. Identifying Sarcasm in Twitter: A Closer Look. *Proceedings of the 49th Annual Meeting of the Association for Computa- tional Linguistics: shortpapers*. Portland, Oregon, June 19-24. 581–586.

Jihen Karoui, Farah Benamara Zitoune, Veronique Moriceau, Nathalie Aussenac-Gilles and Lamia Hadrich Belguith. 2015. Detection automatique de l'ironie dans les tweet en francais. *22eme Traitement Automatique des Langues Naturelles*. Caen, 2015.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco and Nathalie Aussenac-Gilles. 2017. Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain, April 2017.

Roger J. Kreuz and Gina M. Caucci. 2007. Lexical Influences on the Perception of Sarcasm. *Proceedings of the Workshop on Computational Approaches to Figurative Language*. Rochester, NY.1–4.

Christine Liebrecht, Florian Kunneman, and Antal Van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics*. pages 29–37, Atlanta, Georgia, June.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14) (26-31). European Language Resources Association (ELRA)*. Reykjavik, Iceland, 4238–4243.

Antonio Reyes, Paolo Rosso and Tony Veale. 2013. A multidimensional approach for detecting irony in- Twitter. *Language Resources and Evaluation*. 47:239–268.

Graeme Ritchie. 2009. Can computers create humor? *AI Magazine*. Volume 30, No. 3. 71-81.

Helmut Schmid. 1994. Probabilistic Part-of-Speech-Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

Oliviero Stock and Carlo Strapparava. 2006. Laughing with HAHAcronym, a computational humor system. *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. Boston, Massachusetts.

Mirko Tavosanis. 2010. *L'italiano del web.*. Carocci. Roma.

Akira Utsumi. 1996. A unified theory of irony and its computational formalization. *Proceedings of the 16th conference on computational linguistics* Association for Computational Linguistics. Morristown, NJ. 962–967

Tony Veale and Yanfen Hao. 2009. Support structures for linguistic creativity: A computational analysis of creative irony in similes. *Proceedings of CogSci 2009, the 31st annual meeting of the cognitive science society*. 1376–1381.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! A free corpus-based morphological resource for the Italian language. *Proceedings of Corpus Linguistics 2005*. University of Birmingham, Birmingham, UK.

Emerging Sentiment Language Model for Emotion Detection

Anastasia Giachanou[1], Francisco Rangel[2,3], Fabio Crestani[1], and Paolo Rosso[2]

[1]Faculty of Informatics, Università della Svizzera italiana (USI), Lugano, Switzerland
[2]PRHLT Research Center, Universitat Politècnica de València, Spain
[3]Autoritas Consulting, S.A., Spain
{anastasia.giachanou,fabio.crestani}@usi.ch
francisco.rangel@autoritas.es, prosso@dsic.upv.es

Abstract

English. In this paper we present an approach for joy, anger and neutral emotions detection based on an emerging sentiment language model. We propose an approach that can detect specific emotions from positive, neutral and negative sentiments and which favors the tweets that occur at recent sentiment spikes. Our results suggest that our approach can effectively detect joy, neutral and anger emotions and that it performs better compared to the baselines.

Italiano. *In questo articolo presentiamo un approccio per rilevare gioia, rabbia e emozione neutra basato su un modello di sentiment analysis emergente. Proponiamo un approccio in grado di rilevare emozioni specifiche da sentimenti positivi, neutri e negativi e che favorisca i tweets che si verificano nei picchi recenti di sentimento. I risultati suggeriscono che il nostro approccio può rilevare efficacemente le emozioni di gioia, rabbia e neutra che ottiene migliori risultati delle baseline.*

1 Introduction

Recent years have seen the emergence of social media that enable people to share their thoughts and opinions in an easy and fast way. Opinions posted on social media are very useful to understand what people think and how they feel about a specific entity (e.g. a product, a person, a company etc.). For example, companies can mine users' opinions on a product that has just been released to understand if the users are satisfied or not and act accordingly. Therefore, the automatic detection of emotions and sentiments from text has attracted a lot of research interest (Tang et al., 2015; Mohammad, 2015).

Although sentiment and emotion analysis share some similarities, they are two different problems (Munezero et al., 2014). *Sentiment analysis* focuses on understanding the sentiment polarity (positive, neutral, negative) of a text (Pang and Lee, 2008) whereas *emotion analysis* refers to its affectual attitude such as anger, joy, fear etc. (Mohammad, 2015). Most of the previous work have tried to predict sentiments from data annotated with sentiments and emotions from data annotated with emotions. However, preliminary experiments showed that some emotions are related to sentiments. Specifically, negative sentiment is related to anger, positive sentiment to joy and text with no emotion (neutral emotion) to text with no sentiment (neutral sentiment).

In addition, public sentiment towards a specific entity changes over time and in some cases sentiment spikes may occur. Sentiment spikes occur when a large amount of documents of a specific sentiment is posted (Giachanou et al., 2016). The documents that occur at sentiment spikes usually refer to a topic or event that attracted a lot of attention and therefore they can be very helpful for sentiment and emotion analysis. To this end, in this study we propose to incorporate information from the documents that have occurred at sentiment spikes to improve the performance of the emotion analysis task.

In this paper, we focus on Twitter and we propose the *emerging Sentiment Language Model (emerging-SLM)* approach which favors tweets that occur at recent sentiment spikes with the aim to predict joy, anger and neutral emotions from positive, negative and neutral sentiments respectively. We test our approach on a collection of tweets that spans over nine days and we show that the emerging-SLM performs better compared to both state-of-art Sentiment Language Model (SLM) and to a random Sentiment Language Model (random-SLM).

2 Related Work

Sentiment and emotion analysis have both attracted much research attention (Mohammad, 2015; Giachanou and Crestani, 2016). The main difference between the two problems is that sentiment refers to the polarity (e.g. positive, neutral, negative) whereas emotion refers to the affectual attitude that is anger, joy, fear etc. (Mohammad, 2015).

Sentiment analysis has attracted a tremendous research attention over the last years. The proposed approaches can be roughly classified as learning and lexicon based. The lexicon based approaches are typically unsupervised and use lists of words (e.g. *good*,*bad*) whose presence implies a specific sentiment polarity (Turney, 2002; Taboada et al., 2011). The learning based approaches rely on a number of features, usually extracted from text, to build a classifier which is then used to annotate unlabeled text as positive, negative or neutral (Pang et al., 2002). More recently, researchers have proposed deep learning approaches to learn sentiment specific word embeddings (Tang et al., 2014) or semantic representations of user and products (Tang et al., 2015) to address sentiment analysis. A thorough review on opinion retrieval and sentiment analysis can be found in Pang and Lee (2008) whereas Giachanou and Crestani (2016) focused on Twitter sentiment analysis.

With regards to emotion analysis, Mohammad (2012) considered hashtags that refer to an emotion (e.g., #anger, #surprise) to create a collection for emotion analysis and showed that these hashtag annotations matched with the annotations of trained judges. Roberts et al. (2012) extended the list of the six Ekman's basic emotions (joy, anger, fear, sadness, surprise, disgust) (Ekman, 1992) with an additional emotion (love) and created a series of binary SVM emotion classifiers. Also, other researchers have used sentiments or emotions to address other tasks such as irony detection (Farías et al., 2016) or author profiling (Rangel and Rosso, 2016).

In general, language models have been used for text classification problems (Bai et al., 2004). With regards to sentiment analysis, Liu et al. (2012) used manually annotated data to train a language model and then applied smoothing using noisy emoticon data. There are also few works that have considered sentiment dynamics. Bollen et al. (2011) used a psychometric instrument to extract and analyze different moods (tension, depression, anger, vigor, fatigue, confusion) detected in tweets and found that the mood level is correlated to cultural, political and other world global events while An et al. (2014) combined sentiment analysis, data mining and time series methods to track sentiment regarding climate change from Twitter feeds. However, our work is different because we use temporal information to favor documents that were posted recently and attracted a lot of attention with the aim to improve the performance of detecting specific emotions.

3 Methodology

Language Models (LMs) that are widely used in Information Retrieval (IR) and Natural Language Processing (NLP) fields assign probabilities to sequences of words (Ponte and Croft, 1998). The most typical scenario in IR consists in generating a Language Model (LM) for each document and then estimating the likelihood that the query was generated by each document. The documents then can be ranked based on the likelihoods. For a classification problem, we first aggregate all the documents of each specific class and then we estimate the likelihood that a new document is generated from each of the estimated language models. The new document can be annotated with the class for which it has the maximum likelihood.

More formally, let Θ^+, $\Theta^{\bullet}$, Θ^- be the LMs for the positive, neutral and negative classes respectively. Given a test tweet d we can detect its emotion class (joy, anger, neutral) c' as:

$$p(d|c') = \prod_{i=1}^{|d|} p(t_i|\Theta^c)$$

where $|d|$ is the number of words in tweet d and $p(t_i|\Theta^c)$ is a multinomial distribution estimated from the LM of class c (positive, negative, neutral).

To estimate the distributions we use the Maximum Likelihood Estimate (MLE) which computes the probabilities as follows:

$$p(t|\Theta^c) = \frac{n(t,c)}{\sum_{i=1}^{|V_c|} n(t_i,c)}$$

where $n(t,c)$ is the number of times that the term t appears in the collection of documents of class c and $|V_c|$ is the size of the vocabulary of class c.

The *emerging-SLM* combines two different LMs to estimate the probabilities of the terms. The first LM is based on all the tweets of the collection excluding those that occur at recent sentiment spikes whereas the second is based on tweets that occurred at those recent sentiment spikes. Formally, the distributions of the terms using the *emerging-SLM* are estimated as:

$$p(t|\Theta^c) = \lambda * p_{global}(t|\Theta^c) + (1-\lambda) * p_{burst}(t|\Theta^c)$$

where Θ^c is the LM for the class c, $p_{burst}(t|\Theta^c)$ is the probability of that term t appear in the recent sentiment spikes of the class c, $p_{global}(t|\Theta^c)$ is the probability of that term t appear in the class c and λ is the parameter that determines the importance of each LM for the final estimation. Here we should note that $p_{global}(t|\Theta^c)$ is calculated after we excluded the tweets that occurred at sentiment spikes.

One common issue with the LMs is that they assign zero probabilities to terms that do not appear in the training data. To overcome this problem, we apply Jelinek-Mercer smoothing that assigns nonzero probabilities to unseen terms (Zhai and Lafferty, 2004). Jelinek-Mercer smoothing refers to a linear interpolation of the MLE and the collection language model $p(t|\Theta_c)$ and can be defined as:

$$p(t|\Theta) = \mu * p(t|\Theta) + (1 - \mu) * p(t|\Theta_c)$$

where the collection language model is estimated using the maximum likelihood estimate of the whole collection.

To detect the sentiment spikes, we measure the evolution of each sentiment as $r_{t,s} = N_{t,s}/N_t$ where $N_{t,s}$ is the number of documents that express the sentiment s posted at time t and N_t is the total number of documents posted at time t. Figure 1 shows an example of negative spikes that occurred while tracking the sentiment towards *Michelle Obama*.

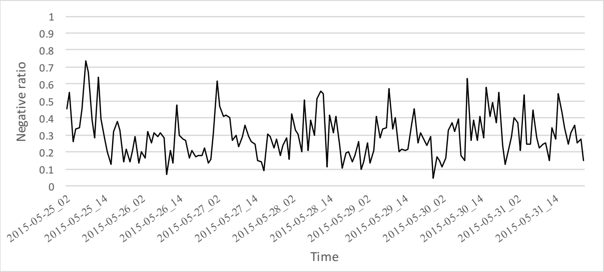


Figure 1: Negative spikes that occurred while tracking the sentiment towards *Michelle Obama*.

4 Experimental Setup

In this section we describe the experimental details of our study that include the description of the dataset, the baselines we used and the experimental settings.

4.1 Dataset

Our collection contains 25,588 tweets about *Michelle Obama* and spans from June 25, 2015 to July 2, 2015. To annotate the collection, we used the Crowdflower platform[1]. Tweets were annotated with regards to sentiment and emotion. For sentiments, annotators could choose among {*positive, no sentiment, negative*} whereas for emotions they could choose among {*anger, fear, sadness, disgust, surprise, happiness, no emotion*}. Each tweet was annotated by three different workers.

To optimize the annotation process and obtain more labels we applied a type of distant supervision, which is a popular technique for obtaining more labels for the data (Go et al., 2009). In our study we used the similarity between the tweets because a large amount of tweets are posted again (retweets). Therefore, first we ranked the tweets by how may times they were retweeted and then we collected annotations for the most popular ones. Next, we disseminated the labels to the rest of the tweets using a similarity threshold set to 0.8. We used cosine similarity to measure the similarity between two tweets.

For all the results reported to this study, we used 429 tweets as a test set which were posted on July 2, 2015. We kept the test and training data always separated.

4.2 Baselines

We used two different baselines to compare the performance of our approach. The first baseline (SLM) is based on sentiment language models and was built from all the data without favoring tweets that occurred at spikes (i.e. $\lambda = 1.0$). The second baseline is the random-SLM approach. In this case, instead of using tweets from recent sentiment spikes, we randomly chose tweets from the whole collection. To build the random LM we select as many tweets as those used to build the LM of sen-

[1] `https://www.crowdflower.com/`

timent spikes. To evaluate the statistical significance of differences we used the McNemar test.

4.3 Experimental Settings

For pre-processing, we removed URLs, mentions, punctuation and the entity-related terms *Michelle* and *Obama*. For the experiments we used only unigrams. To overcome the problem of assigning zero probabilities to unseen terms we used Jelinek-Mercer smoothing with $\mu = 0.1$.

To model the evolution of sentiment and detect any sentiment spikes we split the data hourly. In addition, we defined temporal bins with the size of 8 hours. For the *emerging-SLM* we detected all the sentiment spikes that occurred in the last two days. To detect the spikes, we used the peakutils[2] package setting the threshold to 0.8.

Finally, to tune the λ parameter, we used cross-validation on a rolling basis. Following this approach, we used data published on the first temporal bin as training and data of the second temporal bin as test. Next, data from the first and second temporal bins were used for training and data from the third temporal bin as test and so on. In other words, when we set the number of bins to 9, it means that we were using 3 days as training data (i.e. 8 hours * 9 bins = 72 hours) and one bin as test data. The test bin was always the adjacent temporal bin. The first setup included 3 days, since we wanted to have enough data to build the SLMs. After this process we estimated the best λ parameters using the average performance.

5 Results

Figure 2 shows the performance with regards to the F1-measure for the task of emotion detection using the emerging-SLM on the training data for the different parameters of λ. We show the results for six different temporal bins for reasons of clarity. From this figure, we observe that there is a performance improvement as the λ parameter increases.

Table 1 shows the performance with regards to F1-measure for the task of emotion detection using the emerging-SLM, the SLM and the random-SLM approaches. From the results we observe that the emerging-SLM performs better compared to SLM and random-SLM for all the temporal bins. Also, most of the differences are significant. These results are very important because they show that favoring tweets that have occurred at recent sentiment spikes is very useful. Also, the improvement over the random-SLM validates further this assumption. The results are also shown on Figure 3 for an easier comparison.

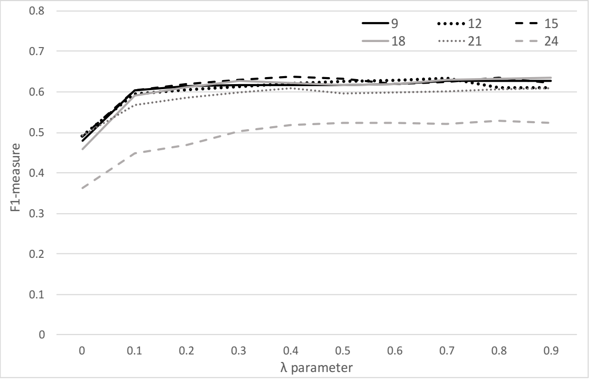


Figure 2: Performance of the *emerging-SLM* on the training data with regards to F1-measure for different λ parameters for six different bins.

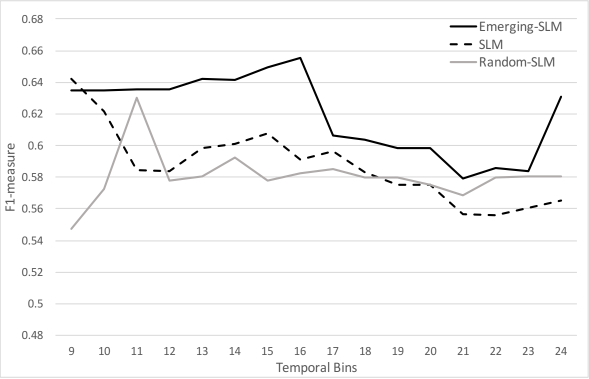


Figure 3: Performance measure with regards to F1-measure using different temporal bins.

6 Conclusions and Future Work

In this paper, we proposed the emerging-SLM approach to detect joy, neutral and anger emotions from positive, neutral and negative sentiments respectively. Emerging-SLM favors tweets that occur at recent sentiment spikes. The results showed that our approach performs better compared to both SLM and random-SLM and can be effectively applied to detect specific emotions.

In future we plan to explore if there is any effect of the temporal bins size on the emotion detection performance and if sentiment language models can be used to detect also other emotions such as fear and surprise.

[2] https://bitbucket.org/lucashnegri/peakutils

Table 1: Performance results over the test data using different size for the temporal bins of the emerging-SLM, SLM and random-SLM approaches. A star (∗) means there is a statistically significant difference between the emerging-SLM and SLM (p<0.05). A † indicates a significance difference between the emerging-SLM and random-SLM (p<0.05).

Bins	emerging-SLM	SLM	random-SLM
9	0.6352†	0.6419	0.5473
10	0.6352∗†	0.6213	0.5725
11	0.6358∗†	0.5842	0.6303
12	0.6358∗†	0.5841	0.5781
13	0.6419∗†	0.5980	0.5803
14	0.6415∗†	0.6012	0.5922
15	0.6493∗†	0.6078	0.5780
16	0.6551∗†	0.5914	0.5821
17	0.6064∗†	0.5961	0.5853
18	0.6034∗†	0.5828	0.5798
19	0.5987∗†	0.5751	0.5797
20	0.5987∗†	0.5751	0.5750
21	0.5792∗†	0.5564	0.5683
22	0.5855∗†	0.5557	0.5796
23	0.5837∗†	0.5604	0.5804
24	0.6311∗†	0.5651	0.5805

Acknowledgement

The work of the first author was partially funded by the Swiss National Science Foundation (SNSF) under the project OpiTrack.
The work of the last author was in the framework of the Spanish MINECO research project SomEMBED (TIN2015-71147-C2-1-P).

References

Xiaoran An, R. Auroop Ganguly, Yi Fang, B. Steven Scyphers, M. Ann Hunter, and G. Jennifer Dy. 2014. Tracking climate change opinions from twitter data. In *Proceedings of the Workshop on Data Science for Social Good held in conjunction with KDD 2014*.

Jing Bai, Jian-Yun Nie, and François Paradis. 2004. Using language models for text classification. In *Proceedings of the Asia Information Retrieval Symposium, Poster Session*, AIRS '04.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 450–453.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3/4):169–200.

Delia Irazú Hernańdez Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology*, 16(3):19:1–19:24.

Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys*, 49(2):28:1–28:41.

Anastasia Giachanou, Ida Mele, and Fabio Crestani. 2016. Explaining sentiment spikes in twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2263–2268.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Standford.

Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. 2012. Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 1678–1684.

Saif M. Mohammad. 2012. #Emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval '12, pages 246–255.

Saif M. Mohammad. 2015. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement*, pages 201–238.

Myriam D. Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, EMNLP '02, pages 79–86.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281.

Francisco Rangel and Paolo Rosso. 2016. On the impact of emotions on author profiling. *Information Processing & Management*, 52(1):73 – 92.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on twitter. In *Proceedings of the 8th International Language Resources and Evaluation Conference*, LREC '12, pages 3806–3813.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 1555–1565.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL '15, pages 1014–1023.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.

Stylometry in Computer-Assisted Translation: Experiments on the Babylonian Talmud

Emiliano Giovannetti[1], **Davide Albanesi**[1], **Andrea Bellandi**[1],
David Dattilo[2], **Felice Dell'Orletta**[1]
[1] Istituto di Linguistica Computazionale, Via G. Moruzzi 1, 56124, Pisa
name.surname@ilc.cnr.it
[2] Progetto Traduzione Talmud Babilonese S.c.a r.l., Lungotevere Sanzio 9, 00153 Roma
david.dattilo@talmud.it

Abstract

English. The purpose of this research is to experiment the application of stylometric techniques in the area of Computer-Assisted Translation to reduce the revision effort in the context of a collaborative, large scale translation project. The obtained results show a correlation between the editing extent and the compliance to some specific linguistic features, suggesting that supporting translators in writing translations following a desired style may actually reduce the number of following necessary interventions (and, consequently, save time) by revisors, editors and curators

Italiano. *Lo scopo di questa ricerca è la sperimentazione dell'applicazione di tecniche stilometriche nell'area della Traduzione Assistita dal Calcolatore per ridurre il lavoro di revisione nel contesto di un progetto di traduzione collaborativo di ampia scala. I risultati ottenuti mostrano una correlazione tra l'entità delle modifiche effettuate e la conformità ad alcune specifiche caratteristiche linguistiche, suggerendo che supportare i traduttori nel processo traduttivo seguendo uno stile desiderato possa effettivamente ridurre il numero di interventi necessari (e, quindi, risparmiare tempo) da parte di revisori, redattori e curatori.*

1 Introduction

The Progetto Traduzione Talmud Babilonese[1] (PTTB) is a research and education project carrying out the digitized Italian translation of the Babylonian Talmud (BT), a fundamental book of the Jewish tradition, covering every aspect of human knowledge: law, science, philosophy, religion and even aspects of everyday life. The translation of the Talmud has been assigned to more than 50 scholars comprising expert translators, trainee translators, instructors, editors and curators.

The translated text is accompanied by the explanations and comments on specific words and subjects, and also by illustrative sheets for the various scientific, historical and linguistic topics addressed inside the Talmudic discussions. However, the Project objectives include more than the translation of the Talmud: the whole work has been set up to be completely digital. Everything, from the very first activities of assigning users to the translation of specific chapters to supporting in the definition of the final printing layout, revolves around Traduco, a collaborative web-based Computer-Assisted Translation (CAT) tool developed within the Project.

Today, many CAT tools, both commercial and freely distributed, are already available, but they have been designed for the translation of technical manuals or domain-specific texts (legislative, medical) with the main purpose of speeding up the translation process.

The BT is a very complex text in many ways: its content, the different, ancient, languages it is composed of (though mainly Babylonian Aramaic and Mishnaic Hebrew), and the history of its composition over the centuries. For these reasons, the approach we adopted for the development of Traduco had to take into account the needs of translators working on a text with very particular interpretative issues. Traduco allows a user to distinguish the literal part of the translation (in bold, see Fig.1) from explicative additions, included by translators to make the most difficult passages clearer to readers. Indeed, a full understanding of this kind of texts requires a translation

[1] www.talmud.it (last access: 25/07/2017)

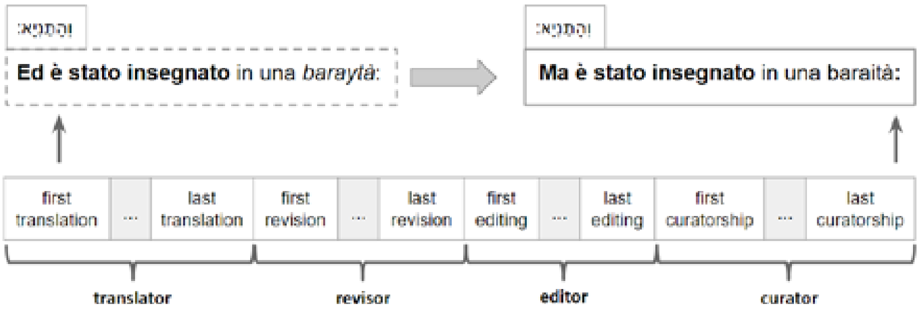


Figure 1: The life cycle of a translated string.

to be enriched with comments, notes, and glossary entries. Furthermore, due to the complexity of the inner structure of the BT, Traduco allows users to split autonomously their translations into "strings" (representing, typically, a sentence, see Fig.1), gathered into "logical units" . Finally, Traduco provides a collaborative and training environment allowing a translator to instantly consult translations done by others, when portions of text (and sometimes even a single word) are difficult to interpret and translate. For a comprehensive description of how Traduco works refer to (Giovannetti et al., 2017). The size and complexity of the text and the need to produce a printed version of the BT translation, required a team of users composed of translators, revisors, editors, curators and supervisors.

The whole translation workflow can be described by following the "life-cycle" of each string (Fig.1). It all starts as soon as the coordinator of the translation assigns a chapter to a specific translator: the first phase of the work, the **translation**, begins. The translation is carried out by scholars having two distinct profiles: expert translators, working autonomously, and trainee translators, these latter being constantly supported by instructors monitoring online their work and providing face-to-face lectures. Once the translation of a specific chapter is concluded, the **revision** phase starts. Revisors are chosen among the most expert scholars involved in the Project and their main task is to verify if translators have understood correctly the meaning of each string. They also have to check if the domain terms (if present) have been appropriately annotated and explained in the relative glossary entry. After the content has been revised, the **editing** starts. In this phase, a formal and linguistic control of the translation is carried out, where the editors ensure that the translated strings are syntactically and orthographically correct. Contextually, each string can be enriched, if needed to help in the understanding of the text, with pictures and tables. The last phase is the **curatorship**, during which one more general control of the translation is done before proceeding with the final exporting and printing of the volume. As we showed in a previous work (Bellandi et al., 2016), the introduction of Natural Language Processing techniques in CAT tools can bring concrete advantages to the translation work and pave the way to innovative research in the area of NLP for Digital Humanities.

One way to ease the translation of a text as the BT is to assist translators in writing, in the first place, good translations requiring as few corrections as possible by revisors, editors and curators. In other words, we want to find a way of alerting a user about to submit a new translation by highlighting specific characteristics of the sentence that may further require a revision and, thus, slow down the overall translation process.

To do that, we chose to experiment the application of stylometric measures to Italian translations. The assumption we would like to prove is that translations being more compliant to the style of revisors will actually require less revisions. If that will be demonstrated, we may develop a strategy to alert translators of potential "unfit" translations and suggest a way to improve them in order to minimize the following editing for revision, editing, and curatorship.

2 Background

Over the last ten years, Natural Language Processing (NLP) techniques combined with machine learning algorithms started being used to investigate the "form" of a text rather than its content. The range of tasks sharing this approach to the analysis of texts is wide, ranging e.g. from native language identification (see among the others (Koppel et al., 2005) and (Wong and Dras, 2009)), author recognition and verification (see e.g. (van Halteren, 2004), authorship attribution (see (Juola, 2008) for a survey), genre identification (Mehler et al., 2011) to readability assessment (see (Dell'Orletta et al., 2014) for an updated survey) or tracking the evolution of written language competence (Richter et al., 2015). Besides obvious differences at the level of the considered task, they share a common approach: they succeed in determining the language variety, the author, the text genre or the level of readability of a text by exploiting the distribution of features automatically extracted from texts. To put it in van Halteren words (van Halteren, 2004), they carry out "linguistic profiling" of texts, i.e. "the occurrences of a large number of linguistic features in a text, either individual items or combinations of items, are counted" in order to determine "how much [...] they differ from the mean observed in a profile reference corpus".

To the best of our knowledge, however, no research has been documented in literature about the application of stylometric or readability techniques to Computer-Assisted Translation. For this reason, a comparison with existing approaches and results was not possible.

On the other hand, the use of stylometry and readability in translation studies is described in several works, especially in the analysis of literary texts (Heydel and Rybicki, 2012), (Kolahi and Shirvani, 2012), (Acar and İŞİSAĞ, 2017), (Huang, 2015) and some of them provide useful indications on how the personal writing style (being it, in our case, that of a translator or a revisor) can influence the final translation (Baker, 2000) and (Rybicki, 2012).

3 Methodology

To construct the dataset we exploited the versioning features of Traduco. As a matter of fact, every version of most of textual resources (currently: strings, notes, and glossary entries) is stored in the database. It is thus possible to compare earlier versions of translations (i.e. those inserted by translators) with the latest ones (i.e. those that have been completely revised) in order to analyse the differences between them. For the experiment, we built two datasets using textual segments of different granularity: blocks for the DS_{bl} dataset and logical units for DS_{lu}.

In more details, each dataset has been built as a set of textual segment pairs extracted from the translations of the tractates Berakhot and Ta'anit, respectively composed, in their revised versions, of 216138 and 81696 tokens. Given a pair (s_1, s_2), the first component s_1 represents the last translation of a block or logical unit inserted by the translator[2] and the second component s_2 its very last version (i.e. that following the revision, editing and curatorship phases). Concerning the size, DS_{bl} was composed of 554 blocks and DS_{lu} of 4303 logical units. Each logical unit is composed, in average, by 5.62 strings, while each string is composed, in average, by 12.5 tokens.

Once the datasets were ready, we had to attribute to each pair a "revision measure" to quantify the difference between s_1 and s_2 in terms of both words and characters. For this purpose we chose to adopt the Levenshtein distance. Since Traduco is equipped with a spell checker, we assumed that the presence of typos should not impact on the revision measure significantly.

As the next step we investigated the presence of linguistic features extracted from those texts belonging to the s_1 component of the pairs correlating with the revision measures. For this purpose, the considered texts were automatically POS tagged by the Part-Of-Speech tagger described in (Cimino and Dell'Orletta, 2016) and dependency parsed by the DeSR parser (Attardi et al., 2009) using multilayer perceptron as learning algorithm. For the specific concerns of this study, we focused on a wide set of features ranging across different linguistic description levels which are typically used in studies focusing on the "form" of a text, e.g. on issues of genre, style, authorship or readability. This represents a peculiarity of our approach: we resort to general features qualifying the lexical and grammatical characteristics of a text, rather than ad hoc features, specifically selected for a given text type or task. The set of selected features is organised into four main categories defined on the basis of the different levels of linguistic analysis automatically carried out (tokenization, lemmatization, morphosyntactic tagging and dependency parsing): i.e. raw text features, lexical features as well as morpho-syntactic and syntactic features.

[2] sometimes translators insert a draft version of a translation, to be completed later: for this reason we chose to take the last translation available.

	$\mathbf{DS_{lu}}$		$\mathbf{DS_{bl}}$	
features	**char**	**token**	**char**	**token**
Number of tokens	**0.65**	**0.68**	**0.84**	**0.85**
Arity of verbs	**0.62**	**0.64**	**0.83**	**0.83**
Number of main verbs	**0.62**	**0.64**	**0.83**	**0.83**
Number of prepositional 'chains'	**0.57**	**0.60**	**0.81**	**0.82**
Number of sentences	**0.49**	**0.53**	**0.80**	**0.80**
Number of verb roots	**0.49**	**0.53**	**0.79**	**0.79**
Number of subord clauses	**0.37**	**0.38**	**0.68**	**0.68**
% of verbs with 5 syntactic dependent	-	-	**0.37**	**0.36**
% of first person singular of verbs	-	-	**0.31**	0.32
% of subjunctive auxiliary-verbs	-	-	0.31	0.30
% of locative modifier	-	-	0.31	0.31
% of second person plural	-	-	0.31	0.31
% of verb in infinitive mood	-	-	0.30	**0.32**
% of demonstrative determiner	-	-	0.30	-
% of "balanced" punctuation	-	**0.33**	-	-
Average of length of dependency links	**0.35**	**0.37**	-	-
Longest dependency links	**0.34**	**0.34**	-	-
Average of main verbs for sentence	**0.33**	**0.32**	-	-
Average length of subord clauses	**0.31**	**0.31**	-	-

Table 1: Spearman's rank correlation coefficients (in bold with $p < 0.001$, otherwise with $p < 0.05$) calculated on both datasets and the two revision measures (distance per character and per token); values below 0.3 have been discarded.

To conclude our experiment we applied the Spearman's rank correlation coefficient to assess the presence of a statistical dependence between our revision measures and the calculated linguistic features.

4 Evaluation

The results (filtered by keeping just the features providing coefficients greater or equal than 0.3) are summarized in Table 1. Apart from the expected correlations between the size of the texts (represented by raw text features such as "Number of tokens" and "Number of sentences") and the revision measures, we found some significative correlations, in relation to morphosyntactic and syntactic features. Most of the morphosyntactic features involve verbs: the presence of main verbs, the mood, the tense, etc.

Some of the syntactic features showing a correlation, such as the length of dependency links, the length of subordinate clauses and the number of prepositional chains, are particularly interesting. As a matter of fact, these linguistic features are typically used as indicators of linguistic complexity: indeed, portions of translated text constituted of long and articulated syntactic structures appear to be more subjected to revisions. As expected, the correlation of some of these syntactic features, such as the number of prepositional chains, appears to be proportional to the size of the analysed text (as in the blocks wrt the logical units in the datasets), since the presence of deeper syntactic structures increases and the text, at least in principle, gets more linguistically complex.

5 Conclusions

The experiment described in this paper proves that the application of NLP to CAT contexts can open new research perspectives and, more importantly, may be of concrete help in real usage translation scenarios. The proposed methodology can be applied, in principle, to any translation project in which a revision phase is a part of the whole translation workflow and where an history of the edits is maintained. The same analysis could be performed on different languages depending solely on the availability of the suitable NLP tools. Some of the NLP techniques adopted for the stylometric analysis of Italian may also be adapted to the processing of Mishnaic Hebrew and Aramaic (the main source languages). The automatic linguistic analysis of Mishnaic Hebrew, for example, is being experimented (Pecchioli, 2017). However, an analysis of the style (or complexity) of the source text, though interesting in a historical text analysis perspective, would be pointless in the specific context of revision support in computer-assisted translation.

The correlation we found between the revision measures and some linguistic features (some of which are actually used as indicators of linguistic complexity) is the first step towards the design of a technique aimed at providing users a way of writing translations less prone to revisions. In this way, the whole translation workflow would benefit from a reduced time in the revision, editing and curatorship phases. Once the approach will be defined, the relative software will be implemented as a new component of Traduco. Moreover, the possibility of suggesting a way of writing "better" translations (at least wrt revisor's style) will be exploited in the education of trainee translators.

6 Acknowledgment

This work was partially supported by the project TALMUD and carried out in the context of the scientific partnership between S.c.a r.l. "Progetto Traduzione del Talmud Babilonese" and ILC-CNR and on the basis of the regulations stated in the "Protocollo d'Intesa" (memorandum of understanding) between the Italian Presidency of the Council of Ministers, the Italian Ministry of Education, Universities and Research, the Union of Italian Jewish Communities, the Italian Rabbinical College, and the Italian National Research Council (21 January 2011).

References

Alpaslan Acar and Korkut Uluç İŞİSAĞ. 2017. Readability and comprehensibility in translation using reading ease and grade indices. *International Journal of Comparative Literature and Translation Studies*, 5(2):47–53.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, and Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of the 2nd Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, (EVALITA 2009).*

Mona Baker. 2000. Towards a methodology for investigating the style of a literary translator. *Target. International Journal of Translation Studies*, 12(2):241–266.

Andrea Bellandi, Giulia Benotto, Gianfranco Di Segni, and Emiliano Giovannetti. 2016. Investigating the application and evaluation of distributional semantics in the translation of humanistic texts: a case study. In *Proceedings of the* 2^{nd} *Workshop on Natural Language Processing for Translation Memories*, pages 6–11.

Andrea Cimino and Felice Dell'Orletta. 2016. Building the state-of-the-art in POS tagging of italian tweets. In *Proceedings of Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. In John Benjamins Publishing Company, editor, *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics, 165:2*, pages 163–193.

Emiliano Giovannetti, Davide Albanesi, Andrea Bellandi, and Giulia Benotto. 2017. Traduco: A collaborative web-based cat environment for the interpretation and translation of texts. *Digital Scholarship in the Humanities*, 32(suppl_1):i47–i62.

Magda Heydel and Jan Rybicki. 2012. The stylometry of collaborative translation. woolf's night and day in polish. In *Digital Humanities 2012 Conference Abstracts*, pages 212–217.

Libo Huang. 2015. Readability as an indicator of self-translating style: A case study of eileen chang. In *Style in Translation: A Corpus-Based Perspective*, pages 95–111. Springer.

Patrick Juola. 2008. Authorship attribution. In *Now Publishers Inc.*

Sholeh Kolahi and Elaheh Shirvani. 2012. A comparative study of the readability of english textbooks of translation and their persian translations. *International Journal of Linguistics*, 4(4):344.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics, vol. 3495, LNCS, Springer–Verlag*, pages 209–217.

Alexander Mehler, Serge Sharoff, and Marina (Eds.) Santini. 2011. Genres on the web. computational models and empirical studies. In *Springer Series: Text, Speech and Language Technology*.

Alessandra Pecchioli. 2017. Elaborazione del linguaggio naturale (nlp) in ebraico: il caso dell'analisi linguistica automatica applicata all'ebraico mishnaico del talmud. Oral communication, sep. XXXI Convegno AISG 2017 - Nuovi studi sullEbraismo, 4-6 settembre 2017, Ravenna, Italy.

Stefan Richter, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2015. Tracking the evolution of written language competence: an nlpbased approach. In Cristina Bosco, Sara Tonelli, and Massimo Zanzotto, editors, *Proceedings of the Second Italian Conference on Computational Linguistics - CLiC-it 2015*, pages 236–240.

Jan Rybicki. 2012. The great mystery of the (almost) invisible translator. *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, 231:231–248.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In John Benjamins Publishing Company, editor, *Proceedings of the Association for Computational Linguistics (ACL04)*, pages 200–207.

Sze-Meng Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*.

Towards an Italian Lexicon for Polarity Classification (polarITA): a Comparative Analysis of Lexical Resources for Sentiment Analysis

Delia Irazú Hernández Farías
PRHLT Research Center
Universitat Politècnica de València
dhernandez1@dsic.upv.es

Irene Laganà
Dipartimento di
Studi Umanistici
Università di Pavia
irene.lagana01@universitadipavia.it

Viviana Patti, Cristina Bosco
Dipartimento di Informatica
Università di Torino
{patti,bosco}@di.unito.it

Abstract

English. The paper describes a preliminary study for the development of a novel lexicon for Italian sentiment analysis, i.e. where words are associated with polarity values. Given the influence of sentiment lexica on the performance of sentiment analysis systems, a methodology based on the detection and classification of errors in existing lexical resources is proposed and an extrinsic evaluation of the impact of such errors is applied. The final aim is to build a novel resource from the filtering applied to the existing lexical resources, which can integrate them with missing lexical entries and more reliable associations of polarity with entries.

Italiano. *L'articolo descrive uno studio preliminare per lo sviluppo di una nuova risorsa lessicale per la sentiment analysis in italiano, i.e. dove alle parole sono associati valori di polarità. Data l'influenza dei lessici di sentiment sulle performance dei sistemi di sentiment analysis, viene proposta una metodologia basata sulla rilevazione e classificazione degli errori presenti nei lessici attualmente disponibili ed una valutazione estrinseca dell'impatto di tali errori sui sistemi. L'obiettivo finale è ottenere un nuovo lessico grazie ad un filtraggio applicato alle risorse lessicali disponibili, e a un'integrazione con le voci lessicali mancanti, ottenendo una maggiore affidabilità nell'associazione delle polarità alle voci.*

1 Introduction

Sentiment Analysis (SA), described as the task of automatically determine the polarity in a given piece of text (Mohammad, 2016), is currently among the most widely investigated topics within NLP. Overall, the approaches for addressing such task are mainly based on techniques ranging from traditional machine learning to novel deep learning ones, as it can be seen also in the context of shared tasks on sentiment polarity classification in Twitter recently proposed, respectively for English (Nakov et al., 2016) and Italian (Barbieri et al., 2016), within the SemEval and Evalita periodical evaluation campaigns. Moreover, the detection of specific words associated with polarity values or emotions has been considered as a powerful information source for identifying the sentiment behind a text. Among the resources which are more commonly exploited by SA systems for performing their task there are therefore sentiment lexica, i.e., lists of words with associated polarity values or emotions.

Several techniques have been applied for the development of lexical resources for SA: they can be built from scratch, manually or automatically, or extracted from corpora (Nissim and Patti, 2017). Nevertheless, the vast majority of these resources are written in English, and a lack of resources currently features several other languages. One of the most commonly applied alternatives for having resources in language other than English is to automatically translate some available English lexicon via tools such as Google translate[1]. But there are many constraints involved in this kind of process, such as handling synonyms and polysemous words, multi-word expressions, but also to deal with cultural differences between source and target language. Apart from this, possible variations of polarity across different contexts and languages should be carefully taken into account, while such approaches rely somehow on the assumption that affective norms related to sentiment are stable across languages.

[1] https://translate.google.com/

In this paper we are interested into evaluate the reliability of the lexical resources currently available for Italian SA and, providing that the most of them are obtained by translation, we will mainly focus on the reliability of automatically translating English resources to Italian language. For doing so, we carried out a methodology involving different facets. Our final aim is to develop a new SA resource for Italian, which comprises pre-existing translated lexical entries enriched with the manual correction of the polarity assigned, as resulting from our analysis, but also includes entries which are featured by a polarity but are missing in the available lexica.

The paper is organized as follows. In the next section, we describe our methodology which mainly consists in three steps: the selection of a sample of tweets from an Italian sentiment corpus and exploited as part of the gold standard in the Sentipolc@Evalita2016 shared task (Stranisci et al., 2016; Barbieri et al., 2016); automatic extraction of the lexical entries polarized according to a set of benchmark sentiment lexica for Italian; the analysis of these entries and the comparison with those expected by a human judge. Section three shows instead an extrinsic evaluation of the impact of the detected errors on the results of the SA system. Some hints about future development of this research are given in the conclusion.

2 Our Methodology

Given the relevance of affective lexica in SA and related tasks, our major aims in the current research are to detect the limits of the currently available lexical resources for Italian and to explore the possibility to develop a novel resource by correcting and extending them. In this paper we focus in particular on the detection of the deficiencies of existing resources and on their motivations. Our methodology consists therefore in: (i) selecting of a sample of tweets from an Italian sentiment corpus featured by political contents (Stranisci et al., 2016) and exploited as part of the gold standard in the Sentipolc@Evalita2016 shared task (Barbieri et al., 2016), with sentiment polarity annotation at the tweet level; (ii) automatically extracting the lexical entries polarized according to a set of benchmark sentiment lexica for Italian and (iii) manually checking the results for each expected lexical entry in the context of the whole tweet (i.e. if the polarity of the entry is that expected by a human annotator or also if there are other entries in the tweet that should appear as polarized but are not in the lexicons).

We take as starting point the SA lexica exploited by (Hernández Farías et al., 2014) in the IRADABE system at Evalita2014's SENTIPOLC (Basile et al., 2014). The same resources where used also in the upgraded system that participated at the same task in Evalita2016 (Buscaldi and Hernández Farías, 2016).

In those works the lexicon AFINN, (Nielsen, 2011), the one developed by Hu and Liu (henceforth HaL) (Hu and Liu, 2004), and SentiWordNet (SWN) (Baccianella et al., 2010) were indeed automatically translated to Italian, to exploit obtained information as features in their supervised system, but no specific evaluation or refining of them was performed. In the present paper we extend our selection by considering, beyond these three, a further resource, i.e. Sentix (Basile and Nissim, 2013) (see Sec. 2.1) which has been developed following a semantics oriented strategy (see Sec. 2.1). Henceforth, we will use the expression *benchmark lexica*) for referring to the four resources. As reference corpus, we considered, instead, TwBuonaScuola (Stranisci et al., 2016), an Italian dataset manually annotated for sentiment polarity and irony, focused on the on-line debate regarding a controversial Italian political reform, which is part of the gold standard provided for the Sentipolc shared task (Barbieri et al., 2016) at Evalita 2016 (Basile et al., 2017).

Our methodology, whose results are shown in Sec. 2.2, includes the steps described below.

Given a random selection of 500 tweets from TwBuonaScuola (henceforth *ItalianTweets*) including 2,706 different words, we manually evaluated the coverage of the *benchmark lexica* for the words included in these tweets. In particular, for each tweet we extracted automatically all the words which are included in each of the *benchmark lexica* and its associated polarity.

Then, for each tweets belonging to *ItalianTweets*, we manually checked the obtained lists of words, considered in the context of the tweet, with a twofold objective:

(i) To deduce which words in the *benchmark lexica* have a wrong polarity associated;

(ii) To identify those words that express certain polarity in the corpus but are not included in the *benchmark lexica*.

2.1 Sentiment Analysis Resources

In this section we describe the *benchmark lexica*.

AFINN (Nielsen, 2011) is an English lexicon composed of 2,477 words and 15 multi-word expressions. Each entry is associated with a score which varies from -5 to +5 in order to respectively introduce negative and positive polarity. The starting point for the development of this resource is a list of obscene words and some positive words; then the lexicon has been extended with words from a corpus of tweets and other lists of words from Urban Dictionary[2] for representing entries typical of Internet language (e.g. "WTF" and "LOL"). After the manual annotation of the entries the lexicon has been evaluated based on a corpus of tweets manually annotated for SA.

HaL, (Hu and Liu, 2004), has been built within a project for developing methods to deal with opinions expressed in reviews about various kinds of goods. A group of 30 adjectives featured by a single and stable polarity and manually annotated has been expanded by including the words which in WordNet's synsets are synonyms or antonyms of these seeds, providing that synonyms are featured by the same polarity and antonyms by the opposite one. The lexicon currently includes 6,800 entries classified as positive or negative.

SentiWordNet 3.0 (Baccianella et al., 2010) is among the larger and more used resources exploited for SA. The main goal of the SentiWordNet project is the fully automated annotation of the polarity of the WordNet's synsets using scores that vary from 0.0 to 1.0 to each of the three basic polarity values (positive, negative, neutral) in order to obtain 1 as the sum of them. By contrast with the other resources, SentiWordNet takes into account different possible senses for each word.

As far as Italian is concerned, only a few resources exist, such as Sentix (Basile and Nissim, 2013) and SABRINA (Borzì et al., 2015). Sentix is the result of the alignment of four semantic database, namely WordNet (Fellbaum, 1998), SentiWordNet, MultiWordNet (Pianta et al., 2002) and Babelnet (Navigli and Ponzetto, 2012). The methodology consists in transferring to the Italian section of WordNet the information about polarity encoded in the English SentiWordNet's synsets, thus aligning Italian and English synsets.

The development of SABRINA instead is based on the application of a prior polarity method on two sets of Italian words, the first composed of 277,000 entries with associated inflexion. However the lexicon is not publicly available.

Finally let us mention ItEM (Passaro et al., 2015), an Italian emotive lexicon which aims at offering information about affect expressed in text according to finer levels of granularity, i.e. referring not simply to positive or negative sentiment polarity but to emotional categories. In ItEM each word is tagged with an emotional label from the height basic emotions of the Plutchik's psychological model (Plutchik, 1980).

Several scholars are devoting their efforts to the development of resources for other languages, by applying translation or other methodologies. Let us cite e.g. FEEL (Abdaoui et al., 2017), a French lexicon where words are associated with polarity and emotions obtained thanks to the application of translation tools to NRC-EmoLEx[3] and a manual validation of results.

2.2 Qualitative Analysis of *Benchmark Lexica*

In order to detect the coverage and correctness of each *benchmark lexicon*, we selected from our reference sample corpus the list of words that according to a human judge are featured by some affective value in the context of the tweet where they appear. Then, for each entry of this list and for each *benchmark lexicon*, we observed if the word is represented in the resource and featured by the same polarity.

Given the preliminary nature of this investigation only a couple of researchers have been involved in the task. Moreover, a further limit of our current research approach depends on the reference to a given context (that determined by our sample corpus); issues related to the context will be accounted for in future investigations.

We observed different coverages of the *benchmark lexica* on our Twitter corpus, first of all in terms of numbers of affective words occurring in the tweets for each lexicon. The full vocabulary of the tweets is composed of 2,706 different words. Only some of these words are featured by some affective value, and focusing on them only we observed the following occurrences: 160 words in AFINN, 190 words in HaL, 302 words in SWN and 551 in Sentix. These word sets are partially overlapped, since 69 words are included in all the

[2] http://www.urbandictionary.com

[3] http://www.saifmohammad.com/WebPages/lexicons.html

lexica.

Resource	Error			
	(i)	(ii)	(iii)	(iv)
AFINN	1.2	2.5	16.8	8.7
HaL	1.5	1.0	12.6	12.6
SWN	5.9	1.6	15.5	13.2
Sentix	5.9	2.1	15.2	16.6

Table 1: Distribution of different errors in the *benchmark lexica* (percentage wrt the coverage of the lexicon).

The total amount of words missing or with an attributed erroneous polarity in the *benchmark lexica* is 388. As far as the erroneous polarization concerns, as summarized in Table 1, these words are featured by four different kinds of errors: (i) a positive word is annotated as negative; (ii) a negative word is annotated as positive; (iii) a neutral[4] word is annotated as positive; and (iv) a neutral word is annotated as negative. The values are expressed in percentage with respect to the coverage of the lexica. As far as the distribution of errors in the four classes, they are for all lexica prevailingly distributed in the last two classes, i.e. iii and iv, laying foundation for the hypothesis that in the automatic transition between English and Italian several non (clearly) polarized Italian words were instead polarized.

Nevertheless, observing Table 1, we can see also that all the lexica are featured by very similar amounts of errors, regardless of the methodology applied for their development (i.e. translation or extraction from semantic databases). Several errors, in particular for what concerns the polarity associated to specific words, can be generated during translation, and a portion of them is therefore motivated by the application of translation tools mainly because they do not consider context where each word occurs. But observing the results extracted from Sentix, which is not obtained simply by translation, and weighting the larger coverage that features this resource, we can see that errors occurs in a percentage that positively compares with that of the other resources. In this case the problem probably depends on misalignment of synsets for different languages. For example, the Italian word "istituto", whose meaning can be "school" or "institution", is aligned with "prison" and "house/prison", with a negative polarity which is not appropriate for the Italian word.

Several errors could be probably avoided in the transition among languages by applying a pre-processing including Part of Speech tagging and considering the grammatical category of the source and target terms. See for instance, the word *tagliando* (cutting) that occurs in the corpus as a Verb and in the *benchmark lexica* is instead aligned with the corresponding noun with the meaning of voucher/coupon. This motivates our decision about the attribution of PoS tags to the words in the first nucleus of a novel resource obtained by extending and correcting the existing ones. The overall impression is that, a manual check, even is a very time-consuming task, is always necessary and unavoidable, both when the new lexicon is obtained by translation, and when it is obtained relying on synset alignment.

3 Lost in Translation: Impact of the Errors

The methodology even if applied on a small set of tweets and based on a manual check of the *benchmark lexica*, confirms the hypothesis that many directions can be followed to improve the quality of existing lexical resources. The first result of this preliminary analysis is the collection of a list of words with associated polarity which will be the nucleus of the novel resource, i.e. polarITA. Each of the words in polarITA has been annotated with an overall polarity value (i.e., positive, negative, or none), and its corresponding Part-Of-Speech (POS) label. Table 2 summarizes the distribution of the words in polarITA in terms of polarity and POS labels.

Experiments on a larger corpus and a quantitative analysis based on a more formal classification of errors is needed for the development of a fully developed reliable lexical resource, together with an in-depth investigation of the relevance of context in the attribution of polarity, which is a very important issue. A comparison of the results that a given SA engine exploiting features extracted from sentiment lexica, for instance IRADABE (Hernández Farías et al., 2014; Buscaldi and Hernández Farías, 2016), obtains using each of the *benchmark lexica* and using polarITA is planned as future work for the evaluation of the novel lexicon, which is not currently suitable because the

[4] We considered neutral a word which is featured by a polarity which may vary across contexts, indicated by None in Table 2.

limited size of our reference corpus and the consequent partial coverage of errors.

Considering the current preliminary stage of development of polarITA, we tried an extrinsic evaluation for detecting the impact on the performance of SA systems of the errors currently featuring the *benchmark lexica* and corrected in the novel lexicon. We compared the words which are missing or assigned to erroneous polarity in the *benchmark lexica* with the Italian words more commonly used and understood by native speakers, whose collection is available in the *Vocabolario di base della lingua italiana (vocItalian)*[5] recently newly released. Like the first version of this resource, published in 1980, (De Mauro, 1980), it includes three word classes: 2,999 High Usage words (HU), 2,231 High Availability words (HA) and 1,979 Foundational words (FO).
In polarITA we collected until now 284 words of the vocItalian, whose distribution across the three classes is shown in Table 2. Among the words in the FO category we found "bene" (good), "mentire" (lie), and "giustizia" (justice). While words like "assassino" (killer), "preoccupato" (worried), and "entusiasta" (enthusiastic) are part of the HU category. Finally, in the HA category it is possible to find words such as "dannoso" (harmful) and "emozionante" (exciting).

This analysis suggests some hints for further investigation, showing that the failures of lexica currently available for Italian SA affect words very commonly used in communication and therefore the improvement of these resources may hopefully result in an advancement for SA and related tasks.

Total words	388

Polarity		
Positive	*Negative*	*None*
225	140	23

Part-of-speech labels	
Adjective	84
Adjective/Noun	1
Adjective/Pronoun	2
Adverb	16
Interjection	3
Noun	187
Noun/Adverb	1
Preposition	1
Pronoun	1
Verb	92

vocItalian	
FO	187
HU	86
HA	11

Table 2: Distribution of the words in polarITA in terms of polarity, POS labels, and vocItalian.

4 Conclusions and Future Work

In this paper we propose the preliminary investigation about a methodology for the development of a novel lexical resource for Italian SA, namely polarITA, which takes advantage of the analysis and filtering of errors occurring in the available lexical resources. We carried out a manual analysis of a set of tweets for determining the reliability of sentiment-related lexica, showing that, even if the transfer of lexical information between two different languages is a common practice to address the lack of resources, information related to sentiment is lost during it. The identified errors are then exploited as a starting point for developing the novel resource.
As future work, we are planning to extend the resource in several directions: by investigating multi-word expressions, extending the coverage to a larger corpus, exploring the impact of figurative language devices such as irony and sarcasm in the use of certain polarized words (Hernández Farías et al., 2016). Moreover, our future effort will be oriented to the automatization of a larger part of the methodology and its application to other languages currently under resourced.

Acknowledgements

C. Bosco and V. Patti were partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01) and by Fondazione CRT (*Hate Speech and Social Media*, 2016.0688).

References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. FEEL: a French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 51:833–855, September.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lex-

[5] `https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana`

ical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta. European Language Resources Association (ELRA).

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In Basile, Cutugno, Nissim, Patti, and Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR Workshop Proceedings.

Valerio Basile and Malvina Nissim. 2013. Sentiment Analysis on Italian Tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, USA. Association for Computational Linguistics.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, Pisa, Italy.

Pierpaolo Basile, Francesco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli. 2017. Evalita goes social: Tasks, data, and community at the 2016 edition. *IJCoL - Italian Journal of Computational Linguistics*, 3(1):93–127.

Valeria Borzì, Simone Faro, Arianna Pavone, and Sabrina Sansone. 2015. Prior Polarity Lexical Resources for the Italian Language. *CoRR*, abs/1507.00133.

Davide Buscaldi and Delia Irazú Hernández Farías. 2016. IRADABE2: Lexicon Merging and Positional Features for Sentiment Analysis in Italian. In *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*. aAcademia University Press.

Tullio De Mauro. 1980. *Guida all'uso delle parole Num. 3 dei Libri di base*. Editori Riuniti, Roma.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Delia Irazú Hernández Farías, Davide Buscaldi, and Belém Priego-Sánchez. 2014. IRADABE: Adapting English Lexicons to the Italian Sentiment Polarity Classification task. In *First Italian Conference on Computational Linguistics (CLiC-it 2014) and the fourth International Workshop EVALITA 2014*, pages 75–81.

Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Irony Detection in Twitter: The Role of Affective Content. *ACM Trans. Internet Technol.*, 16(3):19:1–19:24.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Saif M. Mohammad. 2016. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In Herb Meiselman, editor, *Emotion Measurement*. Elsevier.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, Heraklion, Crete, Greece. CEUR-WS.org.

Malvina Nissim and Viviana Patti. 2017. Semantic aspects in sentiment analysis. In Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu, editors, *Sentiment Analysis in Social Networks*, pages 31–48. Morgan Kaufmann, Boston.

Lucia Passaro, Laura Pollacci, and Alessandro Lenci. 2015. ItEM: A Vector Space Model to Bootstrap an Italian Emotive Lexicon. volume II.

E. Pianta, L. Bentivogli, and C. Girardi. 2002. MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of International Conference on Global WordNet*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In R. Plutchik and H. Kellerman, editors, *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, pages 3–33. Academic press, New York.

Marco Stranisci, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. 2016. Annotating Sentiment and Irony in the Online Italian Political Debate on #labuonascuola. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Multilingual Neural Machine Translation for Low Resource Languages

Surafel M. Lakew
University of Trento, Italy
Fondazione Bruno Kessler

Mattia A. Di Gangi
University of Trento, Italy
Fondazione Bruno Kessler

Marcello Federico
Fondazione Bruno Kessler

via Sommarive, 18, Trento, Italy
{lakew,digangi,federico}@fbk.eu

Abstract

Neural Machine Translation (NMT) has been shown to be more effective in translation tasks compared to the Phrase-Based Statistical Machine Translation (PBMT). However, NMT systems are limited in translating low-resource languages (LRL), due to the fact that neural methods require a large amount of parallel data to learn effective mappings between languages. In this work we show how so-called multilingual NMT can help to tackle the challenges associated with LRL translation. Multilingual NMT forces words and subwords representation in a shared semantic space across multiple languages. This allows the model to utilize a positive parameter transfer between different languages, without changing the standard attention-based encoder-decoder architecture and training modality. We run preliminary experiments with three languages (English, Italian, Romanian) covering six translation directions and show that for all available directions the multilingual approach, i.e. just one system covering all directions is comparable or even outperforms the single bilingual systems. Finally, our approach achieve competitive results also for language pairs not seen at training time using a pivoting (x-step) translation.

Italiano. *La traduzione automatica con reti neurali (neural machine translation, NMT) ha dimostrato di essere più efficace in molti compiti di traduzione rispetto a quella basata su frasi (phrase-based machine translation, PBMT). Tuttavia, i sistemi NMT sono limitati nel tradurre lingue con basse risorse (LRL). Questo è dovuto al fatto che i metodi di deep learning richiedono grandi quantit di dati per imparare una mappa efficace tra le due lingue. In questo lavoro mostriamo come un modello NMT multilingua può aiutare ad affrontare i problemi legati alla traduzione di LRL. La NMT multilingua costringe la rappresentrazione delle parole e dei segmenti di parole in uno spazio semantico condiviso tra multiple lingue. Questo consente al modello di usare un trasferimento di parametri positivo tra le lingue coinvolte, senza cambiare l'architettura NMT encoder-decoder basata sull'attention e il modo di addestramento. Abbiamo eseguito esperimenti preliminari con tre lingue (inglese, italiano e rumeno), coprendo sei direzioni di traduzione e mostriamo che per tutte le direzioni disponibili l'approccio multilingua, cioè un solo sistema che copre tutte le direzioni è confrontabile o persino migliore dei singolo sistemi bilingue. Inoltre, il nostro approccio ottiene risultati competitivi anche per coppie di lingue non viste durante il trainig, facendo uso di traduzioni con pivot.*

1 Introduction

Neural machine translation (NMT) has recently shown its effectiveness by delivering the best performance in various evaluation campaigns (IWSLT 2016 (Cettolo et al., 2016), WMT 2016 (Bojar et al., 2016)). Unlike rule-based or phrase-based MT, the end-to-end learning approach of NMT models the mapping from source to target language directly through a posterior probability. The basic component of an NMT system include an encoder, a decoder and an attention mechanism (Bahdanau et al., 2014). Despite the continuous improvement in performance and translation quality, NMT models are highly dependent on the availability of large parallel data, which in practice can only be acquired for a very limited number of language pairs. For this reason, building effective NMT systems for low-resourced languages becomes a primary challenge (Koehn and Knowles, 2017). Recently, (Zoph et al., 2016) showed how a standard string-to-tree statistical MT system (Galley et al., 2006) can effectively outperform NMT methods for low-resource languages, such as Hausa, Uzbek, and Urdu. In this work, we focus on a so-called multilingual NMT (Johnson et al., 2016; Ha et al., 2016), which considers the use of NMT to target many-to-many language translation. Our motivation is that intensive cross-lingual transfer (Terence, 1989) via parameter sharing should ideally help in the case of similar languages and sparse training data. Hence, in this work we investigate multilingual NMT across Italian, Romanian, and English, and simulate low-resource conditions by limiting the amount of parallel data.

Our approach showed a BLEU increase in various language directions, in a low-resource setting. To compare a single language pair NMT models with a single multilingual NMT (M-NMT) model, we considered six translation directions (i.e English↔Italian, English↔Romanian, and Italian↔Romanian). For evaluating the zero-shot translation (i.e. a translation between language pair with no available parallel corpus), we removed the (Italian↔Romanian) language pairs. In the same way as the six-language-pairs, the performance of the four-language-pairs M-NMT model is comparable with the bilingual models for the language directions with parallel data.

We start in Section 2 with a brief description of NMT and state-of-the-art multilingual NMT approaches. In Section 3, we give a background on our M-NMT model. In Section 4, we present the experimental setting and the NMT model configurations. In Section 5, we show and discuss the results of the experiments. Finally, in Section 6 we present our conclusion and future works.

2 State of The Art

An NMT system consists of three different models called encoder, decoder and attention (Bahdanau et al., 2014). The encoder takes as an input a sequence of words $\mathbf{f} = f_1, \ldots, f_m$ in the form of vocabulary indexes, extract their embeddings and computes a contextual representation of the source words using an RNN implemented with an LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014):

$$\mathbf{h}_t = g(\mathbf{x}_t, \mathbf{h}_{t-1}) \quad t = 1, ..., m$$

where $\mathbf{x}_t$ is the embedding for the word at time step t and m is the length of the source sentence. The decoder receives as input the embedding of the target word at the previous decoding time step, and computes through a RNN a new representation of the current translation, given the representation in the previous step, and a relevant source context computed by the attention model. At each time step, the attention computes normalized weights for the source word positions according to the hidden state of the decoder, which are then used to compute the source context as a weighted sum of all the encoder hidden states. There are several strategies to implement a decoder but all of them end up computing the conditional probability of the next target word depending on the previously translated words and the source sentence:

$$p(e_i = k | e_{<i}, \mathbf{f})$$

The network is trained end-to-end to find the parameters $\hat{\mathbf{\Theta}}$ that maximizes the log-likelihood of the training set $\{(\mathbf{f}_s, \mathbf{e}_s) : s = 1, \ldots, S\}$:

$$\sum_{s=1}^{S} \log p(\mathbf{e}_s | \mathbf{f}_s; \Theta)$$

Based on the end-to-end training approach in NMT, M-NMT models translation across multiple languages with a single model. As such, a multilingual translation task can be categorized into many-to-one, one-to-many, or many-to-many directions, with increasing difficulty. By employing one of these scenarios, recent works in multilingual NMT have shown the possibility of translating language pairs never seen at training time, in addition to improving baseline bilingual NMT models (Ha et al., 2016) (Johnson et al., 2016).

The initial approaches to multilingual NMT required modifications on the standard encoder-decoder architecture (Zoph and Knight, 2016; Firat et al., 2016a; Firat et al., 2016b; Dong et al., 2015; Luong et al., 2015; Lee et al., 2016). Recently, state-of-the-art results are achieved by simply decorating the network inputs with special language tags, to direct the model to a preferred target language at inference time. In this work, following (Johnson et al., 2016) we add a language token at the beginning of every source sentence. This token is unique for the target language and it is a way to impose the target language in which to translate (target-forcing).

3 M-NMT for Low-resource Languages

In this work, we show that it is possible to train a single NMT model for the translation task between multiple language pairs in a low-resource setting. In (Ha et al., 2016; Johnson et al., 2016) it has been shown that a multilingual system trained on a large amount of data improves over a baseline bilingual model, and it is also capable of performing zero-shot translation. In this work we focus on M-NMT in a resource-scarce (Koehn and Knowles, 2017) scenario and show how M-NMT is never worse than a bilingual system for each of the language directions used in the training phase. In fact, the multilinguality can be considered as a way to increase the available amount of data for language directions with small datasets. Moreover, only a single system is needed with respect to several bidirectional NMT systems, thus our setting also represents a way for saving training time and compresses the number of required parameters. The target language can be imposed on the network by using the previously described target forcing.

Furthermore, we use our multilingual model to perform zero-shot translation. We hope that by simply applying the target forcing in the zero-shot scenario, the system can generate sentences in the target language. An alternative zero-shot translation in a resource-scarce scenario can also be performed using a pivot language that is, using an intermediate language for translation. While this is a known technique in machine translation using two or more bilingual models, we expect to achieve a comparable pivoting results using a single multilingual model.

4 Experimental setting

Our NMT model uses embeddings with dimension 1024 and RNN layers based on GRUs of the same dimension. The optimization algorithm is Adagrad (Duchi et al., 2011) with an initial learning rate of 0.01 and mini-batches of size 100. Dropouts are used on every layer, with probability 0.2 on the embeddings and the hidden layers and 0.1 on the input and output layers. All experiments are done using the NMT toolkit Nematus[1] (Sennrich et al., 2017).

Pair	Train	Dev10	Test10	Test17
En-It	231619	1643	929	1147
En-Ro	220538	1678	929	1129
It-Ro	217551	1643	914	1127

Table 1: A total number of parallel sentences used for training and evaluation in a limited low-resource scenario.

For the training set, we used the dataset provided by the latest IWSLT2017[2] multilingual shared task for all possible language pair combinations between Italian, Romanian and English (Cettolo et al., 2012). At the preprocessing stage, we applied word segmentation by jointly learning the Byte-Pair Encoding (Sennrich et al., 2015), merging rules set to 39,500. There is a high overlap between the language pairs (i.e the English dataset paired with Romanian is highly similar to the English paired with Italian). Because of this overlapping, the actual unique sentences in the dataset are approximately the half of the total size. This consequently exacerbates the low-resource aspect in the multilingual models. The size of the vocabulary both in case of the bilingual and the multilingual models stays just under 40,000 sub-words. An evaluation script to determine the BLEU (Papineni et al., 2002) score is used to validate on the dev set and later to choose the best performing models.

We trained models for two different scenarios, the first is the multilingual scenario containing all the available language pairs, while the second scenario is the zero-shot using pivoting, which does not contain parallel sentences for the Romanian↔Italian language pairs. For development and evaluating the models, we used sets from the IWSLT 2010 (Paul et al., 2010) and IWSLT2017 evaluation campaign. The inference is performed using beam search of size 12.

5 Results

5.1 Bilingual Vs. Multilingual

In the first scenario, we compare the translation performance of independently trained bilingual

[1] Nematus- https://github.com/EdinburghNLP/nematus

[2] The International Workshop on Spoken Language Translation - http://workshop2017.iwslt.org/

models against the M-NMT model. In total there are six bilingual models, whereas the M-NMT is trained using the concatenation of all the six languages pair dataset, by just appending an artificial token on the source side. As shown in Table 2, the performance of our systems are evaluated on dev2010 and test2017.

Our preliminary experiments show that the M-NMT system favorably compares with the bilingual systems. Improvements are observed in several language directions, which are likely gained from the cross-lingual parameter transfer between the additional language pairs involved in the source and target side.

Direction	**NMT**	**M-NMT**
English→Italian	26.79	26.34
Italian→English	31.43	31.39
English→Romanian	21.55	**22.13**
Romanian→English	33.84	34.16
Italian→Romanian	15.60	15.92
Romanian→Italian	21.00	**21.60**

Table 2: Comparison between six bilingual models (NMT) against a single multilingual (M-NMT) model. A difference of ≥ 0.5 BLEU score is highlighted as bold.

Specifically, the M-NMT showed an improvement of $+0.58$ and $+0.60$ for En→Ro and It→Ro directions, while having only a small decrease in performance for the En→It and It→En directions (see Table 2).

Direction	**NMT**	**M-NMT**
English→Italian	27.44	**28.22**
Italian→English	29.9	**31.84**
English→Romanian	20.96	**21.56**
Romanian→English	25.44	**27.24**
Italian→Romanian	17.7	**18.95**
Romanian→Italian	19.99	**20.72**

Table 3: Comparison between six bilingual models (NMT) against a single multilingual (M-NMT) model on test2017.

For the evaluation using test2017, however, the M-NMT performed better in all directions than the NMT models (see Table 3). These results show that the M-NMT model performs either in a comparable way or outperforms the single language pair models in this resource-scarce scenario. Moreover, the simplicity of using a single model instead of six leaves a room for further improvements by incorporating more language pairs.

5.2 Pivoting using a Multilingual Model

The pivoting experiment is setup by dropping the Italian-Romanian language pairs from the six directions M-NMT model, which gives us a four directions multilingual model (we call it, PM-NMT), where all the configurations stays the same as in M-NMT. Our main aim is to analyze how a multilingual model can improve a zero-shot translation tasks using a pivoting mechanism, using English as a bridge language in the experiment. Moreover, the use of a multilingual model for pivoting is motivated by the results we acquired using the M-NMT.

Direction	**P-NMT**	**PM-NMT**	**Δ BLEU**
It→Ro	14.14	14.75	$+0.61$
Ro→It	20.16	19.72	-0.44

Table 4: Comparison of pivoting with two bilingual models (P-NMT) against pivoting one multilingual model (PM-NMT). Both approaches use English as the pivoting language. Italian-Romania data was excluded from the training data of the multi-lingual model.

The results in Table 4, show the potential, although partial, of using multilingual models with pivoting for unseen translation directions. The comparable results achieved in both directions speak to us in favor of training and deploying one M-NMT system instead of two distinct NMT.

Direction	**P-NMT**	**PM-NMT**	**Δ BLEU**
It→Ro	16.3	17.58	$+1.28$
Ro→It	18.69	18.66	-0.03

Table 5: Comparison of pivoting with two bilingual models (P-NMT) against pivoting one multilingual model (PM-NMT) using test2017 as the evaluation set.

From the evaluation results on *test2017*, we confirmed that M-NMT can achieve a comparable (Ro→It) or better (It→Ro) result over the two NMT systems used for pivoting. In future work, we will investigate if better performance in pivoting can be achieved by increasing the number of languages covered by the M-NMT system (possibly related to the source and target languages), and/or by different choices of the bridging language.

6 Conclusions

In this paper, we used a multilingual NMT model in a low-resource language pairs scenario. We showed that a single multilingual system achieves comparable performances with the bilingual baselines while avoiding the need to train several single language pair models. Then, we showed how a multilingual model can be used for zero-shot translation by using a pivot language for achieving slightly lower results than a bilingual model trained on that language pair. As a future work we want to explore how the choice of different languages can enable a better parameter transfer in a single model, using more linguistic features of the surface word form, and how to achieve a direct zero-shot translation in a low-resource scenario without the pivoting mechanism.

Acknowledgments

This work has been partially supported by the EC-funded projects ModernMT (H2020 grant agreement no. 645487) and QT21 (H2020 grant agreement no. 645452). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs used for this research.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation (WMT)*, volume 2, pages 131–198.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The iwslt 2016 evaluation campaign. *Proc. of IWSLT, Seattle, pp. 14, WA, 2016.*

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL (1)*, pages 1723–1732.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multilingual neural machine translation. *arXiv preprint arXiv:1606.04164*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 961–968. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT) 2010*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*.

Odlin. Terence. 1989. Language transfer-cross-linguistic influence in language learning. *Cambridge University Press. Cambridge Books Online.*, page 222, June.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

Stem and fragment priming on verbal forms of Italian

Alessandro Laudanna
University of Salerno
Via Giovanni Paolo II, 132 -
84084 - Fisciano (SA)
alaudanna@unisa.it

Giulia Bracco
University of Salerno
Via Giovanni Paolo II, 132 -
84084 - Fisciano (SA)
gcbracco@unisa.it

Abstract

English. In this paper we investigate the respective roles of orthographic and morphological structure in the processing of Italian verbal forms by using the masked priming paradigm.
According to the morphology-based view, in a priming condition the recognition of an inflected word should be facilitated by the presentation of the stem. The cohort model, instead, postulates that the orthographic material from the word's onset to the uniqueness point should be sufficient for the activation of the morphological family.

Italiano. *In questo lavoro indaghiamo il ruolo della struttura ortografica e della struttura morfologica nella elaborazione delle forme verbali dell'italiano, usando il paradigma del priming mascherato. Secondo i modelli basati sulla morfologia, in una condizione di priming il riconoscimento di una forma flessa dovrebbe essere facilitato dalla preventiva presentazione della radice. I modelli coorte, invece, propongono che il materiale ortografico dall'inizio della parola fino al punto di unicità sia sufficiente per attivare la famiglia morfologica.*

1 Introduction

According to the cohort models of visual word recognition (Johnson and Pugh, 1994), all the sources of information that contribute to the identification of a target word proceed from left to right: a single word of the cohort becomes unique at the uniqueness point, when it remains the only candidate corresponding to the orthographic configuration of the stimulus. At that point the recognition takes place. Cohort models use the neighborhood size and the frequency distribution of neighbors as predictors of the competition between candidates and of the consequent recognition latencies. The same models not always deal with the internal morphological structure of the word (but see Marslen-Wilson, 1987). In Italian, verbal families have orthographically similar members, but these words often become 'unique' only at the end, since information about mood, tense and number is carried by affixes in the final part of the word. The N-count is not the best measure to describe their relatedness. On the other hand, morphological parsing accounts (e.g., Baayen, Dijkstra, and Schreuder, 1997; Burani and Caramazza, 1987; Colé, Beauvillain, and Segui, 1989; Taft, 1979) support the argument that the stem is the key of access to the lexicon. Morphological priming shows that, even though morphologically related pairs share some orthographic material (verbs sharing the stem also share letters in initial position, unless they are prefixed), the role played by the morphological structure is different from the one played by the orthographic structure (Pastizzo and Feldman, 2002). In the following experiments, we used the priming paradigm to preactivate both the morphological and the orthographic keys of access. Many studies employing the priming paradigm used an orthographically similar baseline, while others (e.g., Feldman and Soltano, 1999; Marslen-Wilson, Tyler, Waksler, and Older, 1994) employed orthographically and morphologically dissimilar baselines, and others, finally, (Giraudo and Grainger, 2000; Grainger, Colé, and Segui, 1991) estimated morphological facilitation in comparison with both types of baselines. To our knowledge, no study employed the stem priming in Italian, by using both an orthographic and a morphological baseline. We use the term 'stem' to denote the residual part of the word when all inflectional affixes are removed and that can, or cannot be, complex. In contrast, the root is not analysable (Bauer, 1988). The stem does not end in a vowel, and if presented in isolation, it appears as an incomplete word, an orthographic fragment, even if it carries lexical information. By this token, this feature could also turn out to be a benefit: studies employing the priming paradigm usually have to deal with the objection that the base form is a word. The stem primes provide an abstract lexical information in a non-lexicalized form. In order to ascertain if the initial string of letters up to the uniqueness point is the orthographic access code, in Experiment 1 we used Italian verbs with the uniqueness point occurring before the stem (e.g., 'ABBAN' is a fragment present only in the morphological family of 'abbandonare', *to abandon.* All the words starting with the fragment 'ABBAN' (e.g. 'abbandonai', *I abandoned*, 'abbandonato', *abandoned*) belong to the same morphological family). Then, in Experiment 2, we started from the consideration that, when the initial fragment before the stem is shared by more than one family, only a non homographic stem is the true 'uniqueness point': we used verbal forms with the uniqueness point only at the stem boundary (e.g., the fragment 'DISTRI' is shared by two morphological families: 'distribuire', *to distribute*, and 'districare', *to unravel*, whose stems, 'DISTRIB' and 'DISTRIC', respectively, have no homograph). The masked priming paradigm was used in order to avoid intuitions or response strategies in participants (Forster and Davis, 1984; Forster, Davis, Schoknecht and Carter, 1987): this technique avoids the overt detection of any relation between prime and target. Moreover, it has been argued that lexical decision latencies associated with masked priming also reflect the organization of the lexicon in the mind, rather than representing the mechanisms directly involved during single words access (Baayen, 2014). In Experiment 2 cohorts defined by the fragment and by the stem had different frequency distribution, with fragments matching the initial part of lower or higher frequency morphological families. By this token, on the one hand we expected to detect the morphological vs. orthographic nature of the key(s) of access to the lexicon; on the other hand, we indirectly tested the stem frequency effect.

2 Experiment 1

Italian verbs such as 'abbandonare' (*to abandon*) or 'scivolare' (*to slip*) contain fragments ('ABBAN' and 'SCIVO'), that, although shorter than the respective stems 'abbandon*' and '*scivol', can be considered 'morphological uniqueness points', because they belong to just one morphological family. Those stimuli are relevant to decide whether the stem is the necessary key of access to lexical information, or the fragment is sufficient to contact the lexicon. According to the morphology-based view, in a priming condition the recognition of an inflected word should be strongly facilitated by the presentation of the stem. According to the cohort model, instead, the orthographic material from the word's onset to the uniqueness point should be sufficient for the activation of the morphological family.

2.1 Method

Stimuli We selected 16 inflected forms of verbs with a 'unique' initial fragment (e.g., 'abbandonare', to abandon), which served as targets in three different experimental conditions: A) primed by the stem, (e.g., 'ABBANDON'); B) primed by the initial fragment up to the uniqueness point (e.g., 'ABBAN'); C) preceded by an orthographically unrelated fragment which shared no letter with the prime (e.g., 'COTRU'). Mean values for length were 6.9 letters for Stems and 5.3 for Unrelated Prime and Fragments; prime-target orthographic overlap was 67% in Stem Condition and 55% in Fragment Condition. Target mean frequency was 13; root frequency was 462 and initial stem cohort frequency was 245. Three hundred eighty-four items were included in the list as fillers. One hundred eighty-four were words, (40 adjectives, 106 nouns, 38 inflected verbal forms). Those words, together with those in the experimental list, displayed a distribution similar to the one of written Italian (see CoLFIS, Bertinetto et al., 2005). The filler words were matched with experimental targets for their mean length in letters and for their surface frequency. The list included two-hundred items as pseudoword targets. The whole list was composed of 200 words and 200 pseudoword targets preceded in turn by 100 existing primes and 100 non existing primes.

Participants Fifty-four participants, all students of the University of Salerno, and native speakers of Italian, took part into the experiment. They served for a session lasting about 40 minutes. The whole experiment was arranged in three different sessions and each session contained all the targets in one of the three experimental condition (either preceded by the fragment, or preceded by the stem, or preceded by the unrelated fragment). Each participant was submitted to a single experimental session, for a total of 18 'superparticipants'. Each superparticipant was composed of 3 participants, and constituted one data point in the statistical analyses.

Equipment Response box, connected to an IBM PC running the E-Prime 1.1 software (Version 1.1).

Procedure Participants had to press the button corresponding to their dominant hand for the decision 'word', and another one for the decision 'non word'. When the participants reached the 70 % of correct responses in a practice session, the experiment started. All the stimuli appeared in Courier New font, 18 point size in the centre of the computer screen. The fixation was 51 ms, followed by a 51 ms pause. Primes appeared for 51 ms, followed by a 12 characters backward mask ############ (150 ms). The targets remained on the computer screen for a maximum of 1 second. If the participants did not produce any answer within 1 second, the feedback 'Fuori tempo' (*Out of time*) appeared on the screen. The reaction times (RT) were measured from target's onset to subject's response, and the lack of a response was scored as an error.

2.2 Results and Discussion

In Table 1 the mean reaction times and percentage of errors are shown. Table 2 shows the size of Stem and Fragment Priming effects in response latencies and percentage of errors. For 'size of priming effect' we mean the difference between mean Reaction Times (or number of errors) in Stem Condition (or in Fragment Condition) and mean Reaction Times (or number of errors) in Control Condition (Unrelated Fragment Condition).

Condition	Stem	Unrelated Fragment	Fragment
Reaction Times	626 ms	650 ms	626 ms
Errors	12%	15%	13%

Table 1: Mean correct lexical decision latencies and percentage of errors in each priming condition.

Stem Priming Efffect	- 24 ms (-3%)
Fragment Priming Effect	- 24 ms (-2%)

Table 2: Priming effects in response latencies. In parentheses the effect in percentage of errors.

As shown in Table 1, the conditions 'Fragment' and 'Stem' were faster than 'Unrelated Fragment' (Control) condition and they did not differ from each other despite stems were on average longer and with more letters in common with the target word than fragments. The ANOVA on error data did not reveal any significant result (ANOVA by participants $F_{(2,34)}$=.50; p>.6; ANOVA by item $F_{(2,30)}$=.66; p>.5).

The ANOVA on response latencies showed a main effect of prime type only in the analysis by participants ($F_{(2,34)}$= 7.01; p<.002; ANOVA by item $F_{(2,30)}$=1.80; p>.2). Post-hoc analyses based on the ANOVA by participants showed a significant difference between the conditions 'Fragment' vs. 'Unrelated Fragment' (p<.002) and between the conditions 'Stem' vs. 'Unrelated Fragment' (p<.002), but not between 'Fragment' vs. 'Stem' (p >.9). The results are inconsistent with predictions of morphologically-based view: the orthographic uniqueness point is sufficient to contact lexical information.

3 Experiment 2

In Experiment 2 we selected fragments (e.g., DISTRI) shared by verbal families with different frequencies (e.g. 'districare', *to unravel*, lower frequency, LF, and 'distribuire', *to distribute,* higher frequency, HF). Both the accounts (orthographic and morphological) suggest that the fragment is not sufficient for the activation of the morphological family, while the stem, which is also the uniqueness point, should determine a stronger facilitation. The aim of the Experiment 2 was also to address the stem frequency effect.

3.1 Method

Stimuli We selected 32 inflected forms of verbs in 16 pairs with the same initial fragment (e.g., 'distribuito', and 'districato'). One half of the list was composed of 16 targets belonging to higher frequency morphological families (HF, e.g. distribuire, *distributed*); the other half was composed of 16 targets belonging to lower frequency morphological families (LF, e.g., districato, *unraveled*). Target frequency was 2 for LF words, 15 for HF words, 50 for LF Stems and 216 for HF Stems; initial stem cohort frequency was 216 for HF words and 50 for LF words, root frequency was 676 for HF words and 60 for LF words; prime- target orthographic overlap was 72% in Stem condition and 53% in Fragment condition. The same three different experimental conditions of Experiment 2 were arranged. Three hundred sixty-eight items were included in the list as fillers. One-hundred sixty-eight were words, two-hundred items as pseudoword targets. The whole list was composed of 200 words and 200 pseudowords targets preceded in turn by 100 existing primes and 100 non existing primes.

Participants Fifty-four participants, all students of the University of Salerno, and native speakers of Italian, took part into the Experiment. Each participant was submitted to a single session (like in Experiment 1), for a total of 18 superparticipants. Each superparticipant was composed of 3 participants.

Equipment and procedure They were the same as in Experiment 1.

3.2 Results and Discussion

In Tables 3 and 4 the mean reaction times and percentage of errors are shown. Table 5 shows the size of Stem and Fragment Priming effects in response latencies and percentage of errors.

LF			
Condition	Stem	Unrelated Fragment	Fragment
Reaction Times Errors	652 ms 22%	644 ms 24%	647 ms 28%

Table 3: LF verbal forms: mean correct lexical decision latencies and percentage of errors in each priming condition.

HF			
Condition	Stem	Unrelated Fragment	Fragment
Reaction Times Errors	615 ms 14%	627 ms 9%	621 ms 18%

Table 4: HF verbal forms: mean correct lexical decision latencies and percentage of errors in each priming condition.

	LF	HF
Stem Priming Effect	+ 8 ms (-2%)	- 12 ms (-2%)
Fragment Priming Effect	+ 3 ms (+4%)	- 6 ms (+ 3%)

Table 5: Priming effects in response latencies. In parentheses the effect in percentage of errors.

The ANOVA on error data showed an effect of frequency in analyses on both participants ($F_{(1,16)}$=22.33; p<.0005) and items ($F_{(1,30)}$=3,91; p<.05): LF frequency words elicited higher error rates; we also found an effect of prime type in analyses on both participants ($F_{(2,32)}$=5.00; p<.01) and items ($F_{(2,60)}$=4.42; p<.01), but no interaction (ANOVA by participants $F_{(2,32)}$=1.23; p>.3; ANOVA by item $F_{(2,60)}$=1.29; p>.2). The ANOVA on RT showed a main effect of frequency in analyses on both participants ($F_{(1,17)}$=25.80; p<.0001) and items ($F_{(1,30)}$=4.41; p<.04), no effect of prime type (ANOVA by participants: $F_{(2,34)}$=.07; p>.9, ANOVA by item: $F_{(2,60)}$=.18; p>.8), and no interaction (ANOVA by participants $F_{(2,34)}$=1.07; p>.3; ANOVA by item $F_{(2,60)}$=.15; p>.8). On average, HF targets were recognized better than LF targets (621 ms Vs. 647 ms), with faster latencies and a lower percentage of errors (13% Vs. 24%). The lack of priming effect for the Stem condition as compared with the Unrelated Fragment condition is the most surprising result. Post-hoc correlations were performed using main lexical and orthographic variables as predictors, and size of stem and fragment priming effects for RT and errors as criteria. The correlations on results in Fragment condition showed a significant length effect for the fragment prime on HF words (r=-58, p<.02). More interestingly, correlations in Stem conditions showed that the ratio between the surface frequency and the frequency of the stem in initial position is inversely correlated with the size of stem priming (r= -.36, p<.04). The higher the ratio, the faster the latencies: the "relative frequency" of the form in its cohort determines the direction of the effect.The correlation was reliable on LF words (r=-.60, p<.01), while it was not significant on HF words (r=.31 p>.2). The effect did not occur in the Fragment condition, and this might suggest that the effect occurs at the point where the morphological family is selected: the more frequent the cohort, the stronger the inhibition for a verbal form that has a low surface frequency. No effect of cumulative root frequency occurred when the frequency count was obtained by including words embedding the stem in any position (for instance prefixed words), and this allows us to assume that the effect is orthographic in nature. We conclude that not only the word surface frequency, but also the "relative frequency" of the word with respect to its cohort is responsible for recognition.

4 General Discussion

Results of Experiment 1 show that when orthographic information about initial part of the word is exhaustive, it is as reliable as stem priming, and these results are difficult to reconcile with the morphologically- based view which postulates that the stem is critical for lexical access. In addition, the 'relative frequency' effect (Experiment

2), which arises in Stem Condition, suggests that, during recognition, when a low frequency word shares the stem with higher frequency members, it is disadvantaged. In order to gain lexical access, the word has to sustain a harder competition with other members according to their frequency distribution in the morphological family. This effect has been largely described for orthographic neighborhood (Grainger, O'Regan, Jacobs, and Segui, 1989), but, again, it is difficult to reconcile with the stem frequency effect (Burani, Salmaso, and Caramazza, 1984) which states the opposite, with cumulative frequency of morphologically related words facilitating the recognition of a low frequency word. These results are in line with previous data on Italian, which failed in replicating root frequency effects (Laudanna and Bracco, 2009). Root morpheme frequency effects are crucial in the general issue of whether words are accessed through decomposition rather than as full forms, since it has been widely used as the strongest evidence in favor of the hypothesis of the root morpheme representation.

Nevertheless, the morphological parsing account is not the unique explanation for root frequency effects: also full listing models (see Giraudo and Grainger, 2001) can provide a general outline in which this effect is explained, for instance, in terms of lexical connections. This view has been inspected also in linguistics: Bybee (1995) proposed that the activation level of a word is the result of lexical connections and lexical strength, with the first one determined by the frequency of the lexical item, and lexical connections corresponding to the pattern of weight connections associated with the links between related items. For high frequency words the individual lexical strength is elevated; low frequency words have a weak lexical activation and need the support of the activation of lexical connections. This theory is consistent with the claim that whole word frequency effects arise over a precise threshold (Alegre and Gordon, 1999). If the measure for these connections is the stem frequency, the more frequent the stem, the faster the recognition (Burani et al., 1984, Traficante and Burani, 2003, Colombo and Burani, 2002). Meunier and Segui (1999) found that the relative frequency of family members affects the recognition of auditory stimuli: words with high-frequency suffixed candidates derived from the same stem were recognized more slowly than words with morphological family members of a lower frequency. The effects discussed in this paper are also consistent with previous data on stem priming and the "relative frequency" effect in Italian (Bracco and Laudanna, 2012), suggesting that the relations between words in the paradigm need to be taken into account, even if we maintain that whole word representations are the keys for lexical access.

In summary, the results presented in this paper about the processing of Italian verbal forms suggest that it is performed sequentially and it proceeds from left-to-right. Morphological structure does not play a deterministic role, and recognition is guided by the information carried by the initial part of the word, whether it matches a morpheme or not.

Priming induced by 'unique' fragments is as reliable as stem priming. In addition, stem priming is not explainable in terms of a stem frequency effect.

Furthermore, the observation of a 'relative frequency' rather than a 'stem frequency' effect, suggests that we are tapping into a phenomenon concerning connections among whole words.

References

Alegre, M., and Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory & Language*, 40, 41-61.

Baayen, R.H. (2014) Experimental and psycholinguistic approaches to studying derivation. In: R. Lieber and P. Stekauer (Eds), *Handbook of derivational morphology*. Oxford: Oxford University Press, 95-117.

Baayen, H., Dijkstra, T., and Schreuder, R. (1997). *Singulars and plurals in Dutch: Evidence for a parallel dual-route model*. Journal of Memory and Language, 37, 94-117.

Bauer, L. (1988). *Introducing Linguistic Morphology*. Edinburgh, Edinburgh University Press.

Bertinetto, P.M., Burani C., Laudanna, A., Marconi L., Ratti n , D., Rolando, C. and Thornton A.M. (2005). *Corpus e Lessico di frequenza dell'Italiano Scritto* (CoLFIS) http://alphalinguistica.sns.it/CoLFIS/CoLFIS_Presentazione.htm

Bracco G., and Laudanna, A. (2012). "Meccanismi di competizione tra forme verbali nell'accesso al lessico mentale". In: Pier Marco BERTINETTO, Valentina BAMBINI and Irene RICCI e Collaboratori (a cura di). *Linguaggio e cervello / Semantica, Atti del XLII Convegno della Società di Linguistica Italiana* (Pisa, Scuola Normale Superiore, 25-27 settembre 2008). Roma: Bulzoni. Volume 2.

Burani, C., and Caramazza, A. (1987). *Representation and processing of derived words*. Language and Cognitive Processes, 3, 217-227.

Burani, C., Salmaso, D., and Caramazza, A. (1984). *Morphological structure and lexical access*. Visible Language, 18, 348-358.

Bybee, J. L. (1995). *Regular morphology and the lexicon*. Language and Cognitive Processes, 10, 425-455.

Colé, P., Beauvillain, C. and Segui, J. (1989). *On the representation and processing of prefixed and suffixed derived words: A differential frequency effect*. Journal of Memory and Language, 28, 1-13.

Colombo, L., and Burani, C. (2002). *The influence of age of acquisition, root frequency and context availability in processing nouns and verbs*. Brain and Language, 81, 398-411.

Feldman, L.B., and Soltano, E.G. (1999). *Morphological priming: The role of prime duration, semantic transparency and affix position*. Brain and Language, 60, 33-39.

Forster, K. I., and Davis, C. (1984). *Repetition priming and frequency attenuation in lexical access*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 680–698.

Forster, K. I., Davis, C., Schoknecht, C., and Carter, R. (1987). *Masked priming with graphemically related forms: Repetition or partial activation?*Quarterly Journal of Experimental Psychology, 39, 211–251.

Giraudo, H. and Grainger, J. (2001). Priming complex words: Evidence for supralexical representation of morphology. Psychonomic Bulletin & Review, 8, 127-131.

Giraudo, H., and Grainger, J.(2000). *Effects of prime word frequency and cumulative root frequency in masked morphological priming*. Language and Cognitive Processes 15, 421-444.

Grainger, J., Colé, P., and Segui, J. (1991). *Masked morphological priming in visual word recognition*. Journal of Memory and Language, 30, 370-384.

Grainger, J., O'Regan,K., Jacobs,A., and Segui,J. (1989). *On the role of competing word units in visual word recognition: The neighborhood frequency effect*. Perception and Psychophysics, 45, 189-195.

Johnson, N. F. and Pugh, K. R. (1994). *A cohort model of visual word recognition.* Cognitive Psychology, 26, 240-346.

Laudanna A., and Bracco, G. (2009). "Family Frequency, Stem Frequency and Family Size in Visual Word Recognition of Italian Verbal Forms". In: Proceedings of the 6th International Morphological Processing Conference. Turku, June 14-17, 2009, Turku: Morproc, p. 63-64.

Marlsen-Wilson, W.D., Tyler, L., Waksler, R., and Older, L. (1994). *Morphology and meaning in the English mental lexicon*. Psychological Review, 101, 3-33.

Marslen-Wilson, W.D. (l987). *Functional parallelism in spoken word-recognition*. In U. Frauenfelder and L.K. Tyler, Spoken word recognition. Cambridge. MA, MIT Press, 71-102.

Meunier, F. and Segui, J. (1999) Morphological priming effect: the role of surface frequency. Brain and language, 68, 54-60.

Pastizzo, M. J. and Feldman, L. B. (2002) *Discrepancies between orthographic and unrelated baselines in masked priming undermine a decompositional account of morphological facilitation*. Journal of Experimental Psychology: Learning, Memory and Cognition 28, 244–249.

Taft, M. (1979). *Recognition of affixed words and the word frequency effect*. Memory and Cognition, 7, 263-272.

Traficante, D., and Burani, C. (2003). *Visual processing of Italian verbs and adjectives: The role of the inflectional family size*. In R. H. Baayen, R. Schreuder, Morphological Structure in Language Processing, Berlin, Mouton de Gruyter, 45-64.

Metadata annotation for dramatic texts

Vincenzo Lombardo
CIRMA/Dipartimento di Informatica
Università di Torino
vincenzo.lombardo@unito.it

Rossana Damiano
CIRMA/Dipartimento di Informatica
Università di Torino
rossana.damiano@unito.it

Antonio Pizzo
CIRMA /Dipartimento Studi Umanistici
Università di Torino
antonio.pizzo@unito.it

Abstract

English. This paper addresses the problem of the metadata annotation for dramatic texts. Metadata for drama describe the dramatic qualities of a text, connecting them with the linguistic expressions. Relying on an ontological representation of the dramatic qualities, the paper presents a proposal for the creation of a corpus of annotated dramatic texts.

Italiano. *Questo articolo affronta il problema dell'annotazione di metadati per i testi drammatici. I metadati per il dramma descrivono le qualità drammatiche di un testo, connettendole alle espressioni linguistiche. Basandosi su una rappresentazione ontologica delle qualità drammatiche, l'articolo presenta una proposta per la creazione di un corpus di testi drammatici annotati.*

1 Introduction

Drama annotation is the process of annotating the metadata of a drama. Given a drama expressed in some medium (text for screenplays, audiovisual for cinema, interactive multimedia for videogames, etc., termed by Esslin "dramatic media", i.e. media that display characters performing actions): the process of metadata annotation identifies what are the elements that characterize the drama and annotates such elements in some metadata format. For example, in the sentence "Laertes and Polonius warn Ophelia to stay away from Hamlet.", the word "Laertes", which refers to a drama element, namely a character, will be annotated as "Character", taken from some set of metadata. Drama annotation projects, with the sets of metadata and annotations proposed in the scientific literature, rely upon markup languages and semantic encoding.

Recently, there have been many approaches to the annotation of stories (a larger set than drama, including general narrative, not exclusively conveyed by characters performing actions). Annotations are going to enrich drama documents with appropriate metadata. Most of the approaches, e.g., the Story Workbench tool (Finlayson, 2011) and the DramaBank project (Elson, 2012), build upon the linguistic expression of the story, typically some natural language, and annotate story elements, such as characters and conflicts, over the linguistic layer of part-of-speech tagging and verbal frames. Other approaches are more detached from the linguistic expression: they consider the cultural object of the story and rely on conceptual models encoded in logic frameworks, e.g., the Contextus Project[1], the StorySpace ontology (Wolff et al., 2012).

However, most projects work in an isolated fashion: each approach provides its own annotation schema, without connection with the general knowledge, and do not provide the annotated documents with a clear status. This paper presents an overview of the Drammar approach for the metadata annotation of dramatic texts: the gathering of such corpus is relevant for teaching drama through schematic charts (Lombardo et al., 2016b), informing models of automatic storytelling (Lombardo et al., 2015), preserving drama as an intangible form of cultural heritage (Lombardo et al., 2016a). We shortly review the current approaches, before introducing the Drammar ontology under-

[1] `http://www.contextus.net`, visited on 7 July 2017.

lying the annotation schema. Then we describe the crowdsourcing initiative POP-ODE and the current development of the annotated corpus. Finally, we briefly discuss the status of the annotated document, before the conclusion.

2 Drama and annotation

A drama is a story conveyed through characters who perform live actions: for example, theatrical plays (Shakespeare's *Hamlet*), TV series (HBO's *Sopranos*[2]), but even reality shows (CBS's *Survivor*[3]), and games (Ubisoft's *Assassin's Creed* [4]). Metadata annotation for dramatic texts must encode the major concepts and relations of the drama domain, which have been shared by a majority of scholars in the drama literature. Here, we refer to the so–called *dramatic qualities*, that is those elements that are necessary for the existence of a drama, which can be found in several drama analyses, e.g. (Lavandier, 1994; Ryngaert, 2008; Hatcher, 1996; Spencer, 2002). All the initiatives on this topic have shared similar sets of elements, namely story units, characters or agents, actions, intentions or plans, goals, conflicts, values at stake, emotions. These elements are annotated in connection with media chunks (e.g., text paragraphs), often with the goal of constructing corpora of annotated narratives and the study of the relationships between the linguistic expression of the story in the narrative and its content.

Project DramaBank, which has proposed a template based language for describing the narrative content of text documents, is a standalone downloadable application relying on an internal, non-standardized representation format (Elson, 2012). A media-independent model of story is provided by the OntoMedia ontology, exploited across different projects (such as the Contextus Project[5]) to annotate the narrative content of different media objects, ranging from written literature to comics and TV fiction. In the field of cultural heritage dissemination, the StorySpace ontology supports museum curators in linking the content of artworks through stories (Wolff et al., 2012), with the ultimate goal of enabling the generation of user tailored content retrieval. Some initiatives also rely on automatic annotation approaches, which can overcome the difficulties of recruiting annotators, especially when minimal schemata targeted at grasping the regularities of written and oral narratives at the discourse level can be worked out (Rahimtoroghi et al., 2014).

Here, we provide an overview of the Drammar approach[6], an ontology of drama, specifically conceived to annotate dramatic media (Lombardo and Pizzo, 2014), that makes the knowledge about drama available as a vocabulary for the linked interchange of annotations and readily usable by automatic reasoners for implementing many tasks (such as, e.g., the calculation of characters' emotions (Lombardo et al., 2015)).

However, though convenient for its formal account amenable to automatic reasoning, the use of ontology editors and reasoning tools is challenging for drama experts (Varela, 2016). For the accomplishment of the annotation task, it is crucial to provide a friendly environment with metaphors and interfaces that directly descend from the drama scholarship, which abstracts the annotator from the details of the ontology representation. Here we describe a pipeline and system for the metadata annotation of dramatic texts.

3 The Drammar ontology

In order to build a formal encoding of the dramatic elements, Drammar resorts to a set of theories and models that are well established in Artificial Intelligence and Computer Science. Fig. 1 provides an overview of the major classes and properties of the ontology: on the left side, the timeline of incidents grouped into units (upper part, left), connected with the agents' intentions (or plans, lower part, left) through the concept of Action (middle part, left); on the right side, the hierarchical scene structure (upper part, right), connected to the patterns for describing actions (lower part, right), which assign roles to agents; the middle of the figure describes the agent, with its conflicts (lower part, middle), and mental states (middle). Elements in grey levels are referred on external references: *List* and *Treenode*, on top, from abstract data structures; *SituationSchema*, *FramenetSchemata*, and *DescriptionTemplate*, on the left, from linguistic resources; *Agent* and *Object* from general upper

[2] `http://www.hbo.com/the-sopranos`, visited on 21 July 2017

[3] `http://www.cbs.com/shows/survivor/`, visited on 21 July 2017

[4] `https://www.ubisoft.com/en-US/game/assassins-creed/`, visited on 21 July 2017

[5] `http://www.contextus.net`, visited on 21 July 2017

[6] `https://www.di.unito.it/wikidrammar`, visited on 15 October 2017

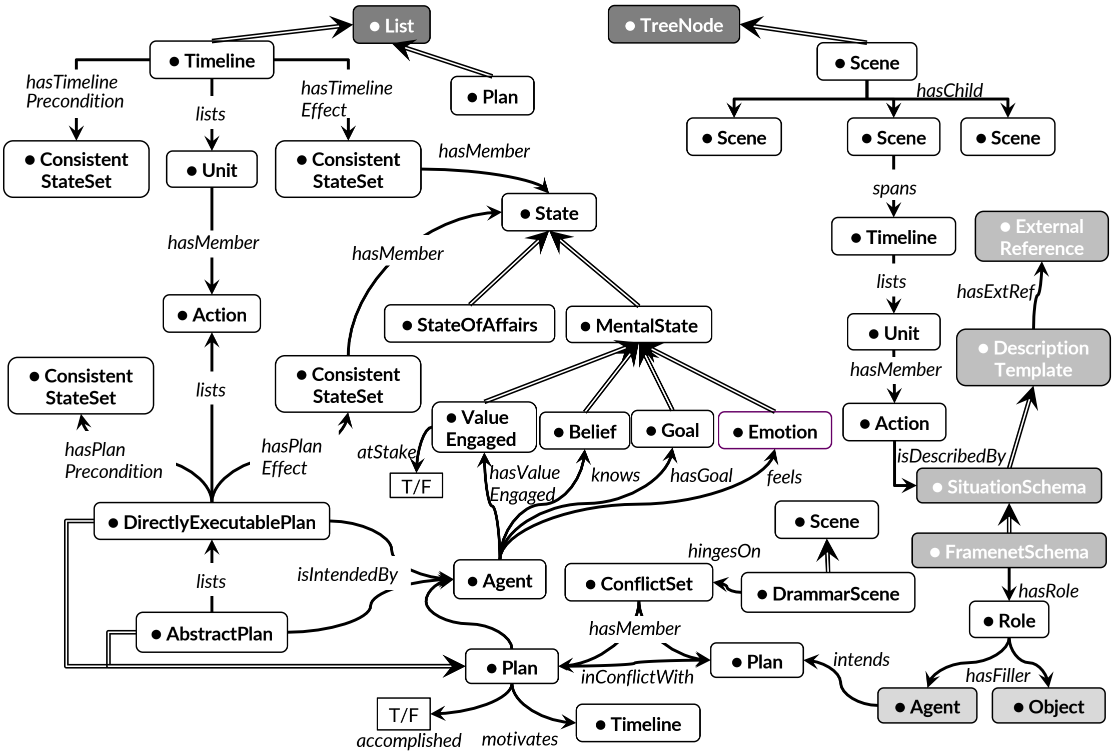


Figure 1: Major classes and properties of ontology Drammar

ontology.

The *Timeline* is the closest element to the drama document (a literary text or an audio-visual medium), a succession of the incidents (or *Actions*) that happen in the drama. Incidents are assembled into discrete structures, called *Units*. Each succession of incidents forms a sub-timeline of the whole timeline of the drama. This level is formalized through the Situation Calculus paradigm (McCarthy, 1986): with sub-timelines that function as operators advancing the story world from one state to another (states aggregated in *ConsistentStateSets*), that work as preconditions and effects of some sub-timeline of incidents.

The actions result from the deliberation process of the characters, named *Agents*, which centers upon the notion of the character's intention in achieving (or trying to achieve) a *Goal*. The intention, or the commitment of the character, is represented by a *Plan*, which consists of the actions that are to be carried out in order to achieve some goal; plans are organized hierarchically, with high-level behaviors (*AbstractPlans*) formulated as lists of lower-level plans, or subplans, until the *DirectlyExecutablePlans*, which directly contain actions. Goals originate from the values of the characters that are put at stake and need to be restored (*ValueEngaged*), given the *Beliefs* (i.e. the knowledge) of the agents. This level is formalized through the rational agent paradigm, or BDI (Belief, Desire, Intention) paradigm (Bratman, 1987) (which is also applied in the computational storytelling community (Norling and Sonenberg, 2004) (Peinado et al., 2008). So, an agent is characterized by goals, beliefs, values engaged, emotions, and plans; values can be *atStake* (true) or in balance (*atStake* false); plans can be in conflict with other plans, possibly of other agents; a conflict set aggregates all the plans, agents, and goals that determine a dramatic scene (*DrammarScene*), through the game of alternate accomplishments. A plan *motivates* the existence of a (sub)timeline, has preconditions and effects, which are consistent sets of states, and can be *accomplished* or not. Finally, scenes, defined by the author or perceived by the audience, to appropriately segment the timeline, are recursively composed of daughter scenes. A scene *spans* a timeline, that is a sequence of units. Some scenes are *DrammarScenes*, meaning that they are motivated by some conflict over the characters' intentions, which is the characterization of scenes according to the Drammar ontology.

The concepts and relations of the ontology Drammar are written in the Semantic Web language OWL (Ontology Web Language), in particular, OWL2 RL (Rule Language), a syntactic and semantic restriction of OWL 2. This allows to address the problem of connecting drama knowledge with the general knowledge. In fact, since Drammar includes classes that are intended as an interface between the drama domain concepts and the linguistic and common sense types of knowledge (see the grey boxes in Fig. 1), it is compliant with the paradigm of linked data (Heath and Bizer, 2011).

4 Crowdsourcing annotation of drama texts: the POP-ODE initiative

POP-ODE consists of a pipeline and a number of tools for the accomplishment of the annotation task of metadata for dramatic texts. A web-based interface supports the feeding of the tables of a data base, built according to the tenets of ontology Drammar: story units, characters, actions, intentions or plans, goals, conflicts, values at stake (emotions are calculated automatically from these data). The ontology axioms have been encoded by the drama scholar (supported by the ontology engineers), through the well-known Protègè editor[7]. A module converts the data base tables into an OWL file, actual a Drammar Instantiated Ontology file (OWL DIO file).

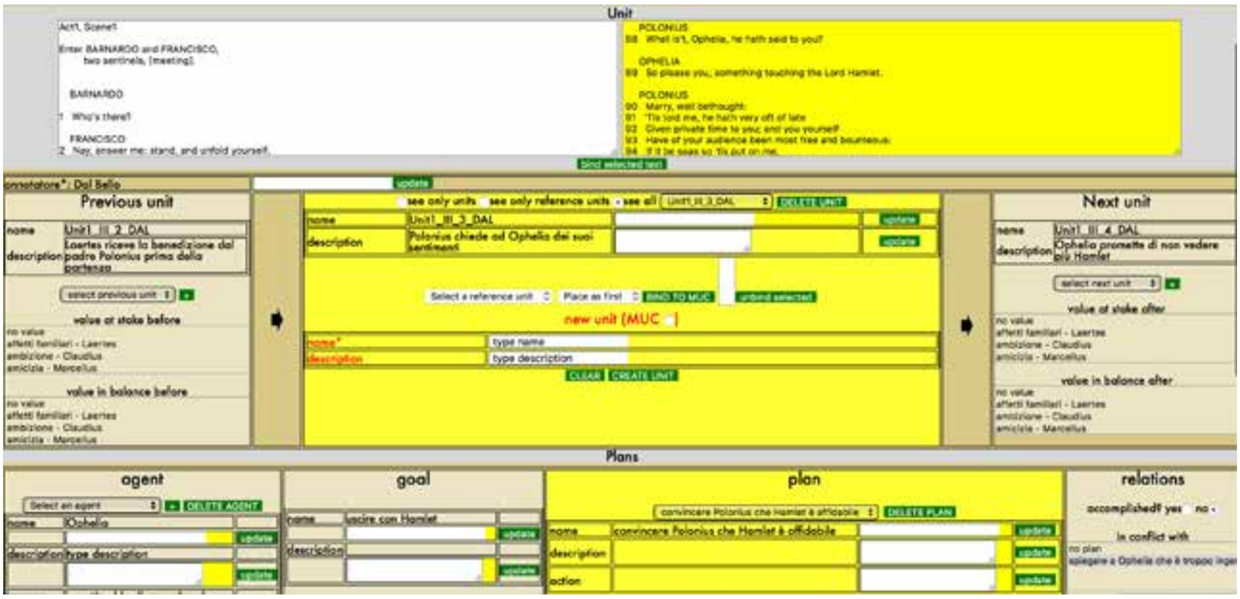
Figure 2 shows the web interface for the annotation. The top of the figure shows the text selector: on the left, the Hamlet text from an authoritative source (Shakespeare's navigators), on the right, the text chunk that pertains to the unit selected below. The middle of the figure shows the unit annotation, that is the actions that have been identified by the annotator in the selected segment of the text, recognized as a bounded unit. On the left and the right of the unit annotation are the previous and the following unit in the story timeline, with the values that are at stake or at balance before and after the current unit. So, in this example, the unit concerns Polonius that asks Ophelia about her feelings; it occurs after Polonius blesses Laertes on his departure and before Ophelia promises to avoid Hamlet. The bottom of the figure concerns the plans that motivate such a unit. In particular, going from left to right, we see that, Ophelia (the agent or character shown at the left), who has the goal of meeting Hamlet, has the plan of convincing her father Polonius that Hamlet is reliable, and this plan is in conflict with Polonius' plan who wants to convince Ophelia that she is too candid for Hamlet. As we know from the following unit, Polonius will succeed in convincing Ophelia, and actually Ophelia's plan fail (see "accomplished? NO" at the far right).

The corpus of annotated drama documents currently consists of a small number of video and textual drama documents, respectively (see table 1). Though we have not carried a thorough evaluation of the annotation, we have employed the annotated documents in two applicative tasks: the first is the calculation of the emotions felt by the characters through automatic reasoning, on the basis of the events and the intentions manually annotated (Lombardo et al., 2015); the second is the realization of printed charts of the characters' intentions, aligned with the timeline of incidents (Lombardo et al., 2016b), currently employed in the didactics of drama writing at the University of Torino. We are going to evaluate the appropriateness of Drammar on the adequacy of description from the point of view of research on the humanities.

The current corpus has been employed in the realization of printed charts of the characters' intentions aligned with the timeline of incidents (Lombardo et al., 2016b), the application of automatic reasoning techniques to compute the emotions felt by the characters on the basis of the events and the intentions manually annotated (Lombardo et al., 2015); the proposal of a model for the preservation of drama as an intangible form of cultural heritage (Lombardo et al., 2016a), the encoding of Stanislavsky's Action Analysis, useful in perspective for supporting actor rehearsals and drama staging (Albert et al., 2016).

Finally, we report a few considerations on the status of a Drammar instantiated file, which contains an annotated drama text, by connecting the Drammar format with the widespread FRBR conceptual model. The FRBR model (Functional Requirements for Bibliographical Entities) (O' Neill, E. T., 2002), designed for capturing the seman-

[7] `http://protege.stanford.edu`, visited on 15 October 2017.



Act1, Scene1

Enter BARNARDO and FRANCISCO,
two sentinels, [meeting].

BARNARDO
1 Who's there?

FRANCISCO
2 Nay, answer me: stand, and unfold yourself.

POLONIUS
88 What is't, Ophelia, he hath said to you?

OPHELIA
89 So please you, something touching the Lord Hamlet.

POLONIUS
90 Marry, well bethought:
91 'Tis told me, he hath very oft of late
92 Given private time to you; and you yourself
93 Have of your audience been most free and bounteous:
94 If it be so as so 'tis put on me,

annotatore*: Dal Bello

Previous unit — name: Unit1_III_2_DAL; description: Laertes riceve la benedizione dal padre Polonius prima della partenza

value at stake before: no value / affetti familiari - Laertes / ambizione - Claudius / amicizia - Marcellus

value in balance before: no value / affetti familiari - Laertes / ambizione - Claudius / amicizia - Marcellus

name: Unit1_III_3_DAL; description: Polonius chiede ad Ophelia dei suoi sentimenti

new unit (MUC ☐) — name*: type name; description: type description

Next unit — name: Unit1_III_4_DAL; description: Ophelia promette di non vedere più Hamlet

value at stake after: no value / affetti familiari - Laertes / ambizione - Claudius / amicizia - Marcellus

value in balance after: no value / affetti familiari - Laertes / ambizione - Claudius / amicizia - Marcellus

agent — name: Ophelia; description: type description

goal — name: uscire con Hamlet

plan — convincere Polonius che Hamlet è affidabile; name: convincere Polonius che Hamlet è affidabile

relations — accomplished? yes ○ no ●

in conflict with: no plan / spiegare a Ophelia che è troppo inge

Figure 2: The web interface of the POP-ODE annotation: top) text selection; middle) unit annotation; bottom) intentions-goals-conflicts annotation.

Medium	Work	Fragment
Text	Hamlet (Shakespeare)	whole text (Arden book)
Text	Mutter Courage und Ihre Kinder (Brecht)	whole text (in Italian - Einaudi)
Text	L'Arialda (Testori's Italian neorealism)	whole text (in Italian - Feltrinelli)
Movie	Apocalypse now	helicopter attack scene (ride of valkyries)
Movie	Taxi driver	"Are you talkin' me?" scene
Movie	Matrix	bullet time scene
Movie	La Dolce Vita	Trevi fountain scene
Movie	The Clockwork Orange	Flat Block Marina scene
Movie	Blade Runner	"I've seen thinks ..." scene
Movie	The deer hunter	Russian roulette scene
Movie	The Godfather	Sollozzo omicide scene
Movie	The Snatch	dog VS. rabbit scene
Movie	Kill Bill - Vol. 2	"losing the other eye" scene
Musical video clip	Taylor Swift's "You belong with me'	3-min video
Advertisement clip	"Zippo" lighter commercial	30-sec video
Animation short	Oktapodi	2:30-min video

Table 1: Corpus of annotated drama documents.

tics of bibliographic information, addresses the abstract ideation (called Work, e.g., Beethoven's idea of the Ninth Symphony), the encoding in a specific language such as the text (called Expression, e.g., Berliner Philarmoniker's interpretation of the Ninth), the concrete representation (called Manifestation, e.g., some Berliner Philarmoniker's recording of the Ninth), and a single instance (called Item, e.g., some published CD of some Berliner Philarmoniker's recording of the Ninth). In our case, the instantiated OWL file is a particular Expression of the underlying drama abstraction (called Work, in FRBR terms), encoded in the ontological format. So, the original textual document is an actual Manifestation of the ontological linguistic Expression that is perfectly compliant with the FRBR model. We can have many manifestations of such a single expression, which however constrains units and timelines to remain unaltered.

5 Conclusion

In this paper, we have described the Drammar approach for the metadata annotation of dramatic texts. We have described the Drammar ontology and the POP-ODE initiative for the annotation pipeline for drama documents, together with the web-based annotation tool. We are going to make a vast and effective test of the annotation tool over several student classes, together with questionnaires and ethnographic observations, to evaluate the functioning of the tool and to create a vast corpus for studies in the digital humanities.

References

Giacomo Albert, Antonio Pizzo, Vincenzo Lombardo, Rossana Damiano, and Carmi Terzulli. 2016. Bringing authoritative models to computational drama (encoding knebel's action analysis). In *Interactive Storytelling. 9th International Conference on Interactive Digital Storytelling, ICIDS 2016*, volume 10045, pages 285–297, Cham – CHE, November 15–18. Springer International Publishing.

Michael E. Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge (MA).

David K. Elson. 2012. Dramabank: Annotating agency in narrative discourse. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Mark Alan Finlayson. 2011. The story workbench: An extensible semi-automatic text annotation tool. In *AAAI Publications, Workshops at the Seventh Artificial Intelligence and Interactive Digital Entertainment Conference*.

Jeffrey Hatcher. 1996. *The Art and Craft of Playwriting*. Story Press, Cincinnati, Ohio.

T. Heath and C. Bizer. 2011. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, pages 1–136.

Yves Lavandier. 1994. *La dramaturgie*. Le clown et l'enfant, Cergy.

Vincenzo Lombardo and Antonio Pizzo. 2014. Multimedia tool suite for the visualization of drama heritage metadata. *Multimedia Tools and Applications*, 75(7):3901–3932.

Vincenzo Lombardo, Cristina Battaglino, Antonio Pizzo, Rossana Damiano, and Antonio Lieto. 2015. Coupling conceptual modeling and rules for the annotation of dramatic media. *Semantic Web Journal*, 6(5):503–534.

Vincenzo Lombardo, Antonio Pizzo, and Rossana Damiano. 2016a. Safeguarding and accessing drama as intangible cultural heritage. *ACM Journal on Computing and Cultural Heritage*, 9(1):1–26.

Vincenzo Lombardo, Antonio Pizzo, Rossana Damiano, Carmi Terzulli, and Giacomo Albert. 2016b. Interactive chart of story characters' intentions. In *Interactive Storytelling, 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings*, volume 10045, pages 415–418, Cham – CHE, November 15–18,. Springer International Publishing.

John C. McCarthy. 1986. Mental situation calculus. In *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning About Knowledge*, TARK '86, pages 307–307, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

E. Norling and L. Sonenberg. 2004. Creating Interactive Characters with BDI Agents. In *Proceedings of the Australian Workshop on Interactive Entertainment IE2004*.

O' Neill, E. T. 2002. Frbr: Functional requirements for bibliographic records; application of the entity-relationship model to humphry clinker. *Library Resources and Technical Services*, 46:150–158.

F. Peinado, M. Cavazza, and D. Pizzi. 2008. Revisiting Character-based Affective Storytelling under a Narrative BDI Framework. In *Proc. of ICIDIS08*, Erfurt, Germany.

Elahe Rahimtoroghi, Thomas Corcoran, Reid Swanson, Marilyn A. Walker, Kenji Sagae, and Andrew Gordon. 2014. Minimal narrative annotation schemes and their applications. In *AAAI Publications, Seventh Intelligent Narrative Technologies Workshop*.

Jean-Pierre Ryngaert. 2008. *Introduction à l'analyse du théâtre*. Collection Cursus. Série Littérature. Armand Colin.

Stuart Spencer. 2002. *The Playwright's Guidebook: An Insightful Primer on the Art of Dramatic Writing*. Faber & Faber.

Miguel Escobar Varela. 2016. Interoperable performance research promises and perils of the semantic web. *The Drama Review*, 60(3).

Annika Wolff, Paul Mulholland, and Trevor Collins. 2012. Storyspace: A story-driven approach for creating museum narratives. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 89–98.

Deep Learning for Automatic Image Captioning in poor Training Conditions

Caterina Masotti and Danilo Croce and Roberto Basili
Department Of Enterprise Engineering
University of Roma, Tor Vergata
caterinamasotti@yahoo.it
{croce,basili}@info.uniroma2.it

Abstract

English. Recent advancements in Deep Learning show that the combination of Convolutional Neural Networks and Recurrent Neural Networks enables the definition of very effective methods for the automatic captioning of images. Unfortunately, this straightforward result requires the existence of large-scale corpora and they are not available for many languages. This paper describes a simple methodology to automatically acquire a large-scale corpus of 600 thousand image/sentences pairs in Italian. At the best of our knowledge, this corpus has been used to train one of the first neural systems for the same language. The experimental evaluation over a subset of validated image/captions pairs suggests that results comparable with the English counterpart can be achieved.

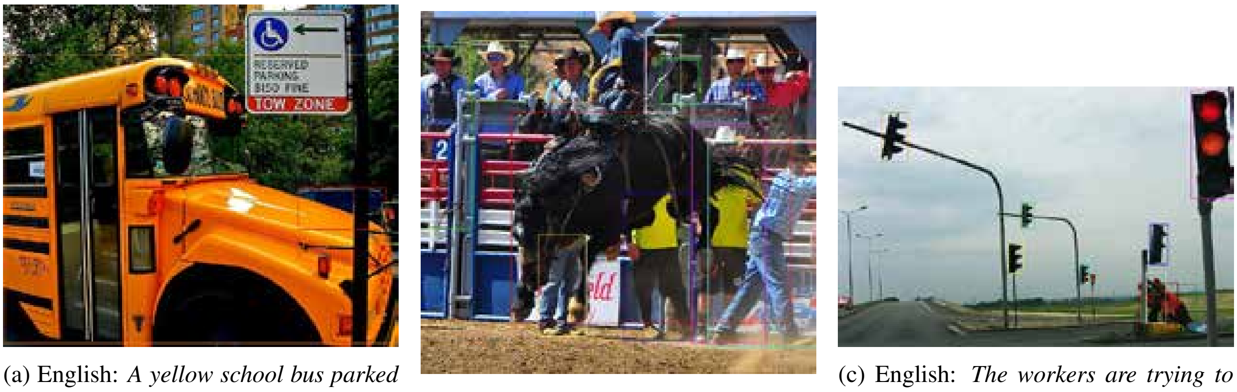
Italiano. *La combinazione di metodi di Deep Learning (come Convolutional Neural Network e Recurrent Neural Network) ha recentemente permesso di realizzare sistemi molto efficaci per la generazione automatica di didascalie a partire da immagini. Purtroppo, l'applicazione di questi metodi richiede l'esistenza di enormi collezioni di immagini annotate e queste risorse non sono disponibili per ogni lingua. Questo articolo presenta un semplice metodo per l'acquisizione automatica di un corpus di 600 mila coppie immagine/frase per l'italiano, che ha permesso di addestrare uno dei primi sistemi neurali per questa lingua. La valutazione su un sottoinsieme del corpus manualmente validato suggerisce che é possibile raggiungere risultati comparabili con i sistemi disponibili per l'inglese.*

1 Introduction

The image captioning task consists in generating a brief description in natural language of a given image that is able to capture the depicted objects and the relations between them, as discussed in (Bernardi et al., 2016). More precisely, given an image I as input, an *image captioner* should be able to generate a well-formed sentence $S(I) = (s_1, ..., s_m)$, where every s_i is a word from a vocabulary $V = \{w_1, ..., w_n\}$ in a given natural language. Some examples of images and corresponding captions are reported in Figure 1. This task is rather complex as it involves non-trivial subtasks to solve, such as object detection, mapping visual features to text and generating text sequences.

Recently, neural methods based on deep neural networks have reached impressive state-of-the-art results in solving this task (Karpathy and Li, 2014; Mao et al., 2014; Xu et al., 2015). One of the most successful architectures implements the so-called *encoder-decoder* end-to-end structure (Goldberg, 2015). Differently by most of the existing encoder-decoder structures, in (Vinyals et al., 2014) the encoding of the input image is performed by a convolutional neural network which transform it in a dense feature vector; then, this vector is "translated" to a descriptive sentence by a Long short-term memory (LSTM) architecture, which takes the vector as the first input and generates a textual sequence starting from it. This neural model is very effective, but also very expensive to train in terms of time and hardware resources[1], because there are many parameters to be learned; not to mention that the model is overfitting-prone, thus it needs to be trained on a training set of annotated images that is as large and heterogeneous

[1] As of now, training a neural encoder-decoder model such as the one presented at http://github.com/tensorflow/models/tree/master/im2txt on a dataset of over $580,000$ image-caption examples takes about two weeks even with a very performing GPU.



(a) English: *A yellow school bus parked in a handicap spot*, Italian: *Uno scuolabus giallo parcheggiato in un posto per disabili*.

(b) English: *A cowboy rides a bucking horse at a rodeo*, Italian: *Un cowboy cavalca un cavallo da corsa a un rodeo*.

(c) English: *The workers are trying to pry up the damaged traffic light*, Italian: *I lavoratori stanno cercando di tirare su il semaforo danneggiato*.

Figure 1: Three images from the MSCOCO dataset, along with two human-validated descriptions.

as possible, in order to achieve a good generalization capability. Hardware and time constraints do not always allow to train a model in an optimal setting, and, for example, cutting down on the dataset size could be necessary: in this case we have *poor training conditions*. Of course, this reduces the model's ability to generalize on new images at captioning time. Another cause of poor training conditions is the lack of a good quality dataset, for example in terms of annotations: the manual captioning of large collections of images requires a lot of effort and, as of now, human-annotated datasets only exist for a restricted set of languages, such as in English. As a consequence, training such a neural model to produce captions in another language (e.g. in Italian) is an interesting problem to explore, but also challenging due to the lack of data resources.

A viable approach is building a resource by *automatically translating the annotations from an existing dataset*: much less expensive than manually annotating images, but of course it leads to a loss of human-like quality in the language model. This approach has been considered in this work to perform one of the first neural-based image captioning in Italian: more precisely, the annotations of the images from the MSCOCO dataset, one of the largest datasets in English of image/caption pairs, have been automatically translated to Italian in order to obtain a first resource for this language: this has been exploited to train a neural captioner and whose quality can be improved over time (e.g., by manually validating the translations). Then, a subset of this Italian dataset has been used as training data for the neural captioning system defined in (Vinyals et al., 2014), while a subset of the test set has been manually validated for evaluation purposes.

In particular, prior to the experimentations in Italian, some early experiments have been performed with the same training data originally annotated in English, to get a reference benchmark about convergence time and evaluation metrics on a dataset of smaller size. These results in English will suggest if the Italian image captioner shows similar performance when trained over a reduced set of examples or the noise induced in the automatic translation process compromises the neural training phase. Moreover, these experiments have also been performed with the introduction of a pre-trained word embedding, (derived using the method presented in (Mikolov et al., 2013)), in order to measure how it affects the quality of the language model learned by the captioner, with respect to a randomly initialized word embedding that is learned together with the other model parameters.

Overall the contributions of this work are threefold: (*i*) the investigation of a simple, automatized way to acquire (possibly noisy) large-scale corpora for the training of neural image captioning methods in poor training conditions; (*ii*) the manual validation of a first set of human-annotated resources in Italian; (*iii*) the implementation of one of the first automatic neural-based Italian image captioners.

In the rest of the paper, the adopted neural architecture is outlined in Section 2. The description of a brand new resource for Italian is presented in Section 3. Section 4 reports the results of the early preparatory experimentations for the English language and then the ones for Italian. Finally, Section 5 derives the conclusions.

2 The Show and Tell Architecture

The Deep Architecture considered in this paper is the *Show and Tell* architecture, described in (Vinyals et al., 2014) and sketched in Figure 2. It follows an encoder-decoder structure where the image is encoded in a dense vector by a state-of-the-art deep CNN, in this case *InceptionV3* (Szegedy et al., 2015), followed by a fully connected layer; the resulting feature vector is fed to a LSTM, used to generate a text sequence, i.e. the caption. As the CNN encoder has been trained over an object recognition task, it allows encoding the image in a dense vector that is strictly connected to the entities observed in the image. At the same time, the LSTM implements a language model, in line with the idea introduced in (Mikolov et al., 2010): it captures the probability of generating a given word in a string, given the words generated so far. In the overall training process, the main objective is to train a LSTM to generate the next word given not only the string produced so far, but also a set of image features. As the first CNN encoder is (mostly) language independent, it can be totally re-used even in the captioning of images in other languages, such as Italian. On the contrary, the language model underlying the LSTM needs new examples to be trained.

In this work, we will train this architecture over a corpus that has been automatically translated from the MSCOCO dataset. We thus speculate that the LSTM will learn a sort of simplified language model, more inherent to the automatic translator than to an Italian speaker. However, we are also convinced that the quality achievable by modern translation systems (Bahdanau et al., 2014; Luong et al., 2015), combined with the generalization that can be obtained by a LSTM trained over thousands of (possibly noisy) translations will be able to generate reasonable and intelligible captions.

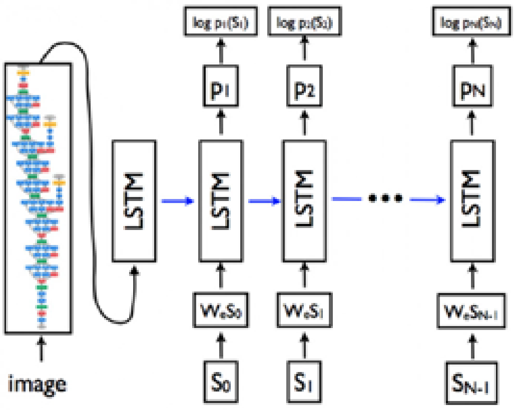


Figure 2: The Deep Architecture presented in (Vinyals et al., 2014). LSTM model combined with a CNN image embedder and word embeddings. The unrolled connections between the LSTM memories are in blue.

3 Automatic acquisition of a Corpus of Captions in Italian

In this section we present the first release of the MSCOCO-it, a new resource for the training of data-driven image captioning systems in Italian. It has been built starting from the MSCOCO dataset for English (Lin et al., 2014): in particular we considered the training and validation subsets, made respectively of $82,783$ and $40,504$ images, where every image has 5 human-written annotations in English. The Italian version of the dataset has been acquired with an approach that automatizes the translation task: for each image, all its five annotations have been translated with Bing[2]. The result is a big amount of data whose annotations are fully translated, but not of the best quality with respect to the Italian fluent language. This automatically translated data can be used to train a model, but for the evaluation a test set of human-validated examples is needed: so, the translations of a subset of the MSCOCO-it have been manually validated. In (Vinyals et al., 2014), two subsets of $2,024$ and $4,051$ images from the MSCOCO validation set have been held out from the rest of the data and have been used for development and testing of the model, respectively. A subset of these images has been manually validated: 308 images from the development set and 596 from the test set. In Table 1, statistics about this brand new corpus are reported, where the specific amount of unvalidated (*u.*) and validated (*v.*) data is made explicit[3].

4 Experimental Evaluation

In order to be consistent with a scenario characterized by *poor training conditions* (limited hardware resources and time constraints) all the experimentations in this paper have been made by training

[2] Sentences have been translated between December 2016 and January 2017.

[3] Although Italian annotations are available for all the images of the original dataset, in the table some images were not counted because they are corrupted and therefore have not been used.

		#images	#sent	#words
training	*u.*	116,195	581,286	6,900,546
valid.	*v.*	308	1,516	17,913
	u.	1,696	8,486	101,448
	p.	(14)	25	304
test	*v.*	596	2,941	34,657
	u.	3,422	17,120	202,533
	p.	(23)	41	479
total		**122,217**	**611,415**	**7,257,880**

Table 1: Statistics about the MSCOCO-it corpus. *p.* stands for *partially validated*, since some images have only some validated captions out of five. The partially validated images are between parentheses because they are already counted in the validated ones.

the model on significantly smaller samples of data with respect to the whole MSCOCO dataset (made of more than $583,000$ image-caption examples).

First of all, some early experimentations have been performed on smaller samples of data from MSCOCO in English, in order to measure the loss of performance caused by the reduced size of the training set[4]. Each training example is a image-caption pair and they have been grouped in data *shards* during the training phase: each shard contains about 2,300 image-caption examples. The model has been trained on datasets of $23,000$, $34,500$ and $46,000$ image-caption pairs (less than 10% of the entire dataset).

In order to balance the reduced size of the training material and provide some kind of linguistic generalization, we evaluated the adoption of pre-trained word embedding in the training/tagging process. In fact, in (Vinyals et al., 2014) the LSTM architecture initializes randomly all vectors representing input words; these are later trained together with the other parameters of the network. We wondered if a word embedding already pre-trained on a large corpus could help the model to generalize better on brand new images at test time. We introduce a word embedding learned through a Skip-gram model (Mikolov et al., 2013) from an English dump of Wikipedia. The LSTM architecture has been trained on the same shards but initializing the word vectors with this pretrained word embedding.

Table 2 reports results on the English dataset in terms of BLEU-4, CIDEr and METEOR, the same used in (Vinyals et al., 2014): in the first five rows, results are reported both in the case of randomly initialized word embedding and pre-trained ones. We compare these results with the ones achieved by the original NIC and NICv2 networks presented in (Vinyals et al., 2014), and the ones measured by testing a model available in the web[5], trained on the original whole training set.

# Shards	BLEU-4	METEOR	CIDEr
1	10,1 / 11,5	13,4 / 13,1	18,8 / 24,4
2	15,7 / 18,9	18,2 / 16,3	36,1 / 51,9
5	22,0 / 22,7	20,2 / 20,4	64,1 / 65,0
10	22,4 / 24,7	22,0 / 21,7	73,2 / 73,7
20	26,5 / 26,2	21,9 / 22,3	79,3 / 79,1
NIC	27,7	23,7	85,5
NICv2	32,1	25,7	99,8
`im2txt`	31,2	25,5	98,1

Table 2: Results on `im2txt` for the English language with a training set of reduced size, without / with and the use of a pre-trained word embedding. Moreover benchmark results are reported.

Results obtained by the network when trained on a reduced dataset are clearly lower w.r.t. the NIC results, but it is straightforward that similar result are obtained, especially considering the reduced size of the training material. The contribution of pre-trained word embeddings is not significant, in line with the findings from (Vinyals et al., 2014). However, it is still interesting noting that the lexical generalization of this unsupervised word embeddings is beneficial, especially when the size of the training material is minimal (e.g. when 1 shard is used, especially if considering the CIDEr metrics). As the amount of training data grows, its impact on the model decreases, until it is not significant anymore.

# Shards	BLEU-4	METEOR	CIDEr
1	11.7 / 12.9	16.4 / 16.9	27.4 / 29.4
2	16.9 / 17.1	18.8 / 18.7	45.7 / 45.6
5	22.0 / 21.4	21.2 / 20.9	62.5 / 60.8
10	22.4 / 22.9	22.0 / 21.5	71.9 / 68.8
20	23.7 / 23.8	22.2 / 22.0	73.0 / 73.2

Table 3: Metrics for the experimentations on `im2txt` for the Italian language with a training set of reduced size, without / with and the use of a pre-trained word embedding.

For what concerns the results on Italian, the experiments have been performed by training the model on samples of $23,000$, $34,500$ and $46,000$ examples, where the captions are automatically

[4] A proper tuning phase was too expensive so we adopted the parameters provided in `https://github.com/tensorflow/models/tree/master/im2txt`

[5] `http://github.com/tensorflow/models/issues/466`

translated with Bing. The model has been evaluated against the validated sentences, and results are reported in Table 3. Results are impressive as they are in line with the English counterpart. It supports the robustness of the adopted architecture, as it seems to learn even from a noisy dataset of automatically translated material. Most importantly, it confirms the applicability of the proposed simple methodology for the acquisition of datasets for image captioning.

When trained with 20 shards, the Italian captioner generates the following description of the images shown in Figure 1: Image 1a: "*Un autobus a due piani guida lungo una strada.*", Image 1b: "*Un uomo che cavalca una carrozza trainata da cavalli.*", Image 1c: "*Una persona che cammina lungo una strada con un segnale di stop.*"

An attempt to use a word embedding that has been pre-trained on a large corpus (more precisely, on a dump of Wikipedia in Italian) has also been made, but the empirical results reported in Table 3 show that its contribution is not relevant but still significant when fewer examples are adopted.

5 Conclusions

In this paper a simple methodology for the training of neural models for the automatic captioning of images is presented. We generated a large scale of about $600,000$ image captions in Italian by using an automatic machine translator. Although the noise introduced in this step, it allows to train one of the first neural-based image captioning systems for Italian. Most importantly, the quality of this system seems comparable with the English counterpart, if trained over a comparable set of data. These results are impressive and confirm the robustness of the adopted Neural Architecture. We believe that the obtained resource paves the way to the definition and evaluation of Neural Models for Image captioning in Italian, and we hope to contribute to the Italian Community, hopefully using the validated dataset in a future Evalita[6] champaign.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442, January.

Yoav Goldberg. 2015. A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726.

Andrej Karpathy and Fei-Fei Li. 2014. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

[6] `http://www.evalita.it/`

E pluribus unum.
Representing compounding in a derivational lexicon of Latin

Silvia Micheli
Università degli Studi di Pavia
Corso Strada Nuova, 75
27100 Pavia
silvia.micheli@unibg.it

Eleonora Litta
Università Cattolica del Sacro Cuore
Largo Gemelli, 1
20123 Milano
e.littamodignani@gmail.com

Abstract

English. This paper describes how compounding is treated in the *Word Formation Latin* derivational lexicon. Through the analysis of some types of Latin compounds, perspectives and limitations of the resource are highlighted; its contribution to theoretical and computational linguistic issues is also outlined.

Italiano. *Questo contributo descrive come viene trattata la composizione nel lessico derivazionale* Word Formation Latin. *Attraverso l'analisi di alcuni aspetti della composizione latina, vengono messi in luce potenzialità e limiti della risorsa e delineato il suo contributo in campo teorico e computazionale.*

1 Introduction: the *Word Formation Latin* lexicon

Word Formation Latin (WFL, (Litta et al., 2016)) is a derivational morphology resource for Latin where words are analysed in their formative components and related to each other on the basis of word formation rules (WFRs).[1] It represents a wide lexical resource not only for the study of Latin derivational morphology (i.e. affixal and conversive processes), but also for compounding, which has often been neglected in other most recent resources for other languages.[2] The lexical basis behind WFL is the same as the morphological analyser and lemmatiser for Latin Lemlat (Passarotti et al., 2017). All lemmas have been collected from three main Classical Latin dictionaries ((Georges and Georges, 1913-1918); (Glare, 1982); (Gradenwitz, 1904)) plus the Onomasticon of Forcellini's (Forcellini, 1940) 5th edition of *Lexicon Totius Latinitatis* (Budassi and Passarotti, 2016). All those lemmas that share a common (not derived) ancestor belong to the same "morphological family", (Litta et al., 2016) represented in the web application (http://wfl.marginalia.it/) as a tree-graph.

The aim of this paper is twofold: on the one hand, it describes how compounding is represented into the WFL derivational lexicon; on the other hand, it aims at highlighting the theoretical and computational contribution of this resource through the analysis of some aspects (i.e. WFRs, input and output lexical categories) of the Latin compounds collected in it.

2 Latin compounding

Compared to other Indo-European languages (e.g. Sanskrit or Greek), compounding in Latin is generally considered to be not very productive. According to (Grenier, 1912) and (Puccioni, 1944), most of Latin compounds are *hapax legomena* and mainly occur in poetic, religious and legal texts. Furthermore, they seem to be strongly influenced by Greek models.

In the last decades, Latin compounding (henceforth LC) has received more attention ((Oniga, 1992); (Oniga, 1988); (Benedetti, 1988); (Fruyt, 2002); (Brucale, 2012)). However, most of the available studies are qualitative descriptions of compounding mechanism, which are based on a small amount of data, usually extracted from dictionaries, and cited as examples of the main types of compounds. These studies have mainly focussed on formal features of LC, which is

[1] The WFL project, still ongoing, received funding from EU Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Individual Fellowship grant agreement No 658332-WFL

[2] Among them, notable ones are the lexical network for Czech DeriNet (Ševčíková and Žabokrtský, 2014) and (Žabokrtský and al., 2016), the derivational lexicon for German DErivBASE (Zeller et al., 2013) and that for Italian derIvaTario (Talamo et al., 2016).

essentially stem-based: Latin compounds are almost always made up of bound units (i.e. roots, stems) connected by a linking element (LE) -*i*-, as in (1).

(1) *purifico*V
pur-i-fic-o
purus+LE+*facio*+INFL
A+V+INFL=V

The nature of the linking element -*i*-,[3] the relationship between compounding and derivation in Latin, and the classification of Latin compounds, are the main theoretical topics on which attention is focused. However, there are still many questions that so far could not be answered exhaustively due to the scarcity of data collected so far: which were the most productive types of compound in Latin? Through which rules were Latin compounds formed? What PoS did Latin compounds consist of most frequently? What kinds of meaning are expressed by compounding in Latin? WFL allows to fill to answer these questions by providing a large account of quantitative data which can help to better understand the mechanisms of LC.

3 Compounding in WFL

The methodology behind WFL is consistent with the Item-and-Arrangement model outlined in (Hockett, 1954), which considers morphemes, not words, the basic units for the study of utterances, containing both form and meaning. The resource relies on a fairly strict morphotactic approach, where, to the basic component of the uninflected word, the so-called les ("LExical Segment"), one derivational morpheme (prefix/suffix) or phenomenon (conversive PoS change) is attached at a time. This means that the output of a WFR is always a lemma richer (containing more morphemes, or different inflection) than the input one.

During the compilation of WFL, an initial list of possible compounds has been drawn by taking into account all possible combinations of V (verb), N (nouns), A (adjectives), PR (pronouns), and I (invariables - e.g. adverbs). Some categories have been filled semi-automatically with the help of SQL queries. These usually matched a string that combines a certain lexical element + -i- + another lexical element or lemma (this one sometimes in the form of a customised string). This method was applicable to morphotactically transparent compounds like those verbs including -*fico* (from verb *facio* 'to make', e.g. *clarifico* 'to make illustrious'), or those adjectives featuring noun *pes* 'foot' as a second constituent (e.g. celer-i-pes, lit. 'fast foot'). However, morphotactically obscure compounds like *fidicina* 'lyre player' (fides 'lyre' + cano 'to sing'), needed to be inserted manually.

The WFL web application allows compounds to be browsed in three ways:

1. By WFR - opens research questions on a specific word formation behaviour; for example, it is possible to view and download a list of all adjectives formed by a A+V=A rule.
2. By PoS - useful for studies on macro-categories, it allows for deeper refinement of constituent PoS.
3. By Lemma - allows for quick search of a specific lemma.

For each compound, a derivational tree-graph is provided (as in Figure 1). In each graph, nodes are lemmas, and edges are relations showing the kind of WFR involved. Special provisions are made in order to collapse and hide compounding relations according to the user's choice. This is useful when very productive constituents are displayed in massive multi-tree graphs.

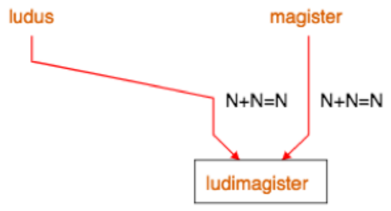


Figure 1: Derivation graph of *ludimagister*

The sample collected from the WFL lexical basis consists of 1744 compounds. The fact that all compounds collected from the three dictionaries mentioned above are for the first time categorised and labelled into a language resource allows for a more in-depth overview and for a quantitative analysis on many aspects of LC (e.g. productivity, WFRs, lexical categories involved in compounding). In the following sections, some preliminary

[3] A survey of the literature on the nature of this linking element is in (Brucale, 2012).

considerations on the data currently included in WFL are provided.

3.1 Word Formation Rules

Compound words collected in WFL are created through 59 WFRs. In table 1, the first twenty most productive WFRs are shown.[4]

	WFRs	Compounds
1	N+V=A	429
2	N+V=N	239
3	N+N=N	135
4	A+V=A	134
5	A+N=A	131
6	N+N=A	120
7	V+V=V	64
8	A+V=V	59
9	N+V=V	56
10	A+N=N	35
11	V+V=A	33
12	A+V=N	32
13	V+N=A	28
14	I+I=I	27
15	A+A=A	22
16	PR+PR=PR	15
17	I+N=N	15
18	N+A=A	14
19	N+A=N	13
20	PR+V=PR	13

Table 1: Compounding WFRs in WFL

The most productive pattern in LC is Noun+Verb: the rule creates both adjectives and nouns, e.g. *soporifer* 'soporific' (*sopor+fero*) or *artifex* 'artisan'(*ars+facio*). This word formation process is no longer productive in Romance Languages, in which the reverse order (i.e. the Verb+Noun pattern, e.g. Italian *portafoglio* 'wallet' or French *porte-parole* 'spokesman') is the most frequent.

In almost all cases, Latin compounds are made up of two constituents. There are only very few (and not productive) cases in which there are three elements, e.g. *turpilucricupidus* (turpis 'vile' + lucrum 'gain' + cupidus 'desirous'; WFR: A+N+N=N) or *suovetaurilia* (sus 'pig' + ovis 'sheep' + taurus 'bull'; WFR: N+N+N=N).

The V+V pattern, that in Italian creates nouns (e.g. *dormiveglia* 'half-sleep', lit. 'to sleep-to stay awake'), in Latin forms mainly new verbs, such as *patefacio* 'to reveal' (*pateo* 'to be evident' + *facio* 'to do').

In addiction to other patterns already identified as productive in previous literature (i.e. A+N=A, N+N=N, N+N=A), it is interesting to notice the presence of a significant number of compounds consisting of two invariable forms (e.g. *etiamtum*, *etiam+tum* 'even then, yet') or two pronouns (e.g. *aliquis*, *alis+quis* 'anyone, someone') which are generally neglected in studies on Latin word-formation.

3.2 Input and output lexical categories

As already pointed out by (Brucale, 2012), verbs and nouns are the most frequent input elements in Latin compounds. While nouns can be found both in first and in second constituent, verbs show a clearer tendency to appear in second position. Data collected in WFL confirms these observations.[5]

Lexical cat.	1° const.	2° const.	Output
A	428	69	942
I	96	55	63
N	1008	491	491
PR	63	32	53
V	141	1089	187

Table 2: Input and output lexical categories in WFL compounds

Table 2 shows the quantitative distribution of the lexical categories (i.e. how many times adjectives are present as the input or as the output PoS) in WFL compounds. More than half of the sample (i.e. 1089 forms, 62.7%) has a verbal second element (e.g. compounds with *-facio* or a related stem, such as *aedifico* 'to build' or *candefacio* 'to whitewash').

As far as the output of whole compounds are concerned, it is worth noticing that LC creates mostly adjectives (e.g. compounds with *-fer* as second constituent, such as *alifer* 'winged'), followed by nouns and verbs. Conversely, in Romance languages, compounding is exploited to create primarily nouns and less frequently adjectives. In Italian, there are very few cases of verbs obtained through compounding, which are made up of a noun and a verb (e.g. *manomettere* 'to tamper

[4] N: noun; V: verb; A: adjective; I: invariable form (i.e. adverb, conjunction); PR: pronoun.

[5] However, as reported below in section 3.3, in order to interpret correctly the data in Table 2, a distinction should be made between adjectives and adjectival participles, which are categorised here as V.

with'); the formation of pronouns and invariable forms through compounding does not seem to be productive anymore.

3.3 Some *caveats*

The main bedrock of WFL methodology lies in its strict relation to the morphological analyser Lemlat and on the PoS categorisation dictated by its lexical basis. As a consequence, the way compound constituents are pigeonholed can sometimes be unconventional. This impacts the representation of compounds in WFL in the following ways:

1. Adjectives that derive or function like participles are not included in the Lemlat lexical basis, because they are seen as part of the verbal paradigm, this means that certain compounds that would be expected to have a A as one of their constituents have a V instead. e.g *altivolans* (altus + volo) 'high flying' can be found among V+V=A compounds rather than among A+V=A.
2. certain type of adverbs ending in *-e* are considered in Lemlat ablative cases of the adjectival declension, so *dulciloquus* (dulce + loquor) 'sweet talking' is to be found among A+V=A, rather than I+V=A.

Another principle lying behind WFL's methodology is that *Oxford Latin Dictionary* acts like a sort of manual for solving a number of theoretical issues. For instance, unlike some traditional studies on Latin word-formation (i.e. (Benedetti, 1988), (Fruyt, 2002) and (Fruyt, 2011)), prepositions (e.g. *cum* 'with' or *in* 'in') are not included among compounding input elements in WFL, due to the overlap with prefixes. However, this can lead to inconsistencies. For instance, in OLD there is a clear distinction between affixes and isolated words where the lemmas' formative elements are specified. This means that words including what OLD considers a prefix, such as *quadriennium* 'period of four years' (*quadri-* 'consisting of four of the things following' + *annus* 'year', and not *quatuor* 'four' + *annus*) are included among prefixes, while other similar lemmas formed by numerals, like *sexennium* 'period of six years', on the other hand, are labelled as N+N=N compounds, because OLD categorises *sex* 'six' as a noun.

Moreover, in certain cases, it was decided to treat certain lemmas, which are generally seen as compounds, as conversions instead. For example, A+V=V and N+V=V compounds ending in *-fico*, i.e. involving the verb*facio* 'to do', which have often a corresponding adjective ending in *-ficus*. The assumption here is that the verbal compound must have been born before the adjective, as the main meaning of such compounds is almost always the result of a performed action (*amplifico* = amplus facio, 'to make (something) bigger'). In WFL, the corresponding adjective *amplificus* 'magnificent', has been connected to 'amplifico' through a conversion relationship V-to-A. This allows the two lemmas to appear in the same derivational tree.

4 Conclusions and future work

This paper has provided an overview of how compounding is represented in WFL, a derivational lexicon for Latin. This preliminary study, with its quantitative analysis in the field of LC, shows the potential for raising new questions and issues offered by a resource that for the first time collects all compounds used in Classical Latin. For instance, representing all compounding rules into a network, as it has been already successfully done for the affixal rules listed in WFL, (Litta et al., 2017), could lead to further research questions. These could be the investigation on constituent typologies or on the productivity of the different types of compounds. Future developments in WFL should be a way of searching through constituents by original lemma (currently still missing), and implementing a way of marking those PoS that appear differently in the resource's lexical basis. This would also allow for a more precise quantitative investigation on constituent typologies.

References

Marina Benedetti. 1988. *I composti radicali latini. Esame storico e comparativo*. Giardini, Pisa.

Luisa Brucale. 2012. Latin compounds. *Probus*, 24: 93-117.

Marco Budassi and Marco Passarotti. 2016. Nomen omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. *Proceedings of the 10^{th} SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, 90-94. Berlin: The Association for Computational Linguistics.

Egidio Forcellini. 1940. *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Cor-*

radini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos, Perin. Typis Seminarii, Padova.

Michèle Fruyt. 2011. Word-Formation in Classical Latin. *A companion to the Latin language*, 157-175.

Michèle Fruyt. 2002. Constraints and productivity in Latin nominal compounding. *Transactions of the Philological Society*, 100(3): 259-287.

Karl Ernst Georges and Heinrich Georges. 1972. *Ausführliches Lateinisch-Deutsches Handworterbuch*, Hahn, Hannover.

Peter G.W. Glare. 1982. *Oxford Latin Dictionary*. Oxford University Press, Oxford.

Otto Gradenwitz. 1904. *Laterali Vocum Latinarum*. Hirzel, Leipzig.

Albert Grenier. 1912. *Ètude sur la formation et l'emploi des composès nominaux dans le latin archaique*. Berger-Levrault, Paris.

Charles F. Hockett. 1954. Two Models of Grammatical Description. *Words*, 10: 210-231.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC–it 2016). Napoli, aAccademia University Press*, 185-189.

Eleonora Litta, Marco Passarotti and Paolo Ruffolo. 2017. Node Formation. Using Networks to Inspect Productivity in Affixal Derivation in Classical Latin. In *Proceedings of DATeCH2017, Göttingen, Germany, June 01-02, 2017*, 6 pages. DOI: http://dx.doi.org/10.1145/3078081.3078092

Renato Oniga. 1992. Compounding in Latin. *Rivista di linguistica*, 4(1): 97-116.

Renato Oniga. 1988. *I composti nominali latini: una morfologia generativa*. Patron, Bologna.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 133, 24-31.

Giulio Puccioni. 1944. L'uso stilistico dei composti nominali latini. *Atti della Accademia d'Italia. Memorie della classe di scienze morali e storiche*, Series 7, 4(10): 372-481.

Magda Ševčíková and Zdeněk Žabokrtský. 2014. Word-Formation Network for Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland*: 1087-1093, ELRA.

Luigi Talamo, Chiara Celata and Pier Marco Bertinetto. 2016. Derivatario: an annotated lexicon of Italian derivatives. *Word Structures*, 9(1): 72-102.

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra and Adéla Limburská. 2016. Merging Data Resources for Inflectional and Derivational Morphology in Czech, In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*: 1307-1314, ELRA.

Britta D. Zeller, Jan Snajder, and Sebastian Padò. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. *ACL*, 1: 1201-1211.

Sanremo's winner is...
Category-driven Selection Strategies for Active Learning

Anne-Lyse Minard, Manuela Speranza, Mohammed R. H. Qwaider, Bernardo Magnini
Fondazione Bruno Kessler, Trento, Italy
{minard,manspera,qwaider,magnini}@fbk.eu

Abstract

English. This paper compares Active Learning selection strategies for sentiment analysis of Twitter data. We focus mainly on category-driven strategies, which select training instances taking into consideration the confidence of the system as well as the category of the tweet (e.g. positive or negative). We show that this combination is particularly effective when the performance of the system is unbalanced over the different categories. This work was conducted in the framework of automatically ranking the songs of "Festival di Sanremo 2017" based on sentiment analysis of the tweets posted during the contest.

Italiano. *Questo lavoro confronta strategie di selezione di Active Learning per l'analisi del sentiment dei tweet focalizzandosi su strategie guidate dalla categoria. Selezioniamo istanze di addestramento combinando la categoria del tweet (per esempio positivo o negativo) con il grado di confidenza del sistema. Questa combinazione è particolarmente efficace quando la distribuzione delle categorie non è bilanciata. Questo lavoro aveva come scopo il ranking delle canzoni del "Festival di Sanremo 2017" sulla base dell'analisi del sentiment dei tweet postati durante la manifestazione.*

1 Introduction

Active Learning (AL) is a well known technique for the selection of training samples to be annotated by a human when developing a supervised machine learning system. AL allows for the collection of more useful training data, while at the same time reducing the annotation effort (Cohn et al., 1994). In the AL framework samples are usually selected according to several criteria, such as informativeness, representativeness, and diversity (Shen et al., 2004).

This paper investigates AL selection strategies that consider the categories the current classifier assigns to samples, combined with the confidence of the classifier on the same samples. We are interested in understanding whether these strategies are effective, particularly when category distribution and category performance are unbalanced. By comparing several options, we show that selecting low confidence samples of the category with the highest performance is a better strategy than selecting high confidence samples of the category with the lowest performance.

The context of our study is the development of a sentiment analysis system that classifies tweets in Italian. We used the system to automatically rank the songs of Sanremo 2017 based on the sentiment of the tweets posted during the contest.

The paper is structured as follows. In Section 2 we give an overview of the state-of-the-art in selection strategies for AL. Then we present our experimental setting (Section 3) before detailing the tested selection strategies (Section 4). Finally, we describe the results of our experiment in Section 5 and the application of the system to ranking Sanremo's songs in Section 6.

2 Related Work

AL (Cohn et al., 1994; Settles, 2010) provides a well known methodology for reducing the amount of human supervision (and the corresponding cost) for the production of training datasets necessary in many Natural Language Processing tasks. An incomplete list of references includes Shen et al. (2004) for Named Entity Recognition, Ringger et al. (2007) for PoS Tagging, and Schohn and Cohn (2000) for Text Classification.

AL methods are based on strategies for sample selection. Although there are two main types of selection methods, certainty-based and committee-based, here we concentrate only on certainty-based selection methods. The main certainty-based strategy used is the uncertainty sampling method (Lewis and Gale, 1994). Shen et al. (2004) propose a strategy which is based on the combination of several criteria: informativeness, representativeness, and diversity. The results presented by Settles and Craven (2008) show that information density is the best criterion for sequence labeling. Tong and Koller (2002) propose three selection strategies that are specific to SVM learners and are based on different measures taking into consideration the distances to the decision hyperplane and margins.

Many NLP tasks suffer from unbalanced data. Ertekin et al. (2007) show that selecting examples within the margin overcomes the problem of unbalanced data.

The previously cited selection strategies are often applied to binary classification and do not take into account the predicted class. In this work we are interested in multi-class classification tasks, and in the problem of unbalanced data and dominant classes in terms of performance.

Esuli and Sebastiani (2009) define three criteria that they combine to create different selection strategies in the context of multi-label text classification. The criteria are based on the confidence of the system for each label, a combination of the confidence of each class for one document, and a weight (based on the F1-measure) assigned to each class to distinguish those for which the system performs badly. They show that in most of the cases this last criteria does not improve the selection.

Our applicative context is a bit different as we are not working on a multi-label task. Instead of computing a weight according to the F1-measure, we experimented with a change of strategy where we focus on a single class.

3 Experimental Setting

The context of our study was the development of a supervised sentiment analysis system that classifies tweets into one of the following four classes: `positive`, `negative`, `neutral`, and `n/a` (i.e. not applicable).

The manual annotation of the data was mainly performed by 25 3rd and 4th year students from local high schools who were doing a one-week group internship at Fondazione Bruno Kessler.

We created an initial training set using an AL mechanism that selects the samples with the lowest system confidence[1], i.e. those closer to the hyperplane and therefore most difficult to classify. In the following we describe the sentiment analysis system, the Active Learning process and the creation of the test and the initial training set. Finally, we introduce the experiments performed on selection strategies for Active Learning.

Sentiment Analysis System. Our system for sentiment analysis is based on a supervised machine learning method using the SVM-MultiClass tool (Joachims et al., 2009)[2]. We extract the following features from each tweet: the tokens composing the tweet, and the number of urls, hashtags, and aliases it contains. It takes as input a tokenized tweet[3] and returns as output its polarity.

AL Process. We used TextPro-AL, a platform which integrates an NLP pipeline, an AL mechanism and an annotation interface (Magnini et al., 2016). The AL process is as follows: (i) a large unlabeled dataset is annotated by the sentiment analysis system (with a small temporary model used to initialize the AL process[4]); (ii) samples are selected according to a selection strategy; (iii) annotators annotate the selected tweets; (iv) the new annotated samples are accumulated in the batch; (v) when the batch is full the annotated data are added to the existing training dataset and a new model is built; (vi) the unlabeled dataset is annotated again using the newly built model and the cycle begins again at (ii).

The unlabeled dataset consists of 400,000 tweets that contained the hashtag #Sanremo2017. The maximum size of the batch is 120, so retraining takes place every 120 annotated tweets.

Training and Performance. The initial training set, whose creation required half a day of work[5], is

[1] The confidence score is computed as the average of the margin estimated by the SVM classifier for each entity.

[2] `https://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html`

[3] Tokenization is performed using the Twokenizer java library `https://github.com/vinhkhuc/Twitter-Tokenizer/blob/master/src/Twokenizer.java`

[4] The temporary model has been built using 155 tweets annotated manually by one annotator. After the first step of the AL process, these tweets are removed from the training set.

[5] The 25 high schools students worked in pairs and trios, for a total of 12 groups.

composed of 2,702 tweets. The class negative is the most represented, covering almost 40% of the total, with respect to positive, with around 30% of the total. The distribution of the two minor classes is rather close, with 18% for neutral and 13% for n/a.

As a test set we used 1,136 tweets randomly selected from among all the tweets which mentioned either a Sanremo song or singer. The test set was annotated partly by the high school students (656 tweets) and partly by two expert annotators (480 tweets); each tweet was annotated with the same category by at least two annotators. 58% of the tweets are positive, 20% are negative, 14% are neutral, and 8% are n/a.

We built the test set selecting the tweets randomly from the unlabeled dataset in order to make it representative of the whole dataset.

The overall performance of the system trained on the initial set is 40.7 in terms of F1 (see EVAL2702 in Table 1). The F1 obtained on the two main categories, i.e. positive and negative, is 54.5, but the system performs more poorly on negative than on positive, with F1-measures of 33.6 and 75.4 respectively.

Experiment. As the evaluation showed good results on positive but poor results on negative, we devised and tested novel selection strategies better able to balance the performance of the system over the two classes. We divided the 25 annotators into three different groups: each group annotated 775 tweets. The tweets annotated by the first group were selected with the same strategy used before, whereas for the other two groups we implemented two new selection strategies taking into account not only the confidence of the system but also the class it assigns to a tweet. As a result we obtained three different extensions of the same size and were thus able to compare the performance of the system trained on the initial training set plus each of the extensions.

4 Selection Strategies

We tested three selection strategies that take into account the classification proposed by the system in order to select the most useful samples to improve the distinction between positive and negative.

S1: low confidence. The first strategy we tested is the baseline strategy, which selects tweets classified by the system with the lowest confidence. The low confidence strategy was also used to build the initial training set (S0: lowC) as described is Section 3.

S2: NEGATIVE with high confidence. The second strategy consists of selecting the samples classified as negative with the highest confidence. We assume that this will increase the amount of negative tweets selected, thus enabling us to improve the performance of the system on the negative class. Nevertheless, as the system has a high confidence on the classification of these tweets, through this strategy we are adding easy examples to the training set that the system is probably already able to classify correctly.

S3: POSITIVE with low confidence. The third strategy aims at selecting the positive tweets for which the system has the lowest confidence. We expect in this way to get the difficult cases, i.e. tweets that are close to the hyperplane and that are classified as positive but whose classification has a high chance of being incorrect.

As the initial system has high recall (82.8) but low precision (69.3) for the class positive, we assume that it needs to improve on the examples wrongly classified as positive. We expect that inside the tweets wrongly classified as positive we will find difficult cases of negative tweets which will help to improve the system on the negative class. On the other hand, recall for the negative class is low (25.7), whereas precision is slightly better (48.7), which is why we decided to extract positive tweets with low confidence instead of negative tweets with low confidence.

5 Results and Discussion

In Table 1 we present the results (in tersm of F1) obtained by the system using the additional training data selected through the three different selection strategies described above. In order to facilitate the interpretation of the results, we also report the performance obtained by the system trained only on the initial set of 2,702 tweets. Additionally, in Table 2, we give the results obtained by the system for each configuration also in terms of recall and precision (besides F1).

The first four lines report the results for each of the four categories, while lines six and seven report respectively the macro-average F1 over the four classes and the macro-average F1 over the

		Eval2702		Experiment on selection strategies					
Strategy used		S0: lowC		S1: lowC		S2: NEG-highC		S3: POS-lowC	
		F1	tweets	F1	tweets	F1	tweets	F1	tweets
NEGATIVE		33.6	1,080	34.8	1,374	32.0	1,669	39.3	1,299
	wrt S0	-	-	(+1.2)	(+294)	(-1.6)	(+589)	(+5.7)	(+219)
POSITIVE		75.4	798	74.8	975	74.8	869	76.5	1,065
	wrt S0	-	-	(-0.6)	(+177)	(-0.6)	(+71)	(+1.1)	(+267)
NEUTRAL		22.3	476	20.9	595	23.3	567	24.6	672
	wrt S0	-	-	(-1.4)	(+119)	(+1.0)	(+91)	(+2.3)	(+196)
N/A		31.3	348	28.6	533	27.6	372	28.6	441
	wrt S0	-	-	(-2.7)	(+185)	(-3.7)	(+24)	(-2.7)	(+93)
Average 4 classes		40.7	2,702	39.8	3,477	39.4	3,477	42.3	3,477
	wrt S0	-	-	(-0.9)	(+775)	(-1.3)	(+775)	(+1.6)	(+775)
Average POS/NEG		54.5	-	54.8	-	53.4	-	57.9	-
	wrt S0	-	-	(+0.3)	-	(-1.1)	-	(+3.4)	-

Table 1: Performance of the system trained on 2,702 tweets and performance of the system trained on the same set of data incremented with 775 tweets selected through three different selection strategies.

	Eval2702			Experiment on selection strategies								
Strategy used	S0: lowC			S1: lowC			S2: NEG-highC			S3: POS-lowC		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
NEGATIVE	25.7	**48.7**	33.6	28.4	45.0	34.8	24.3	46.6	32.0	30.6	54.8	39.3
POSITIVE	**82.8**	69.3	75.4	81.6	69.0	74.8	82.2	68.7	74.8	**85.3**	69.3	76.5
NEUTRAL	20.1	**25.0**	22.3	17.7	25.4	20.9	20.7	26.6	23.3	21.3	**29.2**	24.6
N/A	**32.6**	30.0	31.3	30.4	26.9	28.6	29.3	26.0	27.6	27.2	30.1	28.6
Average 4 classes	40.3	43.2	40.7	39.5	41.6	39.8	39.2	41.9	39.4	41.1	45.9	42.3
Average POS/NEG	54.3	59.0	54.5	55.0	57.0	54.8	53.3	57.6	53.4	57.9	62.1	57.9

Table 2: Performance in terms of precision, recall and F1 of the system trained on the different training set. The two last lines are the average of the recall, precision and F1 over 4 and 2 classes.

two most important classes, i.e. positive and negative. For each selection strategy, we indicate the difference in performance obtained with respect to the system trained on the initial set, as well as the number of annotated tweets that have been added.

With the baseline strategy (S1: lowC, i.e., selection of the tweets for which the system has the lowest confidence) the performance of the system decreases slightly, from an F1 of 40.7 to an F1 of 39.8. Most of the added samples are negative tweets (38%), which enables the system to increase its performance on this class by 1.2 points.

When using the second strategy (S2: NEG-highC, i.e. selection of the negative tweets with the highest confidence), 76% of the new tweets are negative, but the performance of the system on this class decreases. Even the overall performance of the system decreases, despite adding 775 tweets.

We observe that the best strategy is S3 (POS-lowC, i.e., selection of the positive tweets with the lowest confidence), with an improvement of the macro-average F1-measure over the 4 classes by 1.6 points and over the positive and negative classes by 3.4 points. Although we add more positive than negative tweets to the training data (34%), the performance of the system on the negative class increases as well, from F1 33.6 to F1 39.3. This strategy worked very well in enabling us to select the examples which help the system discriminate between the two main classes.

6 Application: Sanremo's Ranking

After evaluating the three different selection strategies, we trained a new model using all the tweets that had been annotated. With this new model, as expected, we obtained the best results. The average F-measure on the negative and positive classes is 58.2, the average F-measure over the 4 classes is 42.1.

For the annotation to be used for producing the automatic ranking, we provided the system with some gazetteers, i.e. a list of words that carry positive polarity and a list of words that carry negative polarity. We thus obtained a small improvement in system performance, with an F1 of 42.8 on the average of the four classes and an F1 of 58.3 on the average of positive and negative.

As explained in the Introduction, the applicative scope of our work was to rank the songs competing in Sanremo 2017. For this, we used only the total number of tweets talking about each singer and the polarity assigned to each tweet by the system. In total we had 118,000 tweets containing either a reference to a competing singer or song that had been annotated automatically by the sentiment analysis system. By doing the ranking according to the proportion of positive tweets of each singer, we were able to identify 4 out of the top 5 songs and 4 out of the 5 last place songs. In Table 3, we show the official ranking versus the automatic ranking. The Spearman's rank correlation coefficient between the official ranking and our ranking is 0.83, and the Kendall's tau coefficient is 0.67

Singer	Official	System
Francesco Gabbani	1	8
Fiorella Mannoia	2	4
Ermal Meta	3	1
Michele Bravi	4	2
Paola Turci	5	5
Sergio Sylvestre	6	6
Fabrizio Moro	7	3
Elodie	8	9
Bianca Atzei	9	13
Samuel	10	7
Michele Zarrillo	11	10
Lodovica Comello	12	12
Marco Masini	13	14
Chiara	14	11
Alessio Bernabei	15	16
Clementino	16	15

Table 3: Sanremo's official ranking and the ranking produced by our system

7 Conclusion

We have presented a comparative study of three AL selection strategies. We have shown that a strategy that takes into account both the automatically assigned category and the system's confidence performs well in the case of unbalanced performance over the different classes.

To complete our study it would be interesting to perform further experiments on other multi-classification problems. Unfortunately this work required intensive annotation work and so its replication on other tasks would be very expensive. A lot of work on Active Learning has been done using existing annotated corpora, but we think that it is too far from a real annotation situation as the datasets used are generally limited in tems of size.

In order to test different selection strategies, we have evaluated the sentiment analysis system against a gold standard, but we have also performed an application-oriented evaluation by ranking the songs participating in Sanremo 2017.

As future work, we want to explore the possibility of automatically adapting the selection strategies while annotating. For example, if the performance of the classifier of one class is low, the strategy in use could be changed in order to select the samples needed to improve on that class.

Acknowledgments

This work has been partially funded by the Euclip-Res project, under the program *Bando Innovazione 2016* of the Autonomous Province of Bolzano. We also thank the high school students who contributed to this study with their annotation work within the FBK Junior initiative.

References

David Cohn, Richard Ladner, and Alex Waibel. 1994. Improving generalization with active learning. In *Machine Learning*, pages 201–221.

Seyda Ertekin, Jian Huang, Léon Bottou, and C. Lee Giles. 2007. Learning on the border: active learning in imbalanced data classification. In Mário J. Silva, Alberto H. F. Laender, Ricardo A. Baeza-Yates, Deborah L. McGuinness, Bjørn Olstad, Øystein Haug Olsen, and André O. Falcão, editors, *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pages 127–136. ACM.

Andrea Esuli and Fabrizio Sebastiani. 2009. Active learning strategies for multi-label text classification. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soulé-Dupuy, editors, *Advances in Information Retrieval, 31th European*

Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings, volume 5478 of *Lecture Notes in Computer Science*, pages 102–113. Springer.

Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. 2009. Cutting-plane training of structural svms. *Mach. Learn.*, 77(1):27–59, October.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proc.International ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pages 3–12, New York, NY, USA. Springer-Verlag New York, Inc.

Bernardo Magnini, Anne-Lyse Minard, Mohammed R. H. Qwaider, and Manuela Speranza. 2016. TEXTPRO-AL: An Active Learning Platform for Flexible and Efficient Production of Training Data for NLP Tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*.

Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Greg Schohn and David Cohn. 2000. Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 839–846, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1070–1079. ACL.

Burr Settles. 2010. Active learning literature survey. Technical report.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March.

Dalla *Word Sense Disambiguation* alla Sintassi: il Problema dell'Articolo Partitivo in Italiano

Ignazio Mauro Mirto
Università degli Studi di Palermo
Dipartimento Culture e Società
V.le delle Scienze Ed. 15 - 90128 Palermo
ignazio.mauro.mirto@unipa.it

Emanuele Cipolla
posta@emanuelecipolla.net

Abstract

Italiano. Fuori contesto, un nesso come *dei professori* non dà certezza di dove collocare *dei* in relazione alle parti del discorso. Il nesso può per esempio valere o *alcuni professori* (per es. in *Dei professori intervennero*) o esprimere appartenenza (per es. *i libri dei professori*). Nel primo caso *dei* è l'articolo partitivo di un nesso *nominale*, nel secondo è la preposizione che introduce un complemento di specificazione. Questo caso di omonimia si può far rientrare nell'area del *Word Sense Disambiguation*, ma la sua rilevanza per la sintassi e per il NLP è evidente. Nonostante ciò, in letteratura di esso non abbiamo trovato tracce. Il lavoro distingue diverse funzioni dei membri della serie e propone un algoritmo per disambiguare i due usi riferiti e altri, per esempio i complementi retti (come in *Approfittano dei tuoi fratelli*) che rendono la disambiguazione ancora più complessa.

English. *Out of context, a phrase of Italian such as dei professori 'of.the teachers' is ambiguous: it can either mean some teachers (e.g. Dei professori intervennero 'Some teachers attended') or carry the value of a Saxon genitive (e.g. i libri dei professori 'the teachers' books'). The part of speech to which dei belongs cannot be identified: dei could be a partitive article in a noun phrase or a preposition in a prepositional phrase. This key difference raises a problem in the area of Word Sense Disambiguation. Despite its relevance for NLP, to the best of our knowledge this case of homonymy has so far been disregarded in the literature. The paper distinguishes a number of functions dei carries and proposes an algorithm that can automatically discriminate between the two uses mentioned above, but also identify others that make the picture more complex.*

1 Introduzione

Questo lavoro verte sull'articolo partitivo in italiano, etimologicamente formato da *di* e da un articolo determinativo. L'intera serie, *del*, *dello*, *dell'*, *della*, *dei, degli*, *delle*, si presenta in superficie identica alle omonime preposizioni articolate.

Anche a un primo sguardo, la varietà di esiti che si ottiene collocando una sequenza come *dei professori* in contesti differenti, con *dei* qui preso come elemento rappresentativo dei sette membri della serie, desta stupore per la numerosità degli usi e le conseguenti difficoltà che ciò crea nel NLP.

Obiettivo del nostro lavoro è la disambiguazione automatica. Le difficoltà che un tale compito pone sono numerose. Lo studio è parte di una ricerca più ampia che ha come fine l'individuazione automatica del Soggetto[1] di una frase semplice (Mirto and Cipolla, 2017). In genere, l'articolo partitivo non è elemento frequente nei testi, ma la sua rilevanza al fine di ottenere maggiore precisione nella ricerca del Soggetto è evidente, come si vedrà nel prosieguo.

La sezione 2 è dedicata alle ambiguità semantiche che l'omonimia genera, derivanti da ambiguità strutturali. La sezione 3 presenta alcuni degli àmbiti grammaticali che creano ostacoli per la corretta identificazione degli articoli partitivi. Ognuno di questi àmbiti ha determinato una parte dello script presentato, che è stato messo alla prova su un corpus formato da 463 occorrenze (casualmente scelte tra le complessive 580) degli ele-

[1] Che, a giudicare dal numero di lavori reperibili in letteratura, non sembra argomento che susciti grande interesse, in particolare per l'italiano. Si veda almeno (Dell'Orletta et al., 2005) e i riferimenti ivi contenuti.

menti del paradigma rinvenute nel romanzo *Palomar* di Italo Calvino. La sezione 4 conclude il lavoro presentando i risultati ottenuti.

2 Ambiguità

La frase *Parlarono dei professori* è semanticamente ambigua: il nesso *dei professori* può infatti essere interpretato come complemento di argomento (i professori sono l'argomento di cui qualcuno parla) oppure come Soggetto post-verbale, con *dei* equivalente, in buona sostanza, ad *alcuni* (*Parlarono dei/alcuni professori*).

Anche la frase *Sono dei professori* risulta ambigua, visto che "oscilla" tra un significato di appartenenza (*Questi libri sono dei professori*, se il Soggetto *questi libri* viene omesso) e un significato equativo, cioè con identità referenziale tra nesso preverbale e nesso postverbale (*Loro/Questi sono dei professori*, con *loro/questi* e *professori* che rimandano allo stesso referente). Già da questi casi è possibile intuire alcune delle difficoltà di parsing per l'italiano generate dall'articolo partitivo, che ricorre in ognuno dei due casi di ambiguità presentati.

Caratteristica precipua dell'articolo partitivo dell'italiano è la frequente possibilità di farne a meno, di ometterlo, a parità di significato e mantenendo inalterata l'accettabilità della frase. È possibile farlo, per esempio, in *Parlarono professori*, ovviamente non più ambigua, così come è possibile farne a meno negli usi equativi: *Loro sono professori*. Di contro, l'omissione risulta impossibile nel significato di appartenenza o di possesso: **Questi libri sono professori* e, chiaramente, anche in quello del complemento di argomento (**Loro parlarono professori*), qualora si desideri mantenere identico il significato e l'ineccepibilità della frase.

Ecco succintamente illustrato uno dei frequentissimi casi di ambiguità che si presentano nelle lingue naturali. Chiamata in causa è l'area di ricerca nota come *Word Sense Disambiguation* (Stevenson and Wilks, 2003). È bene riaffermare che l'ambiguità non è di tipo lessicale, essendo *dei* composto da morfemi grammaticali, quindi privi di contenuto descrittivo.

Un paio di tentativi su demo disponibili online[2], che fanno uso di *dependency parsing*, con frasi come *Degli alunni hanno starnutito* o *Dei ragazzi starnutirono*, entrambe con articolo partitivo, hanno dato per *dei* il lemma *di* e la categoria 'preposizione' (si noti che, di fatto, ciò esclude erroneamente il nesso dalla funzione di Soggetto):

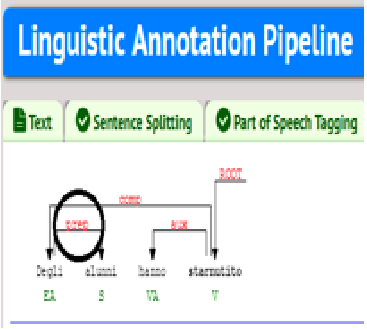
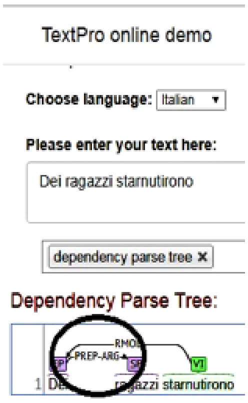

Linguistic Annotation Pipeline

Figure 1: Parsing con LinguA (03.07.2017)


TextPro online demo

Choose language: Italian

Please enter your text here:

Dei ragazzi starnutirono

Dependency Parse Tree:

Figure 2: Parsing con TextPro (11.07.2017)

Al di là dei tentativi di soluzione per fini pratici, si può affermare, più in generale, che a questo problema di omonimia in italiano la linguistica teorica e la semantica formale hanno dedicato molte attenzioni. Di contro, nel campo del NLP esso sembra essere passato inosservato.

L'algoritmo che presentiamo è stato implementato nel linguaggio Python 2.7[3]. Per effettuare *part of speech* e *lemma tagging*, al fine di identificare ad esempio nomi, verbi ed aggettivi, è stato utilizzato *TreeTagger* (`http://www.cis.uni-muenchen.de/` (`~schmid/tools/TreeTagger/`) con il file di parametro per l'italiano realizzati da

[2] Reperibili ai seguenti indirizzi: `http://linguistic-annotation-tool.italianlp.it/syntactic_trees` (figura 1), `http://hlt-services2.fbk.eu/textpro-demo/textpro.php` (figura 2)

[3] A IMM si deve la parte dello script che disambigua i potenziali articoli partitivi. EC si è fatto carico di tutte le indispensabili operazioni di annotazione su *TreeTagger*.

Marco Baroni, richiamato utilizzando il modulo treetagger-python (`https://github.com/miotto/treetagger-python`).[4]

L'algoritmo non si basa sulla nozione di costituente e le strategie adottate non fanno uso di 'alberi' di stampo chomskiano né di *dependency parsing*. Il parsing non è né *bottom up* né *top down*. Riteniamo che ai fini di una maggiore efficacia, cioè per un parsing in grado di identificare e risolvere ambiguità strutturali e semantiche, sarà indispensabile fare ricorso alla struttura argomentale dei predicati, o 'valenza', particolarmente di quelli verbali (Tesnière, 1959).

3 L'Algoritmo di Disambiguazione

Questa sezione mostra la suddivisione dello script di disambiguazione, basata sui diversi contesti di occorrenza dei morfemi della serie indagata. Complessivamente, nel corpus abbiamo identificato sette diversi casi: (I) complementi di specificazione; (II) complementi retti; (III) casi in cui ricorre il verbo *essere* o con funzione di ausiliare perfettivo o come copula; (IV) articoli partitivi con verbi transitivi e intransitivi; (V) comparativi e superlativi; (VI) nessi la cui testa è un pronome indefinito, (VII) locuzioni (*in fin dei conti*, *del resto*, *del tipo* (per es. *un larvato rimprovero del tipo "potresti pensarci un po' tu"*)). Il trattamento degli ultimi tre gruppi (tre occorrenze per (V), cinque per (VI), tre per (VII)) sarà oggetto di un'integrazione successiva.

3.1 *Dei* nel Complemento di Specificazione

Un nesso nominale come *i libri dei professori* esemplifica il complemento di specificazione. La serie che manifesta questo complemento contiene tutti gli elementi già elencati per l'articolo partitivo, ma, significativamente, se ne distingue perché include la forma *di* (*i libri di Leo*). Pur con questa massiccia sovrapposizione di forme, si ottengono distinte parti del discorso: se da un lato il partitivo è una forma di articolo (un determinante), dall'altro ciò che pare lo stesso elemento è invece una preposizione, che può essere articolata o semplice. Con l'unica differenza della preposizione semplice, tuttavia, al parser le forme si presentano identiche, fatto che impone una qualche risorsa che sia in grado di differenziare i due usi. Così, se la frase soggetta al parsing fosse *Abbiamo letto i libri di fisica dei professori*, non si avrebbe difficoltà a collocare *di* tra le preposizioni, mentre per *dei* si rivela necessaria un'operazione di disambiguazione.

Su questo caso di omonimia non siamo stati in grado di trovare in letteratura proposte precedenti. Suggeriamo in questa sede di individuare un complemento di specificazione grazie alla parola che precede la preposizione, che il più delle volte è o un nome o un aggettivo. La parte di codice rilevante, abbreviata e semplificata, è qui di seguito illustrata (`frase[i]` è il pivot):

```
# classificazione: 0=complemento
# di specificazione
for i in range(len(frase)):
    precedenti = frase[0:i]
    successivi = frase[i+2:len(frase)]

    compl_specificazione =
     frase[i] in maybe_partitive \
           and (frase[i-1] in nomi \
           or frase[i-1] in agg)

    if compl_specificazione is True:
            classificazione=0
```

Se tra gli elementi che precedono immediatamente una qualsiasi delle sette forme della serie, inserite nella tupla denominata *maybe_partitive*, si includono (a) i dimostrativi (per es. *il passo delle zampe posteriori [...] quello delle anteriori*), (b) i verbi all'infinito (per es. *l'espandersi della sabbia*), (c) alcune congiunzioni (*l'alfabeto delle onde marine o delle erbe d'un prato*), (d) casi di ricorsività (per es. *del tessuto del fondo*) e, infine, (f) occorrenze multiple con virgola (per es. *la percezione precisa dei contorni, dei colori, delle ombre*), la porzione di script sopra illustrata consente di identificare correttamente 388 complementi di specificazione, pari al $97,7\%$ delle occorrenze. Oltre a questi *true positives* si sono avuti 9 *false negatives*, 3 *false positives* e 63 *true negatives*; ciò dà luogo a una *precision* di 0.99 e ad una *recall* di 0.97; la F_1-*score* è pari a 0.97. Alcuni casi problematici sono: (i) la topicalizzazione del nesso preposizionale (per es. *Della conoscenza mitica degli astri egli capta solo qualche stanco barlume*); (ii) le nominalizzazioni (per es. *tutto il non detto della sua condizione*); oppure (iii) quello di *Ho trovato sul selciato degli uccelli malconci*, in cui *degli* svolge la funzione di articolo partitivo, ma viene erroneamente intercettato come complemento di specificazione a causa del locativo *sul*

[4] Il tratto [±Numerabile] del sostantivo che segue *dei* consentirebbe di escludere che in un nesso come *della penna* ricorra un articolo partitivo (* *Voglio della penna)*. La ricerca ne verrebbe semplificata. Questa risorsa non è stata utilizzata perchè *TreeTagger* non fornisce il tratto.

selciato che ricorre tra il verbo e il nesso nominale post-verbale.

Un paio di osservazioni finali. La prima: dal punto di vista semantico, il complemento di specificazione può esprimere un significato affine a quello di frasi copulative (§ 3.3) come *I libri sono dei professori*, significato cui ci si riferisce comunemente con ’appartenenza’ o ’possesso’. La seconda: è bene ribadire che né in *i libri dei professori* né in *I libri sono dei professori* è possibile sottrarre *dei* (**I libri professori*, **I libri sono professori*), proprio perché la sottrazione a parità semantica è caratteristica esclusiva dell’articolo partitivo, anche se tale opzione non è sempre praticabile.

3.2 *Dei* come Complemento Retto

Si tratta del caso esemplificato con il verbo *parlare*. La già discussa ambiguità della frase *Parlarono dei professori* deriva proprio dal fatto che *parlare* è verbo potenzialmente bivalente (o trivalente: *Leo parlò a Luigi di Ada*). Se l’esempio fosse modificato in *Dei professori parlarono*, con Soggetto anteposto, la frase rimarrebbe ancora ambigua, ma in modo diverso: o *dei professori* è un Soggetto canonicamente pre-verbale oppure, se ancora interpretato come complemento di argomento, esso è allora collocato in una posizione marcata e la frase, segmentata, necessita di un particolare profilo intonativo, cioè di una messa in rilievo tramite enfasi, di seguito richiamata con il maiuscoletto: DEI PROFESSORI *parlarono* (*non degli studenti*). L’esplicitazione del Soggetto porrebbe fine a ogni ambiguità: *Loro parlarono dei professori*.

In italiano i predicati che idiosincraticamente legittimano un complemento in *di* non sono necessariamente verbali. Ecco alcuni dei casi rinvenuti nel corpus, con verbi (il nesso non è né Soggetto né Oggetto diretto), aggettivi, avverbi, nomi e polirematiche (si notino le due topicalizzazioni):

- tener conto degli aspetti complessi
- ripaga del sapere che si propaga
- dell’adeguato innaffiamento approfittano le erbacce
- quello che ha pensato del prato
- spera d’essersi appropriato del pianeta
- faccio parte dei soggetti senzienti
- avrebbe più bisogno del nostro interessamento
- è specifico del sesso femminile
- anche del nulla non si può essere sicuri al cento per cento
- prima della sua nascita
- al di là delle abitudini sensoriali
- in balia della sovrapopolazione di questi lumpen-pennuti *[sic]*

Talvolta lo stesso verbo presenta più valenze, con differenze semantiche come *Chiedono dei professori* vs *Chiedono professori*, dunque con un ulteriore caso di ambiguità: *Chiedono a proposito dei professori* vs *Richiedono professori*. Individuare differenze così sottili richiede soluzioni complesse.

Nello script, i complementi di specificazione sono rilevati dopo i complementi retti. Il motivo è semplice: se la frase sottoposta al parsing fosse *Sandro è degno degli onori più grandi*, la funzione rileverebbe nella posizione precedente a *degli* un aggettivo, restituendo quindi un errore, cioè che *degli onori più grandi* è complemento di specificazione. Lo stesso accadrebbe con una polirematica come *tener conto delle proporzioni*, che nella posizione precedente a *delle* presenta un sostantivo.

I complementi retti introdotti da una delle forme omonime a quelle degli articoli partitivi sono complessivamente 33, pari al $7,1\%$ delle 463 occorrenze indagate.

Per l’individuazione dei complementi retti si è creata una lista, denominata *trigger_di*, contenente verbi, aggettivi, avverbi e locuzioni che legittimano un complemento introdotto dalla preposizione *di*. Con la suddivisione della stringa in ’precedenti’ e ’successivi’ rispetto al pivot l’algoritmo consente di calcolare se il complemento retto è anteposto al predicato che lo regge (ordine marcato) o posposto (ordine canonico):

```
# Classificazione complemento retto:
# 1=posposto, 2=anteposto
for j in range(len(frase)):
    if frase[j] in trigger_di:
        if frase[j] in precedenti:
            classificazione=1
        elif frase[j] in successivi:
            classificazione=2
```

3.3 *Dei* in frasi con *essere* come copula o con *esserci*

È uno dei casi presentati nella sezione 3 con frasi ambigue come *Sono dei professori*. Si noti che la frase *Ci sono dei professori*, in superficie diversa dalla precedente solo per la presenza del clitico *ci*, esemplifica un tipo denominato in letteratura 'esistenziale', che è tutt'altra cosa. Nella frase *Ci sono dei professori* il nesso *dei professori* fornisce un esempio di articolo partitivo. Ne è prova il fatto che *dei* può o essere rimosso senza che la frase collassi (*Ci sono professori*) o essere sostituito con *alcuni* (*Ci sono alcuni professori*). Le due frasi *Sono dei professori* e *Ci sono dei professori* sono dunque diverse dal punto di vista strutturale, al punto che mentre *dei professori* è il Soggetto dell'ultima, nella prima il Soggetto è omesso (*Essi sono dei professori* o *Questi libri sono dei professori*). L'algoritmo deve poter individuare tali differenze strutturali, come si propone nella porzione di codice che segue, che ha individuato due occorrenze di articolo partitivo con *esserci* (*ci sono delle forme e delle sequenze che si ripetono*) senza però essere riuscito ad individuare l'articolo partitivo nel seguente esempio: ([le mani del gorilla] *sono ancora in realtà delle zampe*):

```
# classificazione: 3=nome predicativo,
#                  no soggetto;
# 4=frase esistenziale: partitivo e soggetto

elif is_copulativo is True:
  if is_verbo(tt,frase[i-1],copulativi):
        if frase[i-2] != 'ci':
                classificazione = 3
        else:
                classificazione = 4
elif is_verbo(tt,frase[i-2],copulativi):
    if frase[i-3] != 'ci':
        classificazione = 3
    else:
       classificazione = 4
```

3.4 *Dei* in Soggetti o Oggetti di verbi transitivi e intransitivi

Se, al parsing, un elemento della serie *maybe_partitive* non è riconosciuto come complemento di specificazione, giacché non preceduto né da un nome né da un aggettivo (§ 3.1), oppure se la stringa non contiene né complementi retti (§ 3.2) né un'occorrenza di *essere* copula o di *esserci* (§ 3.3), allora siamo in presenza di un articolo partitivo in un nesso legittimato da un verbo transitivo o intransitivo, come in *Lui per trattenerla le dà dei piccoli morsi a una zampa* e *Esistono delle vie e delle piazze*. In questi casi *essere* può ovviamente ricorrere, ma come ausiliare perfettivo, dunque in combinazione con un participio passato: *Delle ombre silenziose si sono mosse sulla sabbia*. Si tratta in tutto di 10 delle 13 occorrenze complessive di articolo partitivo (2,7% del corpus, tre con *esistere*), così identificate:

```
# classificazione: 5=articolo partitivo
# post-verbale
# 6=articolo partitivo pre-verbale
elif is_forma_verbale is True:
    if frase[j] in precedenti:
      classificazione=5
    elif frase[j] in successivi:
      classificazione=6
```

4 Conclusioni

La procedura di disambiguazione automatica delle sequenze introdotte da *di* + articolo partitivo qui presentata ha dato luogo a risultati promettenti, in particolare per l'identificazione dei complementi di specificazione. Perchè si possa parlare di *information retrieval* è però necessario un campione statistico di una certa rilevanza; l'esiguità del numero di frasi ricadenti nei rimanenti casi di cui alla sezione 3 renderebbe i relativi indicatori privi di utilità, per cui si è scelto di non proporli. Risulta necessario operare ancora sul corpus sia per trattare le rimanenti occorrenze già identificate, sia per arricchirlo di nuove frasi. Inoltre, poichè l'algoritmo lavora per eliminazione, potrebbe essere utile proporre un diverso ordine di valutazione dei casi, in vista di risultati migliori.

References

Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2005. Climbing the path to grammar: A maximum entropy model of subject/object learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, PMHLA '05, pages 72–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ignazio Mauro Mirto and Emanuele Cipolla. 2017. Nooj assisted automatic detection of errors in auxiliaries and past participles in italian. In *Proceedings of the NooJ 2017 International Conference*.

Mark Stevenson and Yorick Wilks. 2003. Word sense disambiguation. *The Oxford Handbook of Comp. Linguistics*, pages 249–265.

Lucien Tesnière. 1959. *Eléments de Syntaxe Structurale*. Paris.

PARSEME-It Corpus
An annotated Corpus of Verbal Multiword Expressions in Italian

Johanna Monti[1], Maria Pia di Buono[2], Federico Sangati[3]

jmonti@unior.it, mariapia.dibuono@fer.hr, federico.sangati@gmail.com
[1]Dep. of Literary, Linguistic and Comparative Studies "L'Orientale" University of Naples, Italy
[2]TakeLab - University of Zagreb, Croatia
[3]Indipendent Researcher, Italy

Abstract

English. This paper describes a new language resource annotated with verbal multiword expressions (VMWEs) in Italian. The paper discusses the state of the art in VMWE identification and annotation in Italian, the methodology adopted, the various VMWE categories annotated, the corpus and the annotation process. Finally, the paper ends with results, conclusion and future work.

Italiano. *Questo contributo descrive una nuova risorsa linguistica annotata con polirematiche verbali per la lingua italiana. Viene presentato lo stato dell'arte relativamente all'identificazione ed all'annotazione di polirematiche per la lingua italiana, la metodologia adottata, le diverse categorie di polirematiche verbali annotate nel corpus, il corpus stesso e il processo di annotazione. Infine vengono illustrati i risultati ottenuti, le conclusioni e le prospettive future.*

1 Introduction

This paper outlines the development of a new language resource for Italian, namely the **PARSEME-It VMWE corpus**, annotated with Italian MWEs of a particular class: verbal multiword expressions (VMWE). The PARSEME-It VMWE corpus has been developed by the PARSEME-IT research group[1] in the framework of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (Savary et al., 2017), a joint effort, carried out within a European research network, to elaborate universal terminologies and annotation guidelines for verbal multiword expressions in 18 languages, among which also the Italian language is represented. Notably, multiword expressions represent a difficult lexical construction to identify, model and treat by Natural Language Processing (NLP) tools, such as parsers, machine translation engines among others, mainly due to their non-compositional property. In particular, among multiword expressions verbal ones are particularly challenging because they have different syntactic structures (*prendere una decisione* 'make a decision', *decisioni prese precedentemente* 'decisions made previously'), may be continuous and discontinuous (*andare e venire* versus *andare in malora* in *Luigi ha fatto andare la società in malora*), may have a literal and figurative meaning (*abboccare all'amo* 'bite the hook' or 'be deceived'). In this paper, we describe the state of the art in VMWE annotation and identification for the Italian language (section 2). We then present the methodology (section 3), the Italian VMWE categories taken into account for the annotation task (section 4), the corpus and the annotation process (section 5), and the results (section 6). Finally, we discuss conclusions and future work (section 7).

2 State of the art in VMWE identification and annotation in Italian

Several scholars have investigated different kinds of Italian VMWEs, focusing on both syntactic and semantic aspects. Among these works, we may distinguish contrastive and comparative analyses, and synchronic and diachronic studies.
In the first group, most of the scholars propose a comparison with Germanic languages (Mateu and Rigau, 2010), mainly for describing verb-particle constructions, that represent a very common phenomenon in this family.
On the other hand, synchronic and diachronic

[1] https://www.researchgate.net/project/PARSEME-IT-Syntactic-Parsing-and-Multiword-Expressions-in-Italian

studies include analyses of: (i) verb-particle constructions (Masini, 2005; Iacobini and Masini, 2005; Quaglia and Trotzke, 2017), (ii) idiomatic constructions (Tabossi et al., 2011; Vietri, 2014c) with either ordinary or support verbs (Vietri, 2014b), (iii) support, or light, verbs, which represent a wider phenomenon and, for this reason, they have been largely analysed (La Fauci, 1980; D'Agostino and Elia, 1998; Cicalese, 1999; Alba-Salas, 2004; Quochi, 2007; Cicalese et al., 2016). Reflexive verbs in Italian have been investigated as occurrences of non-local anaphora (Reuland, 1990) and considering their syntactic classification (Carstea Romascanu, 1977).

To the best of our knowledge only a limited number of monolingual language resources with multiwords for the Italian language have been developed such as a dictionary for Italian idioms (Vietri, 2014a), a series of example corpora and a database of MWEs represented around morphosyntactic patterns (Zaninello and Nissim, 2010), or a corpus annotated with Italian MWEs of a particular class: verb-noun expressions such as *fare riferimento*, *dare luogo* and *prendere atto* (Taslimipoor et al., 2016). At the time of writing, therefore, the PARSEME-It VMWE corpus represents the first sample of a corpus, which includes several types of VMWEs, specifically developed for NLP applications.

3 Methodology

The development of the Italian VMWE corpus is based on the PARSEME annotation guidelines[2], provided for the shared task. The guidelines have been developed with the aim of delivering general definitions and prescriptions for the annotation of VMWEs in 18 languages, but, at the same time, of allowing language-specific descriptions of these linguistic phenomena (Savary et al., 2017). The annotation guidelines include three main categories:

1. a **universal category**, which is common to all the languages involved in the task and holds light-verb constructions (LVCs) and idioms (ID);
2. a **quasi-universal category**, relevant for some languages or language families, that contains inherently reflexive verbs (IReflVs) and verb-particle constructions (VPCs);
3. an **other VMWEs** category, which is a residual category for the occurrences not belonging to any of the previous groups.

In order to ease the identification and categorisation task of VMWEs, a decision tree method was devised with generic and language-specific tests. Generic tests consider general criteria that are valid for all languages, while language-specific tests consider structural, lexical, morphological and syntactic features that are specific for the individual languages. The decision tree includes three steps, (i) identification of a VMWE candidate, i.e., a combination of a verb with at least one other word, which is a potential VMWE; (ii) identification of the lexicalized elements of the expression, (iii) assignment of the VMWE to one of the VMWE categories, using general and language-specific tests.

4 Italian VMWEs

For the Italian VMWE annotation task, according to PARSEME guidelines, multiword expressions are understood as (continuous or discontinuous) sequences of words with the following compulsory properties:

- Their component words include a head word and at least one other syntactically related word. Most often the relation they maintain is a syntactic (direct or indirect) dependency but it can also be e.g., a coordination.
- They show some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is considered general grammar rules of a language.
- At least two components of such a word sequence have to be lexicalized.

In this task we only annotate the lexicalized components and ignore open slots. Collocations, i.e., word co-occurrences whose idiosyncrasy is of statistical nature only (e.g., *the graphic shows, drastically drop*, etc.), are excluded from the scope of this study. The VMWE which have been annotated for the Italian language are:

1. **Light verb constructions** (LVC), which typically consist of a verb and a noun or prepositional phrase, e.g., *fare una domanda* ('to

[2] The guidelines are available at http://parsemefr.lif.univ-mrs. fr/guidelines-hypertext/.

make a question'), *fare una passeggiata* ('to have a walk'). The verb has a purely syntactic operator function (performing an activity or being in a state), whereas the noun is predicative, often referring to an event (e.g., decision, visit) or a state (e.g., fear, courage);

2. **Idioms** (ID), which have at least two lexicalized components including a head verb and at least one of its arguments, e.g., *tirare le cuoia* ('kick the bucket'), *piovere a catinelle* ('rain cats and dogs');

3. **Inherently reflexive verbs** (IReflV), which are those reflexive verbal constructions which (a) never occur without the clitic e.g., *suicidarsi* ('suicide'), or when (b) the REFLV and non-reflexive versions have clearly different senses or subcategorization frames e.g., *riferirsi* ('refer');

4. **Verb particle combinations** (VPC), which are formed by a lexicalized head verb and a lexicalized particle dependent on the verb. The meaning of the VPC is non-compositional. Notably, the change in the meaning of the verb goes significantly beyond adding the meaning of the particle, e.g., *buttare giù* ('swallow'). This type of construction is very frequent in English, German, Swedish, Hungarian, but we can find them also in Italian;

5. **Other Verbal MWEs** (OTH), which gather the types not belonging to any of the categories above, e.g., *corto-circuitare* ('short-circuit').

5 Corpus and annotation task

5.1 PARSEME Italian VMWE corpus

The PARSEME-It VMWE corpus is based on a selection of texts taken from the *PAISA´* corpus of Italian web texts (Lyding et al., 2014). We chose this corpus because it contains documents (i) from different web sources, e.g., Wikibooks, Wikinews, Wikiversity, and several blog services from different websites, collected in 2010 by means of a Creative Commons-focused web crawling, and a targeted collection of documents from specific websites, (ii) dedicated to no specific technical domain, free from copyright issues, so as to be compatible with an open license (iii) annotated in CoNLL format, i.e. lemmatized, POS-tagged and annotated with syntactic dependencies. For our annotation task, we selected a sub-corpus formed by 17,000 sentences (corresponding to 421,848 tokens) randomly taken from blogs, Wikipedia and Wikinews. The corpus was kept in its original state and therefore no errors or inconsistencies were corrected. The pre-annotation of the *PAISA´* was kept in order to ease the annotation work with reference to the identification of verbal MWEs but we asked annotators not to overestimate the system's performances, and to review the whole text, not only the pre-annotated candidates proposed by the system. A dedicated tag in FLAT was defined for this purpose. The objective was to have a final corpus of at least 3,500 annotated VMWEs per language. Since the density of VMWEs highly depend on the particular language, as well as text choice and genre, we were not able to make any reliable estimation of the corpus size needed to reach this goal from the beginning of the task.

5.2 Annotation environment

The annotation environment used for the PARSEME-It VMWE corpus is FLAT, a web-based linguistic annotation environment[3] based around the FoLiA format[4] a rich XML-based format for linguistic annotation. FLAT allows users to view annotated FoLiA documents and enrich these documents with new annotations (Figure 1), a wide variety of linguistic annotation types is supported through the FoLiA paradigm. It is a document-centric tool that fully preserves and visualises document structure. It is open source software developed at the Centre of Language and Speech Technology, Radboud University Nijmegen and is licensed under the GNU Public License v3.

5.3 Annotation task

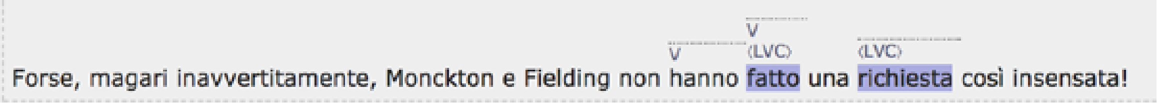
The annotation task for the Italian language was performed in five different stages.

1. The PARSEME Annotation guidelines were agreed on[5] and examples for the Italian language were added in order to ease the annotation task by the Italian annotators. To this end, a two-phase pilot annotation in Italian

[3] http://flat.readthedocs.io/en/latest/

[4] http://proycon.github.io/folia

[5] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/?page=home

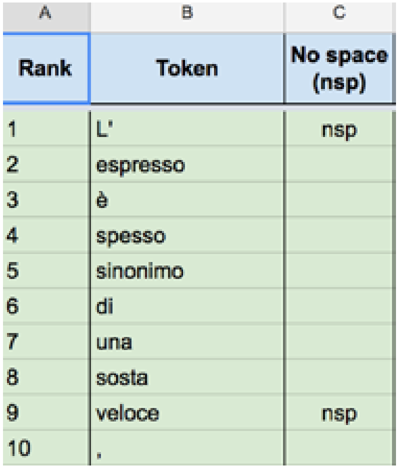


Forse, magari inavvertitamente, Monckton e Fielding non hanno fatto una richiesta così insensata!

Figure 1: Example of annotated data in FLAT

was carried out. This step was useful in identifying the Italian VMWE categories to be annotated, but also to promote cross-language convergences with the other languages foreseen in the shared task. Each pilot annotation phase provided feedback from annotators and was followed by enhancements of the guidelines, corpus format and processing tools.

2. A pre-processing step of the *PAISA´* corpus was needed: a 'no space' column was added to the files in order to add the 'nsp' tag if a token should have been appended to the previous one without a space.



A	B	C
Rank	Token	No space (nsp)
1	L'	nsp
2	espresso	
3	è	
4	spesso	
5	sinonimo	
6	di	
7	una	
8	sosta	
9	veloce	nsp
10	.	

Figure 2: Example of the use of an nsp tag

3. The annotation task of the training set (approx. 16,000 sentences) was manually performed in running texts using the FLAT environment by five Italian native speakers with linguistic background. Each annotator was given a certain number of files, containing 1,000 sentences in CoNLL format. All the doubts about the annotation were collected in a shared file and discussed during the annotation phase. Difficulties in annotating VMWE mainly concerned (i) the boundaries of the VMWE such as in *Sei ovviamente nel pieno diritto di esprimere [...]* where it is difficult to decide if the VMWE should be *sei ... nel ... diritto* or *sei ... nel pieno diritto*, (ii) the category attribution concerning for instance the *fare + N* VMWE type, since in some cases the category is LVC such as in *fare rumore* and in some others is ID such as in *fare schifo*, (iii) the identification of nested VMWEs like in *Mi guardo bene* where the annotator has to decide if in the ID *guardarsi bene* there is also a IReflV *guardarsi* or not.

4. A few files were double-annotated to evaluate the inter-annotator agreement (IAA). Measuring IAA is not a trivial task because of the challenges posed by VMWEs and described in the Introduction. The available IAA results organized per-VMWE F-score (F_{unit}), estimated Cohens K (K_{unit}) and finally standard K(K_{cat}) (Savary et al., 2017) scores are presented in Table 1.

5. Further 1,000 sentences were used as test-set during the shared task. The VMWE annotations were automatically annotated by the systems that took part in the shared task and performed according to the same guidelines.

	#S	#T	#A$_1$	#A$_2$	F$_{\mathrm{unit}}$	K$_{\mathrm{unit}}$	K$_{\mathrm{cat}}$
IT	2000	52639	336	316	0.417	0.331	0.78

Table 1: AA scores for Italian annotation: #S, and #T show the number of sentences and tokens in the corpora used for measuring the IAA, respectively. #A$_1$ and #A$_2$ refer to the number of VMWE instances annotated by each of the annotators (Savary et al., 2017).

6 Results

The PARSEME-It VMWE corpus is composed of 2,454 entries (Table 2), and it is freely available[6], released under Creative Commons licenses.

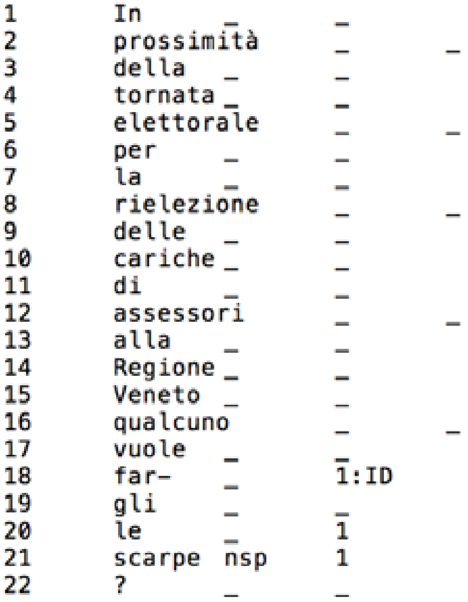
The data have been annotated using the official parseme-tsv format[7] (Figure 3), adapted from the CoNLL format.

[6] http://hdl.handle.net/11372/LRT-2282

[7] http://typo.uni-konstanz.de/parseme/index.php/2-general/184-parseme-shared-task-format-of-the-final-annotation.

Category	Occurrences
ID	1163
IReflV	730
LVC	482
VPC	73
OTH	6
Total	**2454**

Table 2: Overview of VMWEs in the PARSEME-It VMWE corpus, including train and test sets.



```
1       In      _       _
2       prossimità      _       _
3       della   _       _
4       tornata _       _
5       elettorale      _       _
6       per     _       _
7       la      _       _
8       rielezione      _       _
9       delle   _       _
10      cariche _       _
11      di      _       _
12      assessori       _       _
13      alla    _       _
14      Regione _       _
15      Veneto  _       _
16      qualcuno        _       _
17      vuole   _       _
18      far-    _       1:ID
19      gli     _       _
20      le      _       1
21      scarpe  nsp     1
22      ?       _       _
```

Figure 3: Example of annotated data in parseme-tsv format

In the official parseme-tsv format, as described in Savary et al. (2017), the information about each token are represented by 4 tab-separated columns featuring (i) the position of the token in the sentence or a range of positions (e.g., 1-2) in case of multiword tokens such as contractions, (ii) the token surface form, (iii) an optional flag indicating that the current token is adjacent to the next one, and (iv) an optional VMWE code composed of the VMWEs consecutive number in the sentence and for the initial token in a VMWE its category (e.g., 2:ID if a token starts an idiom which is the second VMWE in the current sentence). In case of nested, coordinated or overlapping VMWEs multiple codes are separated with a semicolon. Furthermore, in order to provide data usable as features in the shared task systems, also companion files in a format close to CoNLL-U[8] have been released. These companion files contain extra linguistic information, i.e., lemmas, POS-tags, morphological features, and syntactic dependencies.

7 Conclusion and Future Work

In this paper, we described a linguist resource of Italian VMWE, developed within the PARSEME Shared Task on Automatic Identification of VMWE. We consider this work an initial contribution for elaborating an Italian universal terminology of VMWE. Future work includes the extension of the current corpus and a fine-grained linguistic analysis of the annotation in order to contribute to the description of these phenomena.

Acknowledgments

The work described in this paper has been supported by the IC1207 PARSEME COST action. The annotation work was carried out also thanks to the help of Maarten van Gompel who adapted the FLAT annotation platform to the needs of the community.
Our thanks go also to the Italian annotators, Valeria Caruso, Manuela Cherchi, Anna De Santis, Annalisa Raffone, for their contributions.
Autorship contribution is as follows: Johanna Monti is author of Sections 1, 3, 4 and 5.3; Maria Pia di Buono of Sections 2 and 6 and 7; and Federico Sangati of Sections 5.1 and 5.2.

References

Josep Alba-Salas. 2004. Fare light verb constructions and italian causatives: Understanding the differences. *ITALIAN JOURNAL OF LINGUISTICS*, 16(2):283.

M Carstea Romascanu. 1977. I tipi di verbi riflessivi in italiano. *Revue Roumaine de Linguistique Bucuresti*, 22(2):125–130.

Anna Cicalese, Emilio D'Agostino, Alberto Maria Langella, and Ilaria Villari. 2016. Els verbs locatius com a variants de verbs de suport. *Quaderns d'Italià*, 21:153–166.

Anna Cicalese. 1999. Le estensioni di verbo supporto. uno studio introduttivo. *Studi italiani di linguistica teorica ed applicata*, 28(3):447–485.

Emilio D'Agostino and Annibale Elia. 1998. Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche. *AA. VV, Ai limiti del linguaggio. Bari: Laterza*, pages 287–310.

[8] http://universaldependencies.org/format.htm

Claudio Iacobini and Francesca Masini. 2005. Verb-particle constructions and prefixed verbs in italian: typology, diachrony and semantics. In *Mediterranean Morphology Meetings*, volume 5, pages 157–184.

Nunzio La Fauci. 1980. Aspects du mouvement de wh, verbes supports, double analyse, complétives au subjonctif en italien: pour une description compacte. *Lingvisticae Investigationes*, 4(2):293–341.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice DellOrletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisa corpus of italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43.

Francesca Masini. 2005. Multi-word expressions between syntax and the lexicon: the case of italian verb-particle constructions. *SKY Journal of Linguistics*, 18(2005):145–173.

Jaume Mateu and Gemma Rigau. 2010. Verb-particle constructions in romance: A lexical-syntactic account. *Probus*, 22(2):241–269.

Stefano Quaglia and Andreas Trotzke. 2017. Italian verb particles and clausal positions. In *IATL 31: The 31st annual meeting Israel Association for Theoretical Linguistics*, pages 67–82.

Valeria Quochi. 2007. A usage-based approach to light verb constructions in italian: Development and use.

Eric Reuland. 1990. Reflexives and beyond: Non-local anaphora in italian revisited. *Grammar in progress: glow essays for Henk van Riemsdijk*, 36:351.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemizadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47.

Patrizia Tabossi, Lisa Arduino, and Rachele Fanari. 2011. Descriptive norms for 245 italian idiomatic expressions. *Behavior Research Methods*, 43(1):110–123.

Shiva Taslimipoor, Anna Desantis, Manuela Cherchi, Ruslan Mitkov, and Johanna Monti. 2016. Language resources for italian: towards the development of a corpus of annotated italian multiword expressions. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. ceur-ws.

Simona Vietri. 2014a. The italian module for nooj. In *Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it*, pages 389–393.

Simonetta Vietri. 2014b. *Idiomatic Constructions in Italian: A Lexicon-grammar Approach*, volume 31. John Benjamins Publishing Company.

Simonetta Vietri. 2014c. The lexicon-grammar of italian idioms. In *Workshop on Lexical and Grammatical Resources for Language Processing, COLING 2014*, pages 137–146.

Andrea Zaninello and Malvina Nissim. 2010. Creation of lexical resources for a characterisation of multiword expressions in italian. In *LREC*.

MODAL: A multilingual corpus annotated for modality

Malvina Nissim
CLCG, University of Groningen
The Netherlands
m.nissim@rug.nl

Paola Pietrandrea
University of Tours, CNRS UMR7270
France
pietrandrea-guerrini@univ-tours.fr

Abstract

English. We have produced a corpus annotated for modality which amounts to approximately 20,000 words in English, French, and Italian. The annotation scheme is based on the notion of *epistemic construction* and virtually language-independent. The annotation is rigorously evaluated by means of a newly developed strategy based on the alignment of the entire epistemic constructions as identified and marked up two annotators. The corpus and the agreement scoring tools are publicly available.

Italiano. *Presentiamo un corpus multilingue di circa 20,000 parole annotato per modalità epistemica. La procedura di annotazione è guidata dal concetto di costruzione epistemic. La validità dell'annotazione è valutata attraverso una strategia sviluppata per tenere conto della necessità di allineare intere costruzioni identificate da annotatori diversi. Il corpus e gli strumenti per la valutazione dell'annotazione sono resi disponibili.*

1 Introduction and Background

Modality is a pervasive phenomenon crucial to language understanding, analysis, and automatic processing (Morante and Sporleder, 2012). The creation of modality-annotated data would benefit Natural Language Processing in at least two major aspects: (i) factuality detection, consisting in the automatic distinction between propositions that represent factual events and propositions that represent non factual ones; and (ii) sentiment analysis, which involve the processing of extra-propositional aspects of meaning and the detection of polarised judgements. Additionally, the annotation of modality may also have important repercussions in the field of corpus linguistics, as the techniques developed in the automatic treatment of modality can be used to improve our linguistic knowledge of modality itself.

As far as the detection of polarised judgments goes, there have been substantial annotation efforts in recent years, exemplified by recurring and increasing sentiment analysis tasks within the context of the Semeval evaluation campaign.[1] Attention has also been given to more specific factuality tasks such as the CoNLL-2010 Shared Task on identifying hedges (Farkas et al., 2010), and factuality annotation in languages other than English, such as Italian (Minard et al., 2014), and Dutch (Schoen et al., 2014). However, these are annotation efforts involving specific phenomena rather than modality in general.

Indeed, a major bottleneck in the creation of modality-annotated resources is the very notion of modality itself, as encapsulating this phenomenon in one exhaustive but workable definition is far from trivial (Morante and Sporleder, 2012). Building on the function-based proposal advanced in (Nissim et al., 2013) and (Ghia et al., 2016), we have created a comprehensive annotation scheme for epistemic modality and have applied it to multiple languages. Contextually, we have developed and deployed an evaluation strategy which shows that the corpus is annotated reliably.

Summary of contributions We produced the first multilingual corpus annotated for modality. The annotation scheme is virtually language-independent, and the annotation is evaluated according to a specifically designed methodology

[1] `http://alt.qcri.org/semeval2017/index.php?id=tasks`. Note that in 2017 within the sentiment analysis track there was also a task on truth detection, which goes to show how closely related the two phenomena indeed are.

which is portable to other tasks where annotators are left with substantial freedom in the selection of the tokens to be marked up. The corpus and the tools for scoring agreement are publicly available (`http://modal.msh-vdl.fr/`,`https://bitbucket.org/lennyklb/modality/`).

2 Corpus

The MODAL Corpus is the first corpus of dialogues in multiple languages annotated for phenomena of (epistemic) modality.

MODAL consists of three equivalent resources of English, French and Italian dialogues. These were drawn from the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000) for English, from the ESLO Corpus (Baude and Kanaan, 2014), plus the OTG Corpus and the Accueil UBS Corpus (Antoine et al., 2002) for French, and from the VoLip Corpus (Alfano et al., 2014) for Italian. All data is marked for epistemic modality and amounts to approximately 20.000 words per language for a total of 2824 epistemic constructions (833 for the English Corpus, 1271 for the French Corpus, 720 for the Italian Corpus).

2.1 Approach to annotation

In the construction of MODAL, we were guided by two main principles: *maximum expressivity*, and *cross-lingual validity*. We therefore took an approach to annotation that would simultaneously ensure both.

Specifically, we did not want to annotate a predetermined list of epistemic constructions and assign functions to them. Indeed, this would make the scheme very much language-dependent, as specific tokens/constructions would need to be identified for each language. Additionally, it would restrict the annotation to this pre-selection, which could not be exhaustive.

As an alternative approach, we provided a theoretical meaningful, and operationalisable definition of epistemic modality. On this ground, thus only at a later stage, the annotators identified the linguistic constructions that realise epistemic modality in the three different languages. Thus, rather than going from constructions to functions, we go from functions to constructions.

While this approach has the advantage of being valid cross-linguistically, and maximising expressivity, it also potentially has a major problem. Letting the annotators choose freely the tokens and the constructions to be annotated without controlling for any pre-selection, incurs the risk of a wide range of choices, and substantially low agreement. We discuss this in the Evaluation section. In the remainder of this section we explain the scheme and the procedure we used to annotate the corpus.

Table 1: Annotation categories for the marker

LEMMA	< *lemma* >

ILLOCUTION	`assertion`
	`exclamation`
	`injunction`
	`question`

MORPHOSYNTAX	`morph-conditional`
	`morph-preterite`
	`morph-future`
	`lex-complement-taking-pred`
	`lex-adverb`
	`lex-disc-marker`
	`lex-modal-verb`
	`syn-dependent`
	`syn-list`
	`syn-tag`
	`disc-utterance`
	`prosody-interrogative`

Table 2: Annotation categories for the Relation

DIRECTION	`scope-marker`
	`marker-scope`
	`inside`
	`co-extensive`	

EPISTEMIC TYPE	`direct-auditory`
	`direct-visual`
	`direct-feeling`
	`indirect-infer`
	`indirect-report`
	`quotative`
	`memory`
	`no evidence`	

POLARITY	`positive`
	`neutral`
	`negative`	

DISCOURSE FUNCTION	`qualification`	

	`negotiation`	`acceptation`
		`non acceptation`
		`check`
		`information`

2.2 Procedure and final scheme

We employed a two-fold procedure: epistemic constructions are first identified, and then annotated with their features.

Identification of epistemic constructions In order to annotate epistemic modality in *dialogues*, we subscribed to a communitarian (Stalnaker, 1978), dynamic (Groenendijk and Stokhof, 1991), and interactionist (Ginzburg, 2012) approach to semantics, which led us to refine the traditional definition of epistemic modality. Specifically, we put forward the idea that any construction that explicitly signals the process of shared attribution of a truth value to the propositional tokens that compose a discourse should be considered as an epistemic construction, and thus annotated.

Consequently, we annotated not only constructions in which a marker is realized by a more grammaticalized element, such as a modal verb (Example 1), but also constructions in which a marker is realized lexically (Example 2) or prosodically (Example 3):

(1) A penguin might lay two eggs and at that point [. . .]

(2) And I do believe it was thirty days [. . .]

(3) DON: Oh specifically in the islands?

Besides, we annotated both monological epistemic constructions in which a marker expresses the evaluation of the truth-value of a scope by a single speaker (Example 4), and dialogical epistemic constructions in which two or more markers are used to negotiate the evaluation of the truth-value of one and the same scope among the participants in a conversation (Example 5):

(4) apparently it was very very muddy it was abnormally warm and it was just a big mudbath out there [. . .]

(5) ALIC: I don't think Darren put anything on it .
NICO: Mhm .
ALIC: Right .
ALIC: Okay .

Annotation of epistemic constructions We represented the epistemic constructions identified in the corpus as triadic constructions consisting of a marker, a scope and a relation between the marker and the scope, as shown in Figure 1.

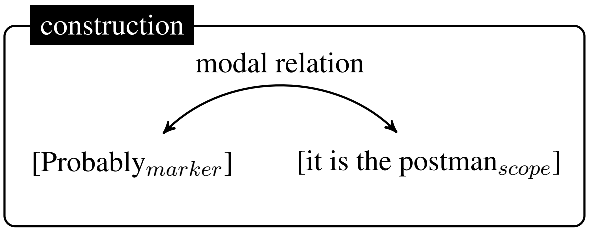


Figure 1: A construction is conceived as a marker, a scope, and a modal relation between them.

We formalised the marker, the scope and the relation between them as three elements each endowed with its own formal and functional properties. Each element is then annotated with syntactic, semantic, and pragmatic features according to the developed annotation scheme.

Building on (Nissim et al., 2013; Ghia et al., 2016), we devise a fully-fledged annotation scheme that is functionally motivated and cross-linguistically valid. Annotation features are specified for all three elements of the modalised construction, namely the marker, the scope, and the relation. The features for the markers and the relations are shown in Tables 1 and 2, respectively. For the scope, we use a property `syntax` and `clause`/`utterance` as features.

Operationalising the annotation task From a theoretical perspective, the scheme is grounded in the Construction Grammar framework (Goldberg, 1995). In practice, the annotators could work with the labels from the annotation schemes, but also with decision trees that guided the process of identification of epistemic constructions as well as feature assignment. The annotation was performed using the Analec annotation tool (Landragin et al., 2012), which produces TEI-compliant XML output. Analec was originally designed for the annotation of anaphoric phenomena and thus lends itself well to the task of annotating a three-way construction, with features for marker, scope, and relation. All data was annotated by three teams of 2 or more annotators (a, b for Italian, a, b, c, d for English, a, e,f for French) and agreement was assessed via a specifically developed evaluation strategy (Section 3).[2]

3 Evaluation

The originality of the general approach and of the annotation procedure led us to develop an origi-

[2] Further information regarding the distribution of categories and examples is available at the project's website (`http://modal.msh-vdl.fr/` and in (Pietrandrea, forthcoming)).

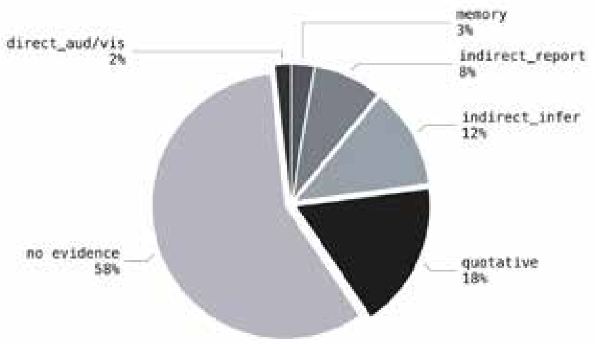


Figure 2: Distribution of EPISTEMIC TYPES for the Relation annotation in the Italian portion of the corpus (see Table 2).

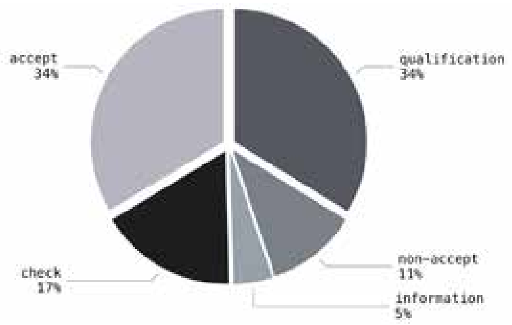


Figure 3: Distribution of DISCOURSE FUNCTIONS for the Relation annotation in the Italian portion of the corpus (see Table 2).

nal technique for testing the inter annotator agreement, essentially based on the percentage of overlap between the spans of text identified as markers or scopes by the annotators (Ghia et al., 2016). In order to assess this, annotations must be *aligned*. We describe how we align the constructions in practice, how we use alignment information in order to assess agreement, and discuss results.

3.1 Alignment and Agreement

Annotators can identify *any* textual element as part of a modalised construction, and each annotator works on their own file. This means that in order to assess agreement, we first need to try to *align* the constructions marked up in the two files. We do so via *anchors*. Anchors can be aligned iff:

- they are of the same type (marker or scope)
- they overlap in content by at least a given proportion of lexical material, which we base on character offset. For example, for a required overlap of 50% and a token length of an anchor A of ten tokens, the content of the candidate anchor from the other file needs to have at least five subsequent words in common with A.

This process results in a collection of pairs of aligned anchors. For example, considering annotator a and annotator b, we would have an aligned pair of marker t_a and marker t_b.

The final step is to iterate through the relations that judge a introduced and align them with relations that judge b introduced. In order to explain the procedure of further alignment to relations, we take judge a as reference, but in terms of scores it doesn't make any difference which direction we go, since $precision_{ab} = recall_{ba}$ so that eventually $fscore_{ab} = fscore_{ba}$. Relations consist of a marker and one or multiple scope portions. Aligning relations is done by pairing up markers and scopes into relations introduced by judge a and check if the aligned counterparts of these markers and scopes by judge b are part of a relation as well. When this is the case, we deem the two constructions as "the same".

Next, we have to assess agreement on the features assigned to relations and markers. While agreement over alignment is measured using precision/recall/f-score as we have to deal with potentially different spans, for the relations' and markers' features, we can then use Cohen's Kappa (Cohen, 1960) over the agreed upon constructions only, as it becomes a plain classification task.

3.2 Results

Because of freedom in the annotation of the extension of anchors, as mentioned above we evaluated alignment at different percentages of overlap. The scores for the alignment of scopes for all three languages is shown in Table 3. While for Italian we observe that even when evaluating alignment of full strings (i.e. requiring 100% overlap), the agreement stays high, this is not the case for French and English. Indeed, if complete overlap of scopes is required to deem the annotations equivalent, F-scores drop quite a bit. We do not include a table for the scores on the markers as they do not change substantially with varying degrees of overlap. This is due to the fact that markers are often just single words, or very short anyway. F-scores range from 0.91 at 10% and 0.90 at 100% for English, from 0.86 at 10% and 0.85 at 100% for French, and stay stable at 0.94 for Italian. For this reason, we can be lenient with markers' align-

Table 3: Agreement for scope identification.

Overlap	FRENCH			ITALIAN			ENGLISH		
	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
10%	0.86	0.92	0.89	0.95	0.93	0.94	0.88	0.85	0.86
20%	0.85	0.92	0.88	0.95	0.93	0.94	0.88	0.85	0.86
30%	0.84	0.92	0.88	0.95	0.93	0.94	0.88	0.85	0.86
40%	0.82	0.92	0.87	0.95	0.93	0.94	0.88	0.85	0.86
50%	0.82	0.92	0.87	0.95	0.93	0.94	0.88	0.85	0.86
60%	0.79	0.91	0.85	0.95	0.93	0.94	0.88	0.85	0.86
70%	0.78	0.9 0	0.84	0.95	0.93	0.94	0.87	0.84	0.85
80%	0.77	0.9 0	0.83	0.95	0.93	0.94	0.87	0.83	0.85
90%	0.75	0.89	0.81	0.94	0.93	0.93	0.85	0.81	0.83
100%	0.72	0.87	0.79	0.94	0.93	0.93	0.81	0.78	0.79

Table 4: Kappa scores for marker's features.

Overlap	FRENCH	ITALIAN	ENGLISH
10%	0.82	0.91	0.85
20%	0.82	0.91	0.86
30%	0.84	0.91	0.87
40%	0.84	0.91	0.88
50%	0.84	0.91	0.88
60%	0.84	0.91	0.88
70%	0.84	0.92	0.87
80%	0.84	0.92	0.87
90%	0.89	0.92	0.86
100%	0.89	0.92	0.87

ment, which we set at 10% when evaluating relations. Interestingly, though, we can observe that when evaluating agreement on the *features*, Kappa increases when stricter alignment is required (Table 4). This is likely due to the fact that on fully agreed upon strings, the assigned features are also agreed upon.

For the Italian annotation of the relation's features, at overlap 100%, we observe $K = 0.86$ for the FUNCTION feature, $K = 0.82$ for TYPE, and $K = 0.72$ for POLARITY.[3]

To provide a more detailed view into the disagreements of the `type` feature, for instance (whose final agreed upon distribution was reported in Figure 2 above), in Figure 4 we show the confusion matrix for Italian. We can observe that the largest number of confusions arise from mixing up the categories `indirect_inferential` and `no_evidence`. Indeed, the precise delimitation between these two categories is a long- and hot-debated issue in the literature on epistemic modality (Pietrandrea, 2005, among others).

Overall, we can see that our annotation, albeit granting the annotators a lot of freedom, is substantially reliable.[4]

	dir_aud	dir_vis	ind_inf	ind_rep	mem	no_ev	quot
dir_aud	**1**	-	-	-	-	-	-
dir_vis	-	**9**	-	-	-	-	-
ind_inf	-	-	**40**	2	-	21	1
ind_rep	-	-	-	**46**	-	1	1
mem	-	-	2	-	**12**	1	-
no_ev	-	1	12	3	1	**316**	3
quot	-	-	-	7	-	7	**90**

Figure 4: Confusion matrix for the annotation of the TYPE feature in Italian.

4 Conclusions

Modality can be reliably annotated in multiple languages by taking a bottom-up, functional approach paired with a solid annotation scheme, trees to guide the annotators' decisions, and a rigorous evaluation strategy. With this approach, we have produced the first multilingual corpus annotated for modality, which can be potentially used to train modality detection models as well as to further study modality itself. By making all of the data publicly available, and by sharing our annotation experience, we also hope to provide a blueprint for creating modality-annotated resources in yet more languages.

Acknowledgments

The research described in this paper was supported by the Maison des Sciences de l'Homme du Val de Loire (grant MoDAL 2015), the IRCOM Consortium, the "Laboratoire Ligérien de Linguistique – UMR7270", and the Groningen Meaning Bank. The authors are also grateful to the anonymous reviewers for their comments.

References

Iolanda Alfano, Francesco Cutugno, Aurelio De Rosa, Claudio Iacobini, Renata Savy, and Miriam Voghera. 2014. Volip: a corpus of spoken italian and a virtuous example of reuse of linguistic resources. In Nicoletta Calzolari et al., editor, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

[3] Very similar scores are observed at different degrees of overlap.

[4] Please note that for all agreement results, for all languages, the reader is referred to the project's website, where examples of disagreement are also included (`http://modal.msh-vdl.fr/`).

J.-Y. Antoine, S. Letellier-Zarshenas, P. Nicolas, and I. Schadle. 2002. Corpus OTG et ECOLE_MASSY : vers la constitution dun collection de corpus francophones de dialogue oral diffusés librement. In *Actes TALN2002*, pages 319–324, Nancy, France.

Olivier Baude and Layal Kanaan. 2014. Le corpus des ESLO à l'ère des Digital Humanities. In *Premières Rencontres FLORAL*, Paris, France, December.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000. Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia: Linguistic Data Consortium.*

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*, CoNLL '10: Shared Task, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

E. Ghia, L. Kloppenburg, M. Nissim, P. Pietrandrea, and V. Cervoni. 2016. A construction-centered approach to the annotation of modality. In *Proceedings of the 12th Workshop on Interoperable Semantic Annotation (ISA-12)*, Portoroz, Slovenia.

Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and philosophy*, 14(1):39–100.

Frederic Landragin, Thierry Poibeau, and Bernard Victorri. 2012. Analec: a new tool for the dynamic annotation of textual data. In *International Conference on Language Resources and Evaluation (LREC 2012)*, pages 357–362.

Anne-Lyse Minard, Alessandro Marchetti, and Manuela Speranza. 2014. Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. In *Proceedings of CLIC-it*, Pisa.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2).

Malvina Nissim, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14, Potsdam, Germany, March. ACL.

Paola Pietrandrea. 2005. *Epistemic modality: functional properties and the Italian system*, volume 74. John Benjamins Publishing.

Paola Pietrandrea. forthcoming. Epistemic constructions at work: A corpus-driven study on spoken Italian dialogues. *Journal of Pragmatics*.

Anneleen Schoen, Chantal van Son, Marieke van Erp, and Hennie van der Vliet. 2014. Newsreader document-level annotation guidelines-dutch. Technical report, Vrije Universiteit Amsterdam, TechReport 2014-8.

Robert C Stalnaker. 1978. *Assertion*. Wiley Online Library.

AHyDA: Automatic Hypernym Detection with feature Augmentation

Ludovica Pannitto
University of Pisa
ellepannitto@gmail.com

Lavinia Salicchi
University of Pisa
lavinia.salicchi@libero.it

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Abstract

English. Several unsupervised methods for hypernym detection have been investigated in distributional semantics. Here we present a new approach based on a smoothed version of the distributional inclusion hypothesis. The new method is able to improve hypernym detection after testing on the BLESS dataset.

Italiano. *Sulla base dei metodi non supervisionati presenti in letteratura, affrontiamo il task di riconoscimento di iperonimi nello spazio distribuzionale. Introduciamo una nuova misura direzionale, basata su un'espansione dell'ipotesi di inclusione distribuzionale, che migliora il riconoscimento degli iperonimi, testandola sul dataset* BLESS.

1 Introduction and related works

Within the Distributional Semantics framework, semantic similarity between words is usually expressed in terms of proximity in a semantic space, where the dimensions of the space represent, at some level of abstraction, the contexts in which the words occur.

Our intuitions about the meaning of words allow inferences of the kind expressed in example (1), and we expect Distributional Semantic Models (DSMs) to support such inferences:

(1) a. Wilbrand *invented* TNT → Wilbrand *uncovered* TNT
b. A *horse ran* → An *animal moved*

The type of relation between semantically similar lexemes may differ significantly, but DSMs only account for a generic notion of semantic relatedness. Furthermore, not all lexical relations are symmetrical (see example (2)), while most of the similarity measures defined in distributional semantics are, like the cosine.

(2) a. I saw a *dog* → I saw an *animal*
b. I saw an *animal* ↛ I saw a *dog*

Hypernymy is an asymmetric relation. Automatic hypernym identification is a very well-known task in literature, which has mostly been addressed with semi-supervised, pattern-based approaches (Hearst, 1992; Pantel and Pennacchiotti, 2006). Various unsupervised models have been proposed (Weeds and Weir, 2003; Weeds et al., 2004; Clarke, 2009; Lenci and Benotto, 2012; Santus et al., 2014), based on the notion of **Distributional Generality** (Weeds et al., 2004) and on the **Distributional Inclusion Hypothesis** (DIH) (Geffet and Dagan, 2005) which has been derived from it.

1.1 The pitfalls of the DIH

The DIH aims at providing a distributional correlate of the extensional definition of hyponymy in terms of set inclusion: x is a hyponym of y iff the extension of x (i.e. the set of entities denoted by x) is a subset of the extension of y. The DIH turns this into the assumption that a significant number of the most salient contexts of x should also appear among the salient contexts of y. While this is consistent with the logical inferences licensed by hyponymy (cf. (2)), it does not take into account the actual usage of hypernyms with respect to hyponyms. Consider for instance the following examples:

(3) a. A *horse* gallops $\overset{?}{\rightarrow}$ An *animal* gallops
b. A *dog* barks $\overset{?}{\rightarrow}$ An *animal* barks

These inferences are truth-conditionally valid: whenever the antecedent is true, the consequent is also true. However, they are not equally "pragmatically" sound. In fact, the fact that one uses

	horse	*dog*	*animal*
gallop	216	–	7
bark	–	869	16

Table 1: Co-occurrence frequency distribution extracted from the ukWaC corpus

a sentence like *A dog barks* does not entail that in the same situation one would have also used the sentence *An animal barks*. The latter sentence would be pragmatically appropriate only in cases in which one knows that something is barking, without knowing which animal is producing this sound. However, the latter condition hardly applies, since barking is a very typical feature of dogs: knowing that something is barking typically entails knowing that it is a dog, since we know that barking is something dogs do. The same argument also applies to the case of *horse* and *galloping*.

The problem of the DIH is that the assumption it rests on, namely that the most typical contexts of the hyponym are also typical contexts of the hypernym, is not borne out in practical language usage because of pragmatic constraints. The most typical contexts of an hyponym are not necessarily the typical contexts of its hypernym. This is also proved by a simple inspection of corpus data, as reported in Table 1. Despite *animal* $(161, 107)$ is more frequent than *dog* $(128, 765)$ and *horse* $(90, 437)$, its co-occurrence with *bark* and *gallop* is much lower than the ones of the hyponyms: *bark* and *gallop* are not typical contexts of *animal*.

If the inferences in (3) are pragmatically odd, the following ones are instead fully acceptable:

(4) a. A *horse* gallops → An *animal* moves
b. A *dog* barks → An *animal* calls

Salient features of the *hypernym* are indeed supposed to be semantically more general than the salient features of the *hyponym*. Santus et al. (2014) tried to capture this fact by abandoning the DIH and introducing an entropy-based measure to estimate of informativeness of the hypernym and hyponym contexts, under the assumption that the former have a higher entropy, because they are more general (e.g. *move* vs. *gallop*).

In this paper, we address the same issue by amending the DIH, to make it more consistent with the actual distributional properties of hyponyms and hypernyms. Therefore, we introduce **AHyDA** (Automatic Hypernym Detection with feature Augmentation), a smoothed version of the DIH: given a context feature f that is salient for a lexical item x, we expect *co-hyponyms* of x to have some feature g that is similar to f, and an *hypernym* of x to have a number of these clusters of features. To remain in the animal sounds area, we expect a *dog* to *bark* and a *duck* to *quack* and an *animal* to produce either of those sounds or to co-occur with a more general sound-emission verb.

2 AHyDA: Smoothing the DIH

All the measures implementing the DIH are based on computing the (weighted) intersection of the features of the hyponym and the hypernym. This is then typically divided by the hyponym features. AHyDA essentially proposes a new way to compute the intersection of the hyponym and hypernym contexts. Given a lexical item x, we call F_x the set of its distributional features. Note that features need not be pure lexical items. In general, we define f as a pair (f_w, f_r) where f_w is typically a lexical item, and f_r is any additional contextual information, in the present case a pattern occurring between x and f_w, as explained in section 3.1. The core novelty of AHyDA is to use a smoothed version of F_x, called F'_x.

The idea is shown in figure 1, which provides a simplified graphical example of the intersection operation. Consider a case where the target *horse* has some feature with *gallop* as a lexical item, for example a feature $f = (gallop, sbj)$ meaning that *horse* is a possible subject of *gallop*. Given what we have said in Section 1.1, we do not expect *animal* to share this *horse*-specific property. So, instead of looking for this particular feature among the ones of *animal*, we generate a new set $N_{horse}(gallop)$ of features $g = (g_w, f_r)$ such that g_w is a neighbor of *gallop* and is a feature (with the same syntactic relation *sbj*) of some neighbor of *horse*. Suppose that *run*, *move*, and *cycle* are neighbors of *gallop*. As *run* and *move* are also features of some neighbor of *horse* (e.g., *lion*), we would have $N_{horse}(gallop) = \{gallop, run, move\}$. Conversely, since *cycle* is not a feature of a close neighbor of *horse*, it would not be included in the expanded feature set.

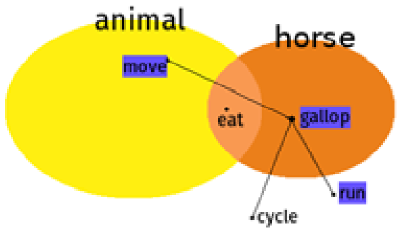


Figure 1: An example of smoothed intersection. Black arrows indicate semantic similarity with *gallop*, items with the blue background are the ones included in $N_{horse}(gallop)$.

Mathematically, we define the expanded feature set F'_x as follows:

$$F'_x = \{(f, N_x(f)) \, \forall f \in F_x\} \tag{1}$$

$$N_x(f) = \{g | g = (g_w, f_r)\} \tag{2}$$

where the following conditions hold for g:

$$d(f_w, g_w) < k \wedge \exists y | d(x, y) < h \wedge g \in F_y \tag{3}$$

where $d(x, y)$ is any distance measure in the semantic space, k and h are empirically set threshold values.

$N_x(f)$ is generated by looking for features g that are similar to f_w, We then check whether this new feature is shared by some neighbor of the target x, and eventually include g in $N_x(f)$. This allows us to redefine the intersection operation between F'_x and F_y as:

$$F_x{}'\hat{\cap}F_y = \{f | f \in F_x \wedge N_x(f) \cap F_y \neq \emptyset\} \tag{4}$$

When expanding a feature f into $N_x(f)$, we expect to find in $N_x(f)$ features that express the same "property" in different ways. We expect these features to be shared by hypernyms more than co-hyponyms, because hypernyms are supposed to collect features from all their hyponyms, while co-hyponyms lack those of other co-hyponyms (e.g. lions *run* but do not *gallop*).

AHyDA is thus defined as follows:

$$AHyDA(x, y) = \frac{\sum_{f \in F_x} |F'_x \cap F_y|}{|F_x|} \tag{5}$$

Importantly, AHyDA only considers the average cardinality of the intersections, without looking at the feature weights. Moreover, the formula is asymmetric (like the others implementing the DIH), and therefore it is suitable to capture the asymmetric nature of hypernymy.

3 Experiments and Evaluation

3.1 Distributional Space

Each lexical item u is represented with distributional features extracted from the *TypeDM* tensor (Baroni and Lenci, 2010). In TypeDM, distributional co-occurrences are represented as a *weighted tuple structure*, a set of $((u, l, v), \sigma)$, such that u and v are lexical items, l is a syntagmatic co-occurrence link between u and v and σ is the *Local Mutual Information* (Evert, 2005) computed on link type frequency. Hence, each lexical item u is represented in terms of features of the kind (l, v).

In addition to the sparse space, we also produced a dense space of 300 dimensions reducing the matrix with Singular Value Decomposition (SVD). This additional space was used to retrieve neighbors during the smoothing operation, as it allowed us to perform faster and more accurate calculations for cosines. The sparse space was instead employed to retrieve features and get their weights.

3.2 Data set

Evaluation was carried on a subset of the BLESS dataset (Baroni and Lenci, 2011), consisting of tuples expressing a relation between nouns.

BLESS includes 200 English concrete nouns as target concepts, equally divided between living and non-living entities. For each concept noun, BLESS includes several relatum words, linked to the concept by one of the following 5 relations: COORD (i.e. co-hyponyms), HYPER (i.e. hypernyms), MERO (i.e. meronyms), ATTRI (i.e. attributes), EVENT (i.e. verbs that define events related to the target). BLESS also includes the relations RANDOM-N, RANDOM-J, RANDOM-V, which relate the targets to control tuples with random noun, adjective and verb relata, respectively.

By restricting to *noun-noun* tuples, we got a subset containing these relations: COORD, HYPER, MERO, RANDOM-N. We preprocessed the dataset in order to exclude lexical items that are not included in TypeDM. As reported in table 2, the distribution (minimum, mean and maximum) of the relata of all BLESS concepts is not even, and therefore we took this into account while

relation	*min*	*avg*	*max*
coord	6	17.1	35
hyper	2	6.7	15
mero	2	14.7	53
ran-n	16	32.9	67

Table 2: Distribution (minimum, mean and maximum) of the relata of all BLESS concepts

evaluating our results.

3.3 Evaluation

We compared AHyDA with a number of directional similarity measures tested on BLESS, with the goal of evaluating their ability to discriminate hypernyms from other semantic relations, in particular co-hyponyms. Given a lexical item x, F_x is the set of its distributional features, $w_x(f)$ is the weight of the feature f for the term x:

WeedsPrec - quantifies the weighted inclusion of the features of a term x within the features of a term y (Weeds and Weir, 2003; Weeds et al., 2004; Kotlerman et al., 2010)

$$\mathit{WeedsPrec}(x,y) = \frac{\sum_{f \in F_x \cap F_y} w_x(f)}{\sum_{f \in F_x} w_x(f)} \quad (6)$$

ClarkeDE - a variation of *WeedsPrec*, proposed in (Clarke, 2009)

$$\mathit{ClarkeDE}(x,y) = \frac{\sum_{f \in F_x \cap F_y} min(w_x(f), w_y(f))}{\sum_{f \in F_x} w_x(f)} \quad (7)$$

invCL - a new measure introduced in (Lenci and Benotto, 2012), to take into account not only the inclusion of x in y but also the non-inclusion of y in x. The measure is defined as a function of *ClarkeDE* (CD).

$$\mathit{invCL}(x,y) = \sqrt{\mathit{CD}(x,y)(1 - \mathit{CD}(x,y))} \quad (8)$$

We used the **cosine** as a baseline, since it is a symmetric similarity measure and is commonly used to evaluate semantic similarity/relatedness in DSMs. In the definition of $N_x(f)$, the target and feature neighbors are identified with the cosine, setting the k and h parameters to 0.8 and 0.9 respectively.

To avoid biases due to the relata distribution among concepts, for each target x, we computed the *minimum* and *maximum* number of items holding a relation with x, and performed $\frac{maximum}{minimum}$ random samples where each relation is presented with *minimum* relata, and then averaged the results.
For example, consider the situation where x has 3 hypernyms, 6 co-hyponyms, 6 meronyms and 12 random nouns. In this situation, the *minimum* number of relata for x would be 3, while the *maximum* would be 12. Therefore, we would perform 4 random sampling for each relation, averaging the results in order to obtain a singular measurement for each relation in the end.

We adopted the same evaluation methods described in Lenci and Benotto (2012): plotting the distribution of scores per relation across the BLESS concepts, and calculating Average Precision (AP).

3.4 Results

Table 3 summarizes the Average Precision obtained by AHyDA, the other DIH-based measures, and the cosine. Although AHyDA's improvement is not big in hypernym detection, *co-hyponyms* get lower values of AP, thus showing that smoothing the intersection allows a better discrimination between the two classes. It is worth remarking that the values for the other measures are generally higher than those reported by Lenci and Benotto (2012), because of the evaluation on the balanced random samples of relations we have adopted. We also reported, in table 4, the AP values obtained through the standard measures, without employing the feature augementation procedure. Altough values for hypernyms do not change much, the main differences are in the *coord* values, which are generally higher without feature augmentation. As mentioned in section 3.1, the results for all the measures are obtained using the sparse space. The reduced space was employed to compute the *Cosine* baseline.

As regards the AP values for hypernyms, we must notice that not all hypernyms in BLESS share the same status: some of them are what we would consider logic entailments (e.g. *eagle* → *bird*), others depict taxonomic relations (e.g. *alligator* → *chordate*), some are not true logic entailments (e.g. *hawk* $\overset{?}{\rightarrow}$ *predator*)

Figure 2 shows the average score produced with the new measure. Here *hypernyms* are neatly set apart from *co-hyponyms*, whereas the distance with *meronyms* and with the control group, *randoms*, is less significative.

measure	*coord*	*hyper*	*mero*	*ran-n*
Cosine	0.77	0.31	0.21	0.14
WeedsPrec	0.29	0.50	0.32	0.16
ClarkeDE	0.31	0.52	0.24	0.14
invCL	**0.28**	**0.52**	0.32	0.17
AHyDA	**0.20**	**0.49**	0.33	0.23

Table 3: Mean AP values for each semantic relation achieved by AHyDA and the other similarity scores

measure	*coord*	*hyper*	*mero*	*ran-n*
Cosine	0.77	0.32	0.21	0.14
WeedsPrec	0.34	0.51	0.28	0.15
ClarkeDE	0.36	0.51	0.27	0.16
invCL	**0.31**	**0.51**	0.29	0.16

Table 4: Mean AP values for each semantic relation achieved by the cited similarity scores, without employing feature augmentation

Figure 3 shows the average scores produced by AHyDA when applied to the reverse hypernym pair. It is interesting to notice that in this case AHyDA produces basically the same results as random pairs. This suggests that AHYDA correctly predicts that hyponyms entail hypernyms, but not vice versa, thereby capturing the asymmetric nature of hypernymy.

4 Conclusion

The Distributional inclusion hypothesis has proven to be a viable approach to hypernym detection. However, its original formulation rests on an assumption that does not take into consideration the actual usage of hypernyms in texts. In this paper we have shown that, by adding some further pragmatically inspired constraints, a better discrimination can be achieved between co-hyponyms and hypernyms. Our ongoing work focuses on refining the way in which the smoothing is performed, and testing its performance on other datasets of semantic relations.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Pro-*

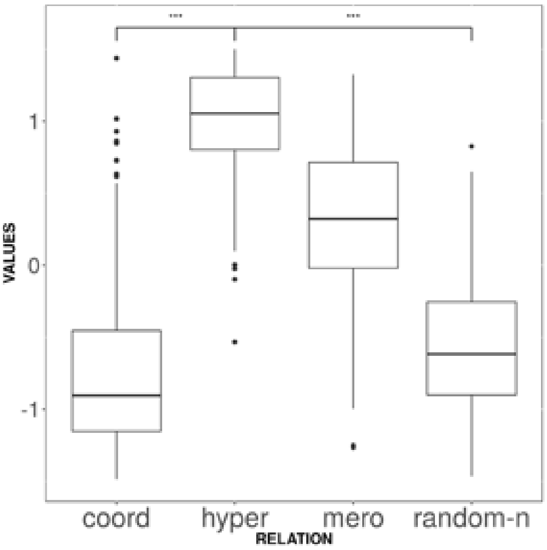


Figure 2: Distribution of relata similarity scores obtained with AHyDA (values are concept-by-concept z-normalized scores)

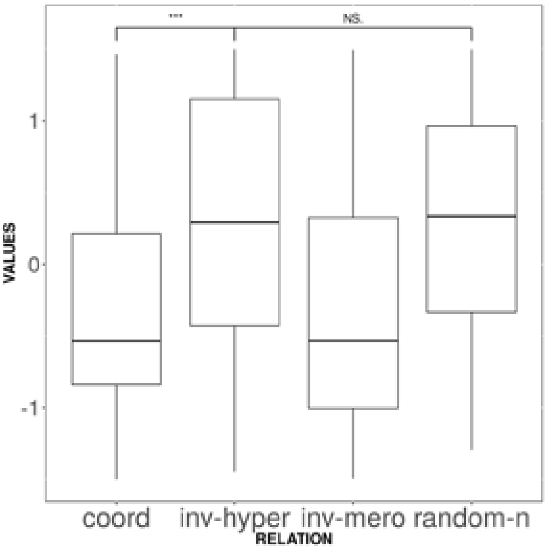


Figure 3: Distribution of relata similarity scores obtained with AHyDA (values are concept-by-concept z-normalized scores), when tested on the inverse inclusion (i.e. *hypernym* does not entail *hyponym*)

ceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, pages 1–10. Association for Computational Linguistics.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the workshop on geometrical models of natural language semantics*, pages 112–119. Association for Computational Linguistics.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114. Association for Computational Linguistics.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79. Association for Computational Linguistics.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte Im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *EACL*, pages 38–42.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 81–88. Association for Computational Linguistics.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1015. Association for Computational Linguistics.

INFORMed PA: A NER for the Italian Public Administration Domain

Lucia C. Passaro***, Alessandro Lenci*****, Anna Gabbolini****
* Dipartimento di Filologia, Letteratura e Linguistica, University of Pisa (Italy)
** ETI^3 | Evolution, Technology & Innovation
lucia.passaro@for.unipi.it
alessandro.lenci@unipi.it
anna.gabbolini@eti3.it

Abstract

English. In this paper, we illustrate the creation of a NER for the Public Administration (PA) domain. We discuss the creation of an annotated corpus with documents from the Italian *Albo Pretorio Nazionale* and provide results of the system evaluation.

Italiano. *In questo lavoro mostriamo la creazione di un NER per il dominio della Pubblica Amministrazione (PA). Presentiamo la creazione del corpus formato da documenti dell'Albo Pretorio Nazionale e mostriamo i risultati della valutazione del sistema.*

1 Introduction

In the Public Administration (PA) domain, the rapid adoption of the new legislation about the governance transparency has been forcing Italian municipalities to produce their acts in a digital form and to make them available for both citizens and authorities. However, the acts delivered by PAs are typically in a free-text electronic format, which is not convenient for searching, decision-support, and data analysis. Therefore, the development of NLP tools to extract high-quality structured information, including Named Entities (NEs) such as Persons and Organizations, represents a key factor to enable the access to the wealth of information produced by PAs, and a crucial step in turning the keyword of "transparency" into reality. The potentialities of NLP tools can be exploited to mine the large document repositories produced by PA daily, with the aim of identifying trends in their activity, suggesting possible synergies to increase their efficiency, and raising "red flags" about suspicious behaviors, especially for their relationships with private companies.

In this paper, we focus on Named Entity Recognition (NER) for PA. Several approaches have been proposed in literature including Rule-based, Machine Learning-based and Hybrid methods.

Hand-made Rule-based NERs focus on extracting names using lots of human-made rules. In general, these systems consist of a set of patterns based on grammatical (e.g., part of speech), syntactic (e.g., word precedence) and orthographic features (e.g., capitalization) in combination with dictionaries (Budi and Bressan, 2003; Appelt et al., 1993; Grishman, 1995). These approaches usually give good results, but require long development time by expert linguists. On the one hand, these systems have better results for restricted domains, being capable of detecting very complex entities, but, on the other one, they lack portability and robustness and do not necessarily adapt well to new domains and languages.

Machine learning techniques, on the contrary, use a collection of annotated documents for training the classifiers. Therefore the development time moves from the definition of rules to the preparation of annotated corpora (Bikel et al., 1997; Borthwick et al., 1998; McCallum and Li, 2003). The systems identify and classify nouns using machine learning algorithms such as Maximum Entropy (Berger et al., 1996), Support Vector Machines (Cortes and Vapnik, 1995) and Conditional Random Field (Lafferty et al., 2001). More recently, also deep learning architectures have been proposed for Named Entity Recognition (Chiu and Nichols, 2015; Strubell et al., 2017).

Finally, Hybrid NER systems, combine rule-based and machine learning-based methods, and make new methods using strongest points from each method (Srihari et al., 2000).

Existing general purpose Italian corpora annotated with NEs such as I-CAB (Magnini et al., 2006) are not optimal for training a NER for the domain of PA because of the gap between bureaucratic language and standard Italian, and also because of the lack of important classes such as act and normative references, that are very useful in PA-oriented applications. To tackle these problems, we decided to create a new corpus from scratch starting from: (i) administrative documents belonging to the *Italian Albo Pretorio*; (ii) the CoLingLab NER, a general NER trained on I-CAB, from which we took the initial configuration of features. The corpus of PA documents written in Italian "bureaucratese", has the characteristics described in Brunato (2015):

1. Pseudo-technicisms or collateral technicisms (e.g., *balneazione, fattispecie*);
2. Abstract nouns with *-zione/-mento* suffixes (e.g., *stipulazione, espletamento*), deverbal nouns, usually with zero suffix (e.g., *subentro, scorporo, utilizzo*) and denominal verbs (e.g., *relazionare, disdettare*);
3. Archaic terms (e.g., *allorché, suddetto*) and latinisms (e.g. *una tantum, pro capite*);
4. Forestierisms (e.g., governance, front office);
5. Uncommon and formal terms (e.g., *diniego* for *rifiuto*);
6. Stereotyped phrases (e.g., *entro e non oltre, in riferimento all'oggetto*);
7. Abbreviations and acronyms.

For the creation of a NER for PA, we decided to exploit the existing architecture employed for the project SEMPLICE[1] and in particular we adopted a statistical method based on the Stanford NER (Finkel et al., 2005), a system implemented in Java and available for download under the GNU General Public License. This choice allowed us to easily compare the gain obtained by enriching the training corpus with PA documents and to speed up the development process. Moreover, using a Conditional Random Field (CRF) (Lafferty et al., 2001) as learning algorithm made it possible for us to compare the PA model with other domain-adapted NERs (Passaro and Lenci, 2014).

This paper is structured as follows: In section 2, we present the CoLingLab NER and we show its performance on a sample of PA documents; in section 3 we describe the adaptation of the system to PA texts and its performances (section 4.1). In section 5, we report on the annotation of relations that we performed on a sample of the corpus and finally discuss the results and ongoing work.

2 The CoLingLab NER

The standard Italian CoLingLab NER was trained on the Italian Content Annotation Treebank (I-CAB (Magnini et al., 2006)), a corpus of Italian news, annotated with semantic information at different levels: Temporal Expressions, Named Entities, relations between entities. I-CAB is composed of 525 news documents taken from the local newspaper 'L'Adige' (time span: September-October of 2004). The NEs annotated in the corpus are: Locations (LOC), Geo-Political Entities (GPE), Organizations (ORG) and Persons (PER).

As we said before, this model is unsatisfactory for the domain of Public Administration in two main respects. First, its classes are insufficient to deal with the type of information in the PA documents, that are full of references to other "linked" acts and legislative reference; second, the language used in these documents is a peculiar and highly complex variant of standard Italian (cf. above). In addition, the performance of the model, attested at ∼**0.66** of F1-score on a portion of I-CAB decreases dramatically on the PA documents, reaching a F1-score of ∼**0.35**. To measure such performances, in the test set we mapped ORG_PA (cf. below) with ORG, and in the training set we mapped GPE with LOC.

3 A NER for PA Documents

The adaptation of the CoLingLab NER to the PA domain included the extension of the standard NE classes (Rau, 1991; Grishman and Sundheim, 1996; Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) to other entity types particularly important in the context of municipalities. In particular, we added the class ACT, to mark other administrative documents (normally, PA texts refer to other documents related to the same procedure), the class LAW for the relevant legislation, and an additional class of organizations, ORG_PA, for municipal departments.

3.1 The PA Corpus

For the creation of the corpus, we used documents taken from the *Albo Pretorio Nazionale* with the

[1] The SEMantic instruments for PubLIc administrators and CitizEns (SEMPLICE; `www.semplicepa.it`) is a 2-year project funded by Regione Toscana in collaboration with IT companies to develop NLP-based tools for knowledge management, information extraction and opinion mining for local public administrations.

aim of capturing the variability of the texts produced by PA. Overall, the corpus includes **460** documents, for a total of **724,623 tokens**, annotated with the following NEs: (i) **ACT**: documents belonging to the Albo Pretorio Nazionale, with their type (optional), number and date: *Determina n. 4 del 12/02/2011*; (ii) **LAW**: legislative references: *art. 183 comma 7 del D.Lgs. n. 267/2000*; (iii) **LOC**: locations and geo-political entities: *Comune di Pisa*; (iv) **ORG_PA**: organizations related to the Public Administration such as municipal Departments: *Sezione Anagrafe*; (v) **ORG**: organizations: *Consip Spa*; (vi) **PER**: physical persons. The corpus has been linguistically annotated by means of a pipeline of general purpose NLP tools and in particular, it has been POS-tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009), dependency parsed with the DeSR parser (Attardi et al., 2009). Finally, complex terms like *forze dell'ordine* (security force) have been identified using the EXTra term extraction tool (Passaro and Lenci, 2016).

3.2 Annotation

NE annotation has been performed by means of an incremental process: first 100 documents have been annotated by 2 annotators (one of them was a domain expert). In a second phase we trained a CRF model on these documents and we used it to automatically annotate new documents. Finally, we identified the most common errors of the classifier and two new annotators manually revised the output. This process has been repeated for each group of 100 documents up to covering the whole corpus that includes 460 distinct documents. The average length of the documents is 1,575.26 tokens and the total number of the tokens is 724,623. Figure 1 shows the distribution of the different NE classes in the corpus.

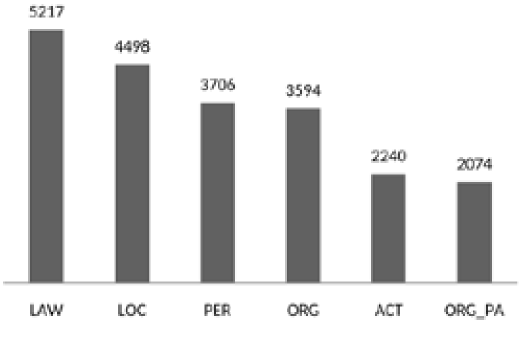


Figure 1: Distribution of the NEs in the corpus.

NEs have been annotated on the CONLL (Nivre et al., 2007) texts using the standard IOB method. In order to deal with acts, we decided to tag them with different "labels" to distinguish their subcomponents: the type (marked with ACT_T), the number (marked with ACT_N), the date (marked with ACT_D), functional tokens (ACT_X) and unparsable tokens (marked wirth ACT_U). For example, the act *Delibera di giunta comunale numero 53 del 23/10/2016* is annotated as follows: *Delibera di giunta comunale (*ACT_T*) numero (*ACT_X*) 53 (*ACT_N*) del (*ACT_X*) 23/10/2016 (*ACT_D*)*, while the act *DD/67/2012* is annotated as ACT_U. This method allows for a simpler normalization of normative references, which is crucial for document retrieval because of the high variability of law mentions in the PA texts.

The inter-annotator agreement between two annotators (attested at ∼0.8) has been calculated using the Cohen's K index on a sample of 25 documents of 25 different municipalities, for a total of 26,190 tokens.

4 System Overview

To train the NER, no information from gazetteers was used. The model includes the following groups of features:

SEQUENCES: Next and previous words and a window of 6 words (3 preceding and 3 following the target word) and their classes;

N-GRAMS: Character-level features, i.e., substrings of the word with a maximum length of 6 letters;

ORTHOGRAPHY: "word shape" features such as spelling, capital letters, presence of non–alphabetical characters etc.;

LINGUISTIC FEATURES: The word position in the sentence (numeric attribute), the lemma, and the PoStag (nominal attribute);

TERMS: We employed complex terms as features to train the model. Terms have been extracted with EXTra (Passaro and Lenci, 2016).

4.1 System's Performances

We trained the CRF model based on the CoLingLab NER on the annotated PA corpus, and we tested its performances first with cross-validation and then on a sample of new 25 documents of 25 different municipalities. This choice stems from the fact that very often different municipalities tend to use different templates and different ways to refer to particular entities. This is particularly common in some NE classes such as ACTS and ORG_PA, that vary a lot across municipalities. For example, some of the analyzed texts contain strings of the form *YYYY/G/NNNNN* to refer to the acts, where the number is actually a string encoding both the date (year: YYYY), a code for the type (G) and the number of the act (NNNNN). Other municipalities instead adopt a less strictly codified pattern to indicate the act such as *Type of act, number N* of DD/MM/YYYY*. Likewise, depending on the writing style (and conventions) of the municipalities, the various departments (i.e., ORG_PA) can include both strings like *Corpo dei Vigili Urbani* and codes like *Tec-01/ICT*. To evaluate the system performance with respect to the variation of the naming conventions adopted by different municipalities, we randomly selected 25 municipalities and one document for each of them balanced for length.

Table 1 reports on the results obtained in cross validation and Table 2 shows the performance on the sample of 25 documents. Figure 2 shows also the confusion matrix for that sample.

In order to investigate the contribution of non-linguistic features, we performed ablation experiments and we tested the results on the sample of 25 documents. The ΔF1-Score for such groups is as follows: SEQUENCES: 3%; N-GRAMS: 1%; ORTHOGRAPHY: 4%. In addition, we performed an additional experiment by training the NER on a combination of I-CAB and the PA documents. In this case, we noticed a ΔF1-Score of 2% by respect to the original model.

	Precision	**Recall**	**F1-Score**
ACT	0.7876	0.8914	0.8356
LAW	0.827	0.8423	0.8343
LOC	0.702	0.7398	0.7196
ORG	0.7085	0.689	0.6977
ORG_PA	0.6158	0.7774	0.6855
PER	0.8373	0.8776	0.8567
MacroAVG	**0.7464**	**0.8029**	**0.7716**

Table 1: System results (10-fold cross validation)

5 Towards a Relational Classifier for PA

For a subset of the corpus, we also annotated the semantic relations occurring between two entities in the domain of the PA, using the following scheme:

	Precision	**Recall**	**F1-Score**
ACT	0.9747	0.8477	0.9068
LAW	0.9494	0.9615	0.9554
LOC	0.799	0.6913	0.7413
ORG	0.8017	0.7686	0.7848
ORG_PA	0.8706	0.7957	0.8315
PER	0.9142	0.8694	0.8912
MicroAVG	**0.914**	**0.8355**	**0.873**
MacroAVG	**0.8849**	**0.8224**	**0.8518**

Table 2: System results (on a sample of 25 texts)

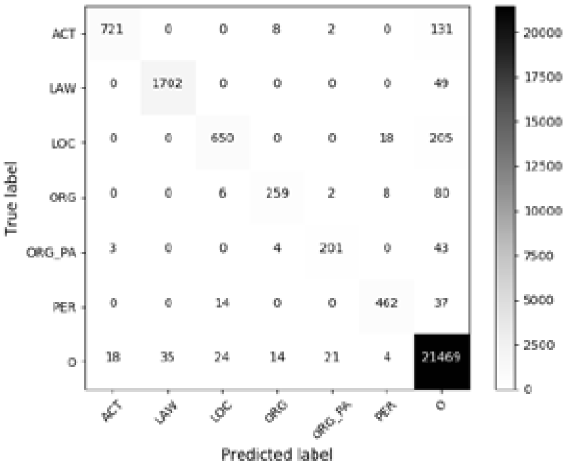


Figure 2: Confusion Matrix (25 texts)

PART OF: the relation of hyponymy, which can occur between: (i) two locations (e.g. a Municipality in Province); (ii) two organizations (e.g. a participated into a holding company); (iii) a person and an organization (e.g. a member of an organization). Implicit attribute for this reation is “work in”.

LOCATION: an entity placed into a particular location, occurring between: (i) an organization and a location (e.g. an organization located in a certain region). Possible attributes for this relation are “work in” and “placed in”; (ii) a person and and a location (e.g. a person living in a particular area). Possible attributes are “work in”, “born in” and “placed in”.

IS RELATED TO: an underspecified relation between any entity pair.

Preliminary experiments have been performed to examine the characteristics of an automatic classifier for extracting relations from administrative acts, and the performance seem to be very promising, despite the size of the training set, which includes in total 100 documents so far. The extension of the annotated corpus and the training of the relational classifier are currently ongoing.

6 Discussion

The results show that the NER reaches satisfactory results for most of the classes, although legging behind in the recognition of PA Organizations, which, among others, tend to have a higher formal variability, including for example both entities like *Corpo dei Vigili Urbani* and *Tec-01/ICT*. Moreover, in the recognition of Location names in the domain of the PA, the system is expected to detect entities with a non-standard detail level going from the name of the municipalities (e.g. *Comune di Pisa*) to very detailed addresses (e.g., *via S. Maria n. 36, 56126 Pisa (PI) interno 15*). A similar problem occurs in the recognition of very small organizations, whose name contains the name of its founder (i.e., *Mario Rossi snc*). In these cases, especially when *snc* is omitted, the system predicts the class PER instead of the correct class ORG. We are confident that adding lexicons and gazetteers will improve the identification of entities of this kind, but it could be interesting to investigate automatic normalization, disambiguation and entity linking approaches (Hoffart et al., 2011; Han et al., 2011).

7 Conclusions and Ongoing Work

Named entities play an important role in administrative acts, especially in those - like the documents in the Albo Pretorio - describing the main actions taken by Municipalities. This kind of information is very useful to fullfil the obligations related to supervisory monitoring, disclosure, periodic self-assessment, and review of the government decisions.

In this paper, we presented a NER for PA that shows a significant ability to identify the relevant entities, and in particular legislative reference and connected acts. It is important to stress the lexical and syntactic complexity of bureaucratic language represents a big challenge for NLP tools and methods. Such a complexity derives from the technical lexis of other domain-specific languages with which PA deals daily, such as education, environment, ICT technologies, public health and so on. In near feature we plan to explore the possibility of re-engineering our system to take advantage of new algorithms for entity extraction such as neural networks and in particular from character level word embeddings. Moreover, we will focus on the development of classifiers for Relation Extraction and Entity Linking.

Acknowledgments

This research has been supported from the Project SEMantic instruments for PubLIc administrators and CitizEns (SEMPLICE), funded by Regione Toscana, and the Company ETI^3| Evolution, Technology & Innovation. Special acknowledgements go to Roberto Battistelli and Francesco Sandrelli (ETI^3) for support, and to the students Roswita Candusso, Carmela Cinquesanti, Federica Semplici and Ludovica Vasile for manual annotation.

References

Douglas E. Appelt, Jerry R Hobbs, John Bear, David Israel, and Mabry Tyson. 1993. Fastus: A finite-state processor for information extraction from real-world text. In *IJCAI*, volume 93, pages 1172–1178.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, and Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy). Springer.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201, Washington, DC. Association for Computational Linguistics.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Nyu: Description of the mene named entity system as used in muc-7. In *In Proceedings of the Seventh Message Understanding Conference (MUC-7*.

Dominique Brunato. 2015. A study on linguistic complexity from a computational linguistics perspective. a corpus-based investigation of italian bureaucratic texts. Ph.D. Thesis, University of Siena.

Indra Budi and Stéphane Bressan. 2003. Association rules mining for name entity recognition. In *Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on*, pages 325–328. IEEE.

Jason P.C. Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297.

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy). Springer.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Ann Arbor, Michigan (USA). Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 466–471. Association for Computational Linguistics.

Ralph Grishman. 1995. The nyu system for muc-6 or where's the syntax? In *Proceedings of the 6th Conference on Message Understanding*, pages 167–175, Columbia, Maryland. Association for Computational Linguistics.

Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: A graph-based method. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 765–774, Beijing (China).

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh (United Kingdom). Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA (USA). Morgan Kaufmann Publishers Inc.

Bernardo Magnini, Emanuele Pianta, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. Italian content annotation bank (i-cab): Named entities.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191, Edmonton (Canada). Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague (Czech Republic). Association for Computational Linguistics.

Lucia C. Passaro and Alessandro Lenci. 2014. "il piave mormorava...": Recognizing locations and other named entities in italian texts on the great war. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 286–290, Pisa (Italy).

Lucia C. Passaro and Alessandro Lenci. 2016. Extracting terms with extra. In *Proceedings of the EUROPHRAS 2015 – Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 188–196, Malaga (Spain).

Lisa F. Rau. 1991. Extracting company names from text. In *Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on*, volume 1, pages 29–32. IEEE.

Rohini Srihari, Cheng Niu, and Wei Li. 2000. A hybrid approach for named entity and sub-type tagging. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 247–254, Seattle, Washington (USA). Association for Computational Linguistics.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2660–2670.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: language-independent named entity recognition. In *Proceedings of the 6th conference on Natural language learning*, volume 31, pages 1–4.

Mining Offensive Language on Social Media

Alessandro Maisto, Serena Pelosi, Simonetta Vietri, Pierluigi Vitale
University of Salerno
Department of Political, Social and Communication Science
Via Giovanni Paolo II, 132
{amaisto,spelosi,vietri,pvitale}@unisa.it

Abstract

English. The present research deals with the automatic annotation and classification of vulgar ad offensive speech on social media. In this paper we will test the effectiveness of the computational treatment of the taboo contents shared on the web, the output is a corpus of 31,749 Facebook comments which has been automatically annotated through a lexicon-based method for the automatic identification and classification of taboo expressions.

Italiano. *La presente ricerca affronta il tema dell'annotazione e della classificazione automatica dei contenuti volgari e offensivi espressi nei social media. Lo scopo del nostro lavoro consiste nel testare l'efficacia del trattamento computazionale dei contenuti tabù condivisi in rete. L'output che forniamo un corpus di 31,749 commenti generati dagli utenti di Facebook e annotato automaticamente attraverso un metodo basato sul lessico per l'identificazione e la classificazione delle espressioni tabù.*

1 Introduction

Flaming, *trolling*, *harassment*, *cyberbullying*, *cyberstalking*, *cyberthreats* are all terms used for referring to vulgar and offensive contents shared on the web. The shapes can be different and the focus can be on various topics, such as physical appearance, ethnicity, sexuality, social acceptance and so forth.

Although taboo language is generally considered to be the strongest clue of harassment in the web, it must be clarified that the presence of bad words in posts does not necessarily indicate the presence of offensive behaviors. The words collected in vulgar lexicons, in some cases, are neutral or even positive. Moreover, profanity can be used with comical or satirical purposes, and bad words are often just the expression of strong emotions (Yin et al., 2009).

In this paper, we propose a system for the automatic treatment of vulgar and offensive utterances in Italian. The strength of our method is that lexical items are not considered in isolation. Instead, we recognize the power of the local context of the words, which can modulate the meaning of words, phrases and sentences.

Section 2 briefly illustrates the state of the art contributions on offensive language modeling. Next, Section 3 describes the Italian lexical and grammatical resources for the automatic detection of taboo language in Italian. Then, Section 4 explains how we tested our method and resources on a Facebook corpus and describes the results of the taboo expressions automatic annotation. Finally, Section 5 reports the future works that will enhance our research.

2 State of the Art on the Computational Treatment of Offensive Language

As it is anticipated, taboo words are basically considered a strong clue of online hate speech (Chen et al., 2012; Reynolds et al., 2011; Xu and Zhu, 2010; Yin et al., 2009; Mahmud et al., 2008). Nevertheless, the methods that simply match offensive words stored in blacklists, are clearly not meant to reach high levels of accuracy. Consistent with this idea, in the recent years many studies on offensive cyberbullying and flame detection integrated the bad words context in their methods and tools. Chen et al. (2012) exploited a Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potential offensive users in social media. Xu and Zhu (2010) proposed a sentence-level semantic filtering approach that combined grammatical relations with offensive words. Insulting phrases and derogatory comparisons of human beings with insulting items or animals were clues used by Mahmud et al. (2008). Razavi et al. (2010) proposed an automatic flame detection method based on the variety of statistical models and the rule-based patterns. Among the flame topics that they identified, there are attacks and abuses that embarrass the readers. Xiang et al. (2012) learned topic models from a dataset of tweets through Latent Dirichlet Allocation (LDA) algorithm. Waseem and Hovy (2016) and Kwok and Wang (2013) focused on racist and sexist slurs on Twitter; Waseem and Hovy (2016) made reference to hate speech expressed without any derogatory term, and Kwok and Wang (2013) focused on the relation between the tweet content and the identity of the user, on the base of which a post is considered to be racist or not. Badjatiya et al. (2017) also used Twitter in order to investigate the application of deep neural network architectures.

3 Lexical and Grammatical Resources

In this paragraph we will describe the Italian lexical database and the grammatical rules which have been used as indicators for the automatic identification of the taboo language.

The items of the lexicon are labeled through the use of the following three main categories:

- *Trait*, that specifies if the taboo expression is addressed to other users, to events or to things;
- *Type*, that verifies if an expression is *offensive*, if it represents a *threat* or if it is just *rudeness*;
- *Semantic Field* which specifies the taboo domain (namely *sex, sexism, aesthetics, behavior, homophobia, racism, scatology*).

Such tags have been collected and classified by a team of four annotators (one linguist and three Italian native speakers), which annotated the linguistic resources through an agreement of 92%.

Taboo words, which were impossible to classify through a defined semantic field, have been annotated with the residual category "N.C.". Our taboo lexicon is composed of

- *Simple Words*, which include nouns, adjectives, verbs and adverbs collected from the Sentiment Lexicon *SentIta* (Pelosi, 2015) and manually evaluated with reference to the categories described above;
- *Multiword Expressions* (MWE), that are nouns automatically annotated through the integrated use of the simple words list and *ad hoc* regular expressions (e.g. see section 3.2);
- *Idiomatic Structures*, which are verbs + frozen complement collected from Vietri (2014) and manually annotated on the grounds of the hate speech tags.

This choice is due to the fact that in colloquial and informal situations, a taboo expression can work simply as intensifier, also for positive sentences (e.g. *it's fucking nice*!). This is why the words' semantic orientation must be, case by case, modulated when occurring into the context of (semi)frozen structures. Concrete examples are idiomatic structures that involve concrete nouns indicating body part (with a vulgar meaning) as fixed constituents (e.g. *essere culo e camicia*, "to be thick as thieves").

3.1 Simple Vulgar Words

Our project is grounded on a collection of 342 taboo simple words that include the following grammatical categories: nouns, verbs, adjectives, adverbs and exclamations. Nouns count 242 entries, among which 216 are simple words (e.g. *cozza*, "mussel", addressed to ugly women) and 26 are monorematic compounds (e.g *rompiballe* "pain in the ass"). Verbs count 72 entries, among which 27 are verbs indicating bodily predicates that involve acts of violence, e.g. *violentare*, "to rape", and 21 are pro-complementary and pronominal verbs (e.g. *incazzarsi* "to get mad"). Adjectives count 16 entries (e.g. *cazzuto* "die-hard"), adverbs 4 entries (e.g. *incazzosamente* "grumpily") and exclamations 8 entries (e.g. *vaffanculo* "fuck off").

3.2 Taboo Multiword Structures

The simple words listed in our database, especially the ones with an uncertain semantic orientation (see "N.C. in Figure 1"), can be part of frozen or semi-frozen expressions that can make clear, for each occurrence, the actual meaning of the words in context.

Idioms are particularly interesting in a work on online harassment, because they are open to word-plays and trolls. Indeed, it must be reported a higher than expected presence of idiomatic structures in our corpus. Nevertheless, their syntactic flexibility and the lexical variations make them very difficult to automatically locate, if compared with other multiword expressions. A very typical Italian example is *cazzo* "dick", with its, more or less vulgar, stilistic and regional variants (e.g. *minchia*, *pirla*, *cavolo* "cabbage", *cacchio* "dang", *mazza* "stick" , *tubo* "pipe", *corno* "horn", etc...). The context systematically gives the word under examination a clear connotation. Examples are (negative) adverbial and adjectival expressions (e.g. *a cazzo*, "fucked up"); (emphatic) exclamations and interrogative forms (e.g. *che cazzo* "what the hell"); intensification of negations (e.g. *non* V *un cazzo*, "don't *V* shit").

Multiwords Expressions. With Multiwords Expressions, we mean sequences of simple words separated by blanks, characterized by semantic atomicity, restriction of distribution, shared and established use and lack of ambiguity. In this research, we automatically located and annotated MWEs through the combined use of the taboo simple words that trigger the recognition and a set of regular expressions (based on part of speech patterns) that locate the MWEs (e.g. *culo rotto*, "lucky" from the simple noun *culo* and the pattern *NA*). Other MWEs are those ones related to idioms (see next paragraph, e.g. *rottura di palle* "nuisance"). The regular expressions used to identify the taboo MWEs are summarized below:

- *Taboo Noun* + *Preposition* + *Noun* (NPN)
- *Noun* + *Preposition* + *Taboo Noun* (NPN)
- *Taboo Noun* + *Adjective* (NA)
- *Noun* + *Taboo Adjective* (NA)
- *Adjective* + *Taboo Noun* (AN)
- *Taboo Adjective* + *Noun* (AN)

Idiomatic Expressions. Among the possible idiomatic structures, the present research focuses on those idioms (verb and at least one frozen complement) which have vulgar nouns of body part as frozen complement.

The lexical resources used in this research are composed of 52 items that include 28 ordinary verb structures (e.g. *girare le palle* "to bust the balls") and 23 support verb idioms (*avere culo* "to be lucky"). The classes to whom they belong (Vietri, 2014) are various and can be in systematic correlation, as it happens with *girare le palle/avere le palle girate*, "to bust the balls/to have the balls busted".

The idioms under examination can be also related to some derived nominals in *-tore,-trice,-ura,-ata* (e.g. *rottura di palle* "pain in the arse") and/or with VC compounds (verb + fixed constituent e.g. *rompipalle* "ball-buster"). These compounds occur in the corpus as both simple words and multiword units.

The automatic recognition of taboo idioms, similar to MWEs, start from the nouns indicating vulgar body parts, and proceed with another lexical anchor that is associated to the idiom in the lexical resources (e.g. *girare le palle* "to piss off" is annotated in the corpus when the tool locates at the same time *palle* e *girare* with a maximum distance of three word forms). This procedure streamlines the automatic recognition of the idioms, guaranteeing high levels of recall in spite of the large variety of syntactic transformations that the frozen structure can go through (causative constructions, infinitive forms preceded by *da*, dislocation, modification, among others (Vietri, 2014)).

4 Experiment and Evaluation

The linguistic resources described so far have been tested on a large corpus of User-Generated-Contents scraped by Facebook. We chose an Italian Facebook page called *Sesso Droga e Pastorizia*, which became popular for its explicit and offensive contents. The page has been shut down the 10/03/2017 for the social network policy violation; therefore, the page's administrators created a set of connected pages in order to continue the activity in case of temporal or definitive closing. For our experimentation, posts and comments have been extracted from three pages correspondent to the following indices: *sessodrogapastorizia1*, *sessodrogapastorizia3* and *sessodrogapastoriziariserva*. The corpus includes 31,749 comments published between 28 March 2017 and 13 April 2017 by over 20 thounsand users, replying to 122 status. We extracted 2,797 taboo expressions with a Recall of 97% and a Precision of 83% by applying dictionaries and grammars to the generated corpus[1].

Figure 1 represents a bubble chart which illus-

[1] The Recall has been evaluated on the entire corpus of over 31,000 comments, while the Precision has been calculated on the extracted 2,700+ sentences.

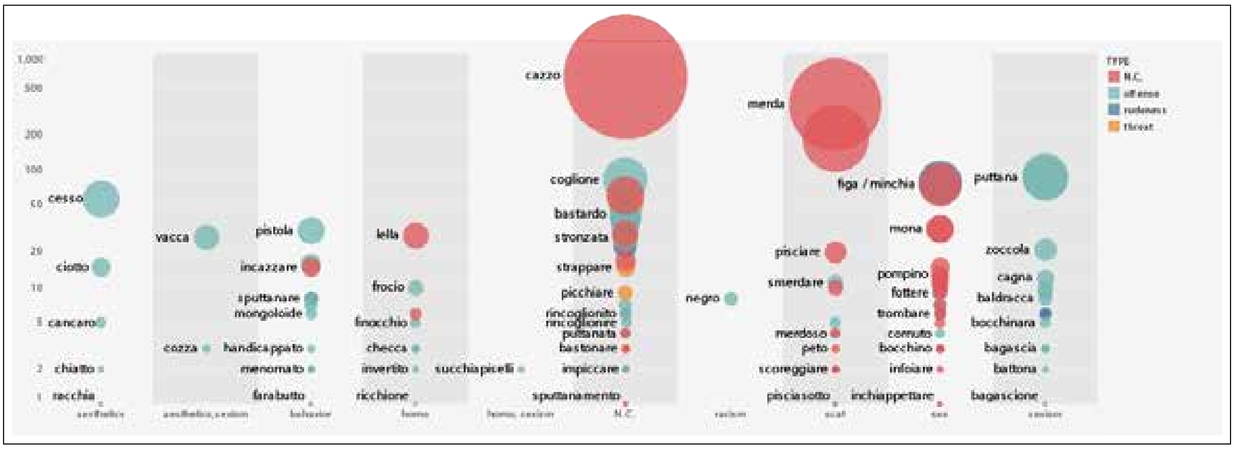


Figure 1: Extracted words occurrence, types and fields

trates the distribution of *Semantic Fields* and *Types* of the extracted words. The fields are listed in the horizontal axis. The vertical axis and the size of the circles describe together the frequency of the extracted items. Finally, the colors of the circles represent the words' types.

As far as MWEs are concerned, we extracted 134 idioms; moreover 597 MWEs have been annotated as NPN structures, 213 as NA and 175 as AN. Among the most frequent MWEs we mention some items which were already listed into the dictionaries (e.g. 9 occurrences of *testa di cazzo*, "dickhead"; 7 occurrences of *pezzo di merda*, "piece of shit").

Also new vulgar structures belonging to various fields have been automatically located through our strategy (e.g. 51 occurrences of *cazzo duro*, "hard-on" from the field *sex*; 3 occurrences of *gran troia* "total slut" from the *sexism* field; 2 occurrences of *busta di piscio* "box of piss" from the *scat* field).

The extracted patterns underline the relevance of the local context in the disambiguation of some words which have classified N.C. as simple words, because of their ambiguity out of the context. An example is *cazzo* which, alone, did not receive any field or type label, but as a MWE clearly belongs to defined categories. *Cazzo duro* belongs to the *sex* field. *Cazzo di + Noun* "this fucking + N" is a generic offense (e.g. *cazzo di pagina* "this fucking page") and *cazzo di + Taboo Noun* represents an intensification of the expressed offensive term (e.g. *cazzo di zingaro* "this fucking gypsy").

5 Conclusion

In this paper we described an experiment on the detection and classification of offenses, threats and insults shared through User Generated Contents.

As a matter of fact, in May 2016, the European Commission, together with companies like Facebook, Twitter, YouTube and Microsoft, underlined the relevance of these topics by presenting a code of conduct[2] which aimed to constrain the virality of illegal online violence and hate speech, with a special focus on utterances fomenting racism, xenophobia and terrorist contents. The negative impact of such practices is not limited to individuals, but strongly affects the freedom of expression and the democratic discourse on the Web.

Our research focused on a particular Facebook page, which became famous in Italy for the number of times it has been shut down due to its disturbing content. More than 31,000 users' comments downloaded from this page have been automatically annotated according to a dataset of taboo expressions, in the form of simple words and multiword expressions. This operation has led to a hate speech annotated corpus which distinguishes eight harassment *semantic fields*, four *types* of insult and four hate targets (*traits*). The evaluation of the experiment performances confirmed the hypothesis that the local context of words represents an essential feature for an effective hate speech mining on the web.

In future works we will test the interaction of the taboo item located in the corpus with some Italian Contextual Valence Shifters (Maisto and Pelosi, 2014) in order to verify if the sentence context of the insult indicators affects the semantic orientation of the items into an Opinion Mining

[2] `http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf`

view.

Furthermore, it would be interesting to verify the efficacy of our resources and our method on different domains, Political Communication, among others.

In the end, just because the automatic extraction has been done in this paper on a very polarized corpus, future analyses will focus on testing the reliability of this research on more neutral collections of texts.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. 2008. Detecting flames and insults in text. In *Proceedings of the Sixth International Conference on Natural Language Processing*. BRAC University.

Alessandro Maisto and Serena Pelosi. 2014. A lexicon-based approach to sentiment analysis. the italian module for nooj. In *Proceedings of the International Nooj 2014 Conference, University of Sassari, Italy*. Cambridge Scholar Publishing.

Serena Pelosi. 2015. Sentita and doxa: Italian databases and tools for sentiment analysis purposes. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 226–231. Accademia University Press.

Amir Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. *Advances in Artificial Intelligence*, pages 16–27.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE.

Simonetta Vietri. 2014. *Idiomatic Constructions in Italian: A Lexicon-grammar Approach*, volume 31. John Benjamins Publishing Company.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@ HLT-NAACL*, pages 88–93.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM.

Zhi Xu and Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the WEB*, volume 2, pages 1–7.

Automatic Evaluation of Employee Satisfaction

Marco Piersanti **Giulia Brandetti** **Pierluigi Failla**

Data Modeling and Analysis – Enel Italia S.R.L.
Rome, Italy
{name}.{surname}@enel.com

Abstract

English. Human Resources are one of the most important assets in modern organizations. Their capability of facing employees' needs is critical in order to have an effective and efficient company, where people are the center of all business processes. This work is focused on developing new techniques that, leveraging a data driven approach, can help Human Resources to find a more precise employee satisfaction categorization, to easily identify possible issues and to act in a proactive fashion.

Italiano. *Le Risorse Umane sono una delle funzioni piú importanti nelle aziende moderne. La loro capacità di affrontare le necessità dei dipendenti è fondamentale per avere un'azienda efficiente, dove le persone sono al centro di tutti i processi di business. Il presente lavoro è focalizzato sullo sviluppo di nuove tecniche che, facendo leva su un approccio data driven, possano aiutare le Risorse Umane a dare una categorizzazione della soddisfazione dei dipendenti piú precisa, ad identificare piú facilmente possibili problemi condivisi e ad agire in maniera proattiva.*

1 Introduction

Every modern organization has a dedicated function which takes care of its employees, commonly called Human Resources (HR). HR duties are related to the capability of creating value through people, ensuring that everyone can express his own potential and has a productive and comfortable office environment.

Nowadays, HR can rely on data to create a new paradigm based on a *data driven* approach, where analysts can leverage data in order to get more complete, detailed and data-supported decisions.

Being able to monitor employees' engagement and satisfaction is critical in order to maintain a positive and constructive office environment. The benefit for the company is in the capability of retaining the best employees and keeping the overall workforce strong and motivated. Furthermore, recent surveys (Globoforce, 2015) show the issues that companies are facing when they try to do retention or improve engagement.

This paper is organized as follows. Section 2 presents a literature review on both themes of HR Management and text mining, Section 3 summarizes the motivations that drove the present study, Sections 4 and 5 discuss data and methodology, respectively, and Section 6 presents the results. Finally, Section 7 discusses the implications of the findings and further possible developments.

2 Related Works

Despite the great interest that is arising around the application of Data Science methods and Natural Language Processing (NLP) to HR problems, very few studies exist on the topic.

The entire field of corporate HR Management has been revolutionized by the pioneering work done by People Operations at Google (well described in Bock (2015)), that first put a spotlight on the benefits of having a more scientific and rigorous approach to these areas which have been traditionally more reluctant to adopt change.

Employee satisfaction has been linked to long-run stock returns (Edmans, 2011), consistently with human relations theories which argue that employee satisfaction brings a stronger corporate performance through improved recruitment, retention, and motivation. Furthermore, Moniz and Jong (2014) followed an interesting approach to link employee satisfaction and firm earnings, based on sentiment analysis of employees' reviews from the career community website `www.glassdoor.com`.

Text clustering, and more generally text classification, is a well established topic in the NLP research area (Sebastiani, 2002; Aggarwal and Zhai, 2012; Kadhim et al., 2014). The automated categorization of texts, although dating back to the early '60s (Maron, 1961; Borko and Bernick, 1963), went through a booming interest in the last twenty years, due to the explosion of the amount of documents available in digital form and the impelling need to organize them. Nowadays text classification is used in many applications, ranging from automatic document indexing and automated metadata generation, to document filtering (e.g., spam filters (Drucker et al., 1999)), word sense disambiguation (Navigli, 2009), population of hierarchical catalogs of Web resources (Dumais and Chen, 2000), and in general any application requiring document understanding.

Flourished in the last decade, sentiment analysis aims to classify the polarity of a given text – whether the expressed opinion in a document or a sentence is positive, negative, or neutral (Pang et al., 2002; Pang and Lee, 2008; Baccianella et al., 2010; Liu, 2012). The growing interest on the subject reflects on the success of the tasks of sentiment analysis on Twitter data at SemEval since 2013 (Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov, 2016). Even if the driving language for most of those techniques is English, we started to see an increasing trend also in Italy (Basile and Nissim, 2013; Basile et al., 2014; Basile et al., 2015), confirming the great interest of the Italian NLP community in sentiment analysis techniques.

3 Task Description

Enel HR Business Partners' (HR-BPs) job consists in monitoring employees' well-being, acting when necessary to solve issues. In doing so, they periodically interview employees and register information about their satisfaction, motivation, work-life balance and other personal issues in textual notes.

Currently, employees are manually classified by HR-BPs in three main categories: *Demotivated*, *Neutral* and *Motivated*. Unfortunately, employee motivation is not a very reliable indicator of employee well-being, since it may mask an underlying dissatisfaction, or more generally the presence of issues that HR department should act on. Indeed, one can face several problems in the office everyday life but still be motivated. We therefore chose to consider the sentiment, as it shows through interviews, as a proxy of employee satisfaction.

With the present study, we aim to categorize employee satisfaction in a more detailed and automatic way, identifying common trends among employees and clustering them into groups that share similar problems. The goal is to help HR-BPs in having an overall view of their resources' mood and make effective adjustments in critical situations. It will also help in such situations when new HR-BPs take over a group of already interviewed resources, allowing them to have a clearer understanding of the employees and their criticalities without having to read all interviews.

For all the aforementioned reasons, we performed a classification of the interviews based on their sentiment (Section 5.1) prior to send them into the text clustering algorithm (Section 5.2). In the present study, we chose to focus only on negative moods, since they include the biggest issues HR should monitor. Nevertheless, the practical usage of this system involves the whole set of sentiment classes, since HR is interested in monitoring the entire workforce well-being and in following its evolution over time.

In choosing methods, we had to tackle the challenge to balance the scientific rigor and the need of ease of interpretation and communication to all actors involved in the process. We therefore chose to use well understood and controllable techniques, like *sentiment analysis* and *k-means clustering*.

4 Experiments and Data

4.1 Data Description

HR System Integration provided interviews data, a file containing 53k textual notes in more than 5 languages taken by HR-BPs during interviews. Interviews spanned approximately 1 year, from June 2015 to July 2016, and they were performed by 142 different HR-BPs.

For the present study, we focused only on Italian interviews (25k interviews) and selected a single interview for each employee (23k interviews), since in the few cases of repeated interviews texts were not relevant (e.g., "See previous interview").

Notes shorter than 5 words (the 5th percentile of the distribution of the number of words in each note) were considered irrelevant. As a result, in the present study we considered a dataset of 22k interviews.

4.2 Data Preprocessing

Data preparation includes removing punctuation, numbers and stop words (we removed 300 common Italian stop words, including some peculiar words that are not relevant in this context, like “Enel”, “colloquio”, etc.), changing letters to lower case and lemmatization (Schmid, 1994). We assumed all unrecognized words to be typos, and we corrected them by using a dictionary composed by 110k Italian words and 650 English words commonly used in business daily-life[1]. In order to have an effective correction, we used Optimal String Alignment distance (Brill and Moore, 2000) (OSA distance), an extension of Levenshtein distance that, together with insertion, deletion and substitution, includes transpositions among its allowable operations.

5 Model Description

5.1 Sentiment Analysis

We performed sentiment classification of texts by customizing and improving a publicly available lexicon[2]. In total, we used 3428 Italian labeled unigrams and 10451 bigrams, categorized as positive (4736), neutral (4367) or negative (4776) based on their polarity.

The sentiment classification model proposed in this paper is based on a score φ_{sent} that weights differently unigrams and bigrams with a factor α:

$$\varphi_{sent} = (1-\alpha)\cdot\varphi_{uni} + \alpha\cdot\varphi_{bi}$$

where $0 \leq \alpha \leq 1$, φ_{uni} is the difference between the number of positive and negative unigrams, normalized by the number of words in the text and φ_{bi} is the difference between the number of positive and negative bigrams, normalized by the number of bigrams in the text. Final sentiment was then calculated according to the formula

$$\text{Sent} = \begin{cases} +1 & \text{if } \varphi_{sent} > \theta \\ -1 & \text{if } \varphi_{sent} < -\theta \\ 0 & \text{otherwise.} \end{cases}$$

Model calibration (i.e. the choice of parameters α and θ) was performed by comparing model results with the ones produced by manually annotating a subset of 200 (randomly chosen) texts (*training set*): two judges classified texts independently and a third one solved the cases where there wasn’t agreement. Agreement between the two independent judges was measured by calculating Cohen’s Kappa ($\kappa = 0.6$).

We chose $\alpha = 0.7$ and $\theta = 0.0004$ so that accuracy, recall and precision of the sentiment model were maximized. Although we may have chosen to optimize parameters in order to maximize negative texts recognition, we chose to consider the overall accuracy on the three classes, because from a business perspective it is more valuable to monitor the entire workforce satisfaction and to follow its evolution over time. While for α we tried manually different settings, weighting more bigrams than unigrams, for θ we used the ROC curve and the area under it, picking the one with maximal sum of true-positive and false-negative values.

5.2 Text Clustering

For notes’ clustering, we focused only on those classified as negative from the sentiment model (Section 5.1).

Since we didn’t have a target variable to model (unsupervised classification), we chose to adopt the k-means clustering algorithm, using k-means++ technique to seed the initial cluster centers (Arthur and Vassilvitskii, 2007).

The clustering model was applied on the TF-IDF matrix, built with bigrams appearing in at least 2 documents. In this way, we reduced our dimensionality from the initial 37k bigrams to 5k. To calculate proximity among documents, we used cosine similarity.

Additionally, *Silhouette distance* has been chosen to select the best number of clusters: different models were computed by varying the number of clusters between 2 and 30 and the respective Silhouette scores were compared, fixing the number of clusters at 12 (corresponding to the highest score).

6 Results

The application of this sentiment model (Section 5.1) classified interviews in 3655 negatives, 956 neutrals and 17297 positives. As we can see in Table 1, sentiment classification is more clearly related to employee satisfaction than motivation classes provided by HR-BPs, although they some-

[1] `https://github.com/napolux/paroleitaliane`

[2] `https://github.com/opener-project/public-sentiment-lexicons`

Text (after preprocessing)	HR-BP Motivation	Sentiment
risorsa brillante neodirigente clima positivo ansioso molto positivo (*brilliant resource new executive positive mood anxious very positive*)	Motivated	+1
assenteista risorsa molto critico non riuscire nulla (*absentee very critical resource don't succeed in anything*)	Demotivated	-1
non valorizzare poco riconoscimento non potere rimanere (*don't valorize inadequate recognition can't stay*)	Motivated	-1
molto scontento non credere azienda reale meritocrazia interessare piano esodo (*very unhappy don't believe company real meritocracy interest retirement plan*)	Motivated	-1
stabile routinario non proattivo scarso impegno (*stable routine not proactive scarce effort*)	Neutral	-1
assumere direttamente assistente seguire particolare sicurezza vedere capo (*hire directly assistant follow particular safety see boss*)	Neutral	0

Table 1: Examples of sentiment classification and comparison with HR-BPs motivation classes.

True/Predicted	-1	0	1	All
-1	12	11	3	26
0	3	20	18	41
1	1	37	95	133
All	16	68	116	200

Table 2: Confusion matrix. True values here represent manually labeled texts.

times are aligned.

A different subset of 200 manually labeled texts (*test set*), labeled with the same methodology as described in Section 5.1, was used for evaluating model performance. Accuracy and recall were both 64%, while precision was 70%. For more details about the sentiment classification performance, see confusion matrix in Table 2.

The clustering algorithm was applied only on the 2392 negative interviews and it identified 8 clusters that we were able to precisely label, while for the remaining 4 clusters labeling was unfeasible (see Table 3). Labels were applied by manually looking at the most frequent bigrams within clusters, trying to identify common significant topics.

The most frequent identified issues preventing employee satisfaction were *health problems*, the will to *change activity*, *compensation* and the high *workload*. The most frequent bigrams for clusters 0–3 were not specific enough to lead to a precise labeling, since they refer to work activity and job in general and they don't focus on clear issues.

In Figure 1, we represented clustering results by means of t-SNE, a popular method for exploring high-dimensional data (Maaten and Hinton, 2008). By this mean, we reduced the high-dimensionality space of bigrams to an artificial two-dimensional space (since dimensions here don't have a real meaning, we excluded them from the plot). For the sake of clarity, we chose not to show unlabeled clusters; the resulting plot shows that clusters are well separated and on average quite dense.

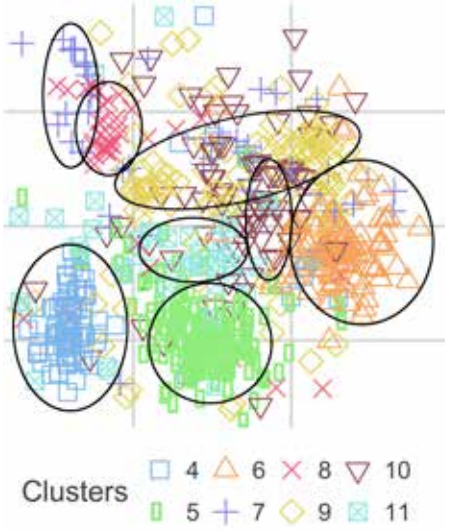


Figure 1: Clustering results represented with t-SNE. Only labeled clusters are shown.

7 Conclusions

The proposed approach could be a powerful tool for HR-BPs to better understand the main issues related to the lack of employees' satisfaction. Furthermore, it could help HR analysts to quickly decide which are the best actions to solve those issues, analyzing whether a complaint is isolated or shared by a group, whether it's trivial or urgent and act accordingly. As an example, HR Departments could test different actions over a group of unsatis-

Cluster id	Docs #	Label	Most frequent bigrams
0	382	(NA)	lavoro svolgere (*do work*)
1	76	(NA)	persona supporto (*support person*) supporto dipendente (*employee support*) carico lavoro (*workload*)
2	1985	(NA)	lavoro piacere (*enjoy work*)
3	33	(NA)	attività poco (*activity low*) solo attività (*only activity*) attività dovere (*activity must*)
4	149	Workload	carico lavoro (*workload*) eccessivo carico (*exaggerated load*) lamentare eccessivo (*complain about exaggerated*)
5	297	Health issues	problema salute (*health issue*) grave problema (*difficult problem*) serio problema (*serious problem*)
6	206	Change activity	cambiare attività (*change activity*) volere cambiare (*want to change*)
7	81	Low productivity	poco produttivo (*low productivity*)
8	67	Not productive	rispetto compito (*compliance with task*) compito non produttivo (*not productive task*)
9	173	Compensation	mancato riconoscimento (*lacking recognition*) lamentare mancato (*complain about lacking*)
10	134	Don't change activity	svolgere attività (*do activity*) volere continuare (*want to go on*) continuare svolgere (*keep doing*)
11	72	Change job	cambio attività (*activity change*) cambiare lavoro (*change job*)

Table 3: Clustering results. Cluster id, number of documents within clusters, cluster labels and most frequent bigrams inside clusters are shown. Labels were applied by manually looking at the most frequent bigrams within clusters.

fied employees, in order to understand which one is the most effective for a given issue.

The very same model could also be used on neutral and positive subjects, so that HR could check whether the quality of life at work of these employees could be somehow improved, and understand which are the essential key factors for the employees' well-being.

From a technical point of view, one possible improvement in order to strengthen the solidity of the present approach could be to manually annotate a subset of (anonymized) texts, developing a gold standard of HR interview clusters, to be used as a test set for techniques like the one presented in this study. This gold standard may be made available company-wise, in order to encourage collaboration and to foster the creation of a data science community, to help bring a data driven way of thinking even to those areas which have been traditionally more reluctant to adopt a rigorous digital transformation.

This is a first step to improve how HR Departments operate nowadays. We strongly believe that the introduction of a data driven approach can support critical HR decisional processes and improve companies' productivity, without having to sacrifice each individual's quality of life.

Acknowledgements

This research was supported by Enel. We thank our colleagues from HR System Integration dep. who provided the data analyzed in this study.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th*

Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 100–107.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the evalita 2014 sentiment polarity classification task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*.

Pierpaolo Basile, Valerio Basile, Malvina Nissim, and Nicole Novielli. 2015. Deep tweets: from entity linking to sentiment analysis. In *Proceedings of the Italian Computational Linguistics Conference (CLiC-it 2015)*.

Laszlo Bock. 2015. *Work rules!: Insights from inside Google that will transform how you live and lead*. Hachette UK.

Harold Borko and Myrna Bernick. 1963. Automatic document classification. *Journal of the ACM (JACM)*, 10(2):151–162.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 286–293. Association for Computational Linguistics.

Harris Drucker, Donghui Wu, and Vladimir N. Vapnik. 1999. Support vector machines for spam categorization. *IEEE Transactions on Neural networks*, 10(5):1048–1054.

Susan Dumais and Hao Chen. 2000. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263. ACM.

Alex Edmans. 2011. Does the stock market fully value intangibles? employee satisfaction and equity prices. *Journal of Financial Economics*, 101(3):621–640.

Globoforce. 2015. 2015 employee recognition report – culture as a competitive differentiator. Technical report.

Ammar Ismael Kadhim, Yu-N Cheah, and Nurul Hashimah Ahamed. 2014. Text document preprocessing and dimension reduction techniques for text document clustering. In *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*, pages 69–73.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Melvin Earl Maron. 1961. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8(3):404–417.

Andy Moniz and Franciska Jong. 2014. Sentiment analysis and the impact of employee satisfaction on firm earnings. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416*, ECIR 2014, pages 519–527, New York, NY, USA. Springer-Verlag New York, Inc.

Preslav Nakov. 2016. Sentiment analysis in twitter: A semeval perspective. In *Proceedings of NAACL-HLT*, pages 171–172.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 73–80. Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 154–164.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Hate Speech Annotation: Analysis of an Italian Twitter Corpus

Fabio Poletto
Dipartimento di StudiUm
University of Turin
f.poletto91@gmail.com

Marco Stranisci
Acmos
marco.stranisci@acmos.net

Manuela Sanguinetti,
Viviana Patti,
Cristina Bosco
Dipartimento di Informatica
University of Turin
{msanguin,patti,bosco}@di.unito.it

Abstract

English. The paper describes the development of a corpus from social media built with the aim of representing and analysing hate speech against some minority groups in Italy. The issues related to data collection and annotation are introduced, focusing on the challenges we addressed in designing a multifaceted set of labels where the main features of verbal hate expressions may be modelled. Moreover, an analysis of the disagreement among the annotators is presented in order to carry out a preliminary evaluation of the data set and the scheme.

Italiano. *L'articolo descrive un corpus di testi estratti da social media costruito con il principale obiettivo di rappresentare ed analizzare il fenomeno dell'hate speech rivolto contro i migranti in Italia. Vengono introdotti gli aspetti significativi della raccolta ed annotazione dei dati, richiamando l'attenzione sulle sfide affrontate per progettare un insieme di etichette che rifletta le molte sfaccettature necessarie a cogliere e modellare le caratteristiche delle espressioni di odio. Inoltre viene presentata un'analisi del disagreement tra gli annotatori allo scopo di tentare una preliminare valutazione del corpus e dello schema di annotazione stesso.*

1 Introduction

Hate is all but a new phenomenon, yet the global spread of Internet and social network services has provided it with new means and forms of dissemination. Online hateful content, or Hate Speech (HS), is characterised by some key aspects (such as virality, or presumed anonymity) which distinguish it from offline communication and make it potentially more dangerous and hurtful (Ziccardi, 2016). What is more, HS is featured as a complex and multifaceted phenomenon, also because of the variety of approaches employed in attempting to draw the line between HS and free speech (Yong, 2011). Therefore, despite the multiple efforts, there is yet no universally accepted definition of HS.

From a juridical perspective, two contrasting approaches can be recognised: while US law is oriented, quite uniquely, towards *granting freedom of speech* above all, even when potentially hurtful or threatening, legislation in Europe and the rest of the world tends to *protect the dignity and rights of minority groups* against any form of expression that might violate or endanger them.

Several European treaties and conventions ban HS: to mention but one, the Council of European Union condemns publicly inciting violence or hatred towards persons or groups defined by reference to race, colour, religion, descent or national or ethnic origin. The *No Hate Speech Movement*[1], promoted by the Council of Europe, is also worth-mentioning for its efforts in endorsing responsible behaviours and preventing HS among European citizens.

The main aim of this paper is at introducing a novel resource which can be useful for the investigation of HS in a sentiment analysis perspective (Schmidt and Wiegand, 2017). Providing that among the minority groups targeted by HS, the present socio-political context shows that some of them are especially vulnerable and garner constant attention - often negative - from the public opinion, i.e. immigrants (Bosco et al., 2017), Roma and Muslims, we decided to focus our work on HS against such groups. Furthermore, providing the spread of HS in social media together with their

[1] https://www.nohatespeechmovement.org

current relevance in communication, we focused on texts from Twitter, whose peculiar structure and conventions make it particularly suitable for data gathering and analysis.

2 Related Work

One of the earlier attempts to develop a corpus-based model for automated detection of HS on the Web is found in Warner and Hirschberg (2012): the authors collect and label a set of sentences from various websites, and test a classifier for detecting anti-Semitic hatred. They observe that HS against different groups is characterised by a small set of high frequency stereotypical words, also stressing the importance of distinguishing HS from simply offensive content.

The same distinction is at the core of Davidson et al. (2017), where a classifier is trained to recognise whether a tweet is hateful or just offensive, observing that for some categories this difference is less clear than for others.

An exhaustive list of the targets of online hate is found in Silva et al. (2016), where HS on two social networks (Twitter and Whisper) is detected through a sentence structure-based model.

One of the core issues of manually labelling HS is the reliability of annotations and the inter-annotator agreement. The issue is confronted by Waseem (2016) and Ross et al. (2017), who find that more precise results are obtained by relying on expert rather than amateur annotations, and that the overall reliability remains low. The authors suggest that HS should not be considered as a binary ”yes/no” value and that finer-grained labels may help increase the agreement rate.

An alternative to lexicon-based approaches is suggested in Saleem (2016), where limits and biases of manual annotation and keyword-based techniques are pointed out, and a method based on the language used within self-defined hateful web communities is presented. The method, suitable for various platforms, bypasses the need to define HS and the inevitable poor reliability of manual annotation.

While most of the available works are based on English language, Del Vigna et al. (2017) is the first work on a manually annotated Italian HS corpus: here the authors apply a traditional procedure on a corpus crawled from Facebook, developing two classifiers for automated detection of HS.

3 Dataset Collection

The dataset creation phase was divided into three main stages.

We first collected all the tweets written in Italian and posted from 1st October 2016 to 25th April 2017.

Then we discussed in order to establish *a*) which minority groups should be identified as possible HS targets, and *b*) the set of keywords associated with each target, in order to filter the data collected in the previous step. As for the first aspect, we identified three targets that we deemed particularly relevant in the present Italian scenario; based also on the terminology used in European Union reports[2], the targets selected for our corpus were immigrants (class: ethnic origin), Muslims (class: religion), and Roma. As regards the second aspect mentioned above, we are aware of the limits of a keyword-based method in HS identification (Saleem et al., 2016), especially regarding the amount of noisy data (e.g. off-topic tweets) that may result from such method; on the other hand, the choice to adopt a list of explicitly hateful words[3] may prevent us from finding subtler forms of HS, or even just tweets where a hateful message is expressed without using a hate-related lexicon. With this in mind, we then filtered the data by retaining a small set of neutral keywords associated with each target. The keywords selected are summarised below:

ethnic group	**religion**	**Roma**
immigrat*	terrorismo	rom
(immigrant*)	(terrorism)	(roma)
immigrazione	terrorist*	nomad*
(immigration)	(terrorist*)	(nomad*)
migrant*	islam	
stranier*	mussulman*	
(foreign)	(muslim*)	
profug*	corano	
(refugee*)	(koran)	

The dataset thus retrieved consisted of 370,252 tweets about ethnic origins, 176,290 about religion

[2] See the 2015 Eurobarometer Survey on discrimination in the EU: http://ec.europa.eu/justice/fundamental-rights/files/factsheet_eurobarometer_fundamental_rights_2015.pdf

[3] Such as the ones extracted for the Italian HS map (Musto et al., 2016): http://www.voxdiritti.it/ecco-la-nuova-edizione-della-mappa-dellintolleranza/

and 31,990 about Roma.

The last stage consisted in the creation of the corpus to be annotated. In order to obtain a balanced resource, we randomly selected from the previous dataset 700 tweets for each target, with a total amount of 2,100 tweets.
However, a large number of tweets were further removed from the corpus, during the annotation stage (because of duplicates and off-topic content). Despite the size reduction, though, the distribution of the targets in the corpus remained quite unchanged, resulting in a balanced resource in this respect.

At present, the amount of annotated data consists of 1,828 tweets. In the next section, we describe the whole annotation process and the scheme adopted for this purpose.

4 Data Annotation: Designing and Applying the Schema

Being HS a complex and multi-layered concept, and being the task of its annotation quite difficult and prone to subjectivity, we undertook some preliminary steps in order to make sure that all annotators share a common ground of basic concepts, starting from the very definition of HS.
When determining what can, or cannot, be considered HS (thus in a *yes-no* fashion), and based on the juridical literature and observations reported above in Section 1, we considered two different factors:

- the **target** involved, i.e. the tweet should be addressed, or just refer to, one of the minority groups identified as HS targets in the previous stage (see Section 3), or even to an individual considered for its membership in that category (and not for its individual characteristics);

- the **action**, or more precisely the illocutionary force of the utterance, in that it is capable of spreading, inciting, promoting or justifying violence against a target.

Whenever both factors happen to co-occur in the same tweet, we consider it as a HS case, as in the example below:

target	**tweet**
religion	*Ci vuole la guerra per salvare l'Italia dai criminali filo islamici* ("We need a war to save Italy from pro-Islamic criminals")

In case even just one of these conditions is not detected, HS is assumed not to occur.

In line with this definition, we also attempted to extend the scheme to other annotation categories that seemed to significantly co-occur with HS; this in order to better represent the (perceived) meaning of the tweet, and to help the annotator in the task, by providing a richer and finer-grained tagset[4]. The newly-introduced categories are described below.

Aggressiveness (labels *no - weak - strong*): it focuses on the user intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target; if the message reflects an overtly hostile attitude, or whenever the target group is portrayed as a threat to social stability, the tweet is considered *weakly* aggressive, while if there is the reference – whether explicit or just implied – to violent actions of any kind, the tweet is *strongly* aggressive.

tweet	**aggressiveness**
nuova invasione di migranti in Europa (A new migrant invasion in Europe)	weak
Cacciamo i rom dall'Italia (Let's kick Roma out of Italy)	strong

Offensiveness (labels *no - weak - strong*): conversely to aggressiveness, it rather focuses on the potentially hurtful effect of the tweet content on a given target. A tweet is considered *weakly* offensive in a large number of cases, among these: the given target is associated with typical human flaws (laziness in particular), the status of disadvantaged or discriminated minority is questioned, or when the members of the target group are described as unpleasant people; on the other hand, if an overtly insulting language is used, or the target is addressed to by means of outrageous or degrading expressions, the tweet is expected to be considered as *strongly* offensive.

[4] The whole scheme description along with the detailed guidelines are available at https://github.com/msang/hate-speech-corpus

tweet	offensiveness
I migranti sanno solo ostentare l'ozio (Migrants can only show off their laziness)	weak
Zingari di merda (You fucking Roma)	strong

Irony (labels *no* - *yes*): it determines whether the tweet is ironic or sarcastic rather than based on the literal meaning of words. The introduction of this category in the scheme was led by preliminary observations of the data, which highlighted how it was a fairly common linguistic expedient used to mitigate or indirectly convey a hateful content.

tweet	irony
ora tutti questi falsi profughi li mandiamo a casa di Renzi ??! (shall we send all these fake refugees to Renzi's house??!)	yes

Stereotype (labels *no* - *yes*): it determines whether the tweet contains any implicit or explicit reference to (mostly untrue) beliefs about a given target. There is a whole host of stereotypes and prejudices associated with the target groups selected for our research; from an exploratory observation of the data in the corpus, the following cases were identified: the members of a given target are referred to as invaders, freeloaders, criminals, filthy (or having filthy habits), sexist/mysoginist, undemocratic, violent people.
Furthermore, we also take into account the role that conventional media may have in spreading stereotypes and prejudices while reporting news on refugees, migrants, and minorities in general. Based on what suggested in the Italian journalists' Code of Conduct, known as "Carta di Roma"[5], in order to ensure a correct and responsible reporting about these topics, we also applied this criterion to any tweet containing a news headline that implicitly endorses, or contributes to the spread of, such stereotypical portrayals (see the example below).

tweet	stereotype
Roma in bancarotta ma regala 12 milioni ai rom (Rome is bankrupt but still gives 12 millions to Roma)	yes

Annotation process The annotation task consisted in a multiple-step process, and it was carried out by four independent annotators after a preliminary step where the guidelines were discussed and partially revised.
The corpus was split in two, and each part was annotated by two annotators. The annotator pairs then switched to the other part, in order to provide a third (possibly solving) annotation to all those tweets where at least one category was labelled differently by the previous two annotators. A further subset of around 130 tweets still received different labels by the different annotators (namely for aggressiveness and offensiveness). In order to solve these remaining cases, a fifth independent annotator was finally involved. As a result, the final corpus only contains tweets that were fully revised.

Regarding the results of the annotation in terms of label distribution, we found that 16% of all tweets have been considered containing HS, of which 23% against immigrants, 38% against Muslims and 39% against Roma. When considered alone, aggressiveness occurs in 14% , offensiveness in 10%, irony in 11% and stereotype in 29% of tweets. However, the labels that co-occur more frequently with hate speech are those indicating the presence of aggressiveness (81%), stereotypes (81%), and offensiveness (56%), and, overall, they co-occur altogether 52% of the times; irony is labelled in 11% of HS tweets. While, within the whole corpus, 57% of cases are just tweets with a "neutral" content, which means that no one of the categories were annotated as such.

4.1 Agreement Analysis

The development phase related to the inter-annotator agreement (IAA) is not only a necessary step for validating the corpus and evaluating the schema adopted, but also a tool that provides more details about the trends and biases of individual annotators with respect to specific annotation categories.

In this study, we measured the IAA right after the first annotation step was completed, i.e. the one where just two annotators were involved (see Section 4). In line with related cases[6], our data showed a very low agreement: in 47% of cases, the annotator pair annotated at least one of the five

[5] See `https://www.cartadiroma.org/`

[6] See (Del Vigna et al., 2017), (Gitari et al., 2015), (Kwok and Wang, 2013), (Ross et al., 2017), (Waseem, 2016), to mention a few.

categories using different labels. In fact, the disagreement took place mostly in one (40%) or two (16%) categories, while just 4 tweets received a completely different annotation by the annotator pairs. More specifically, we measured the agreement coefficient, using Cohen's kappa (Carletta, 1996), for each individual category. Results – also reported in Table 1 – show that the category with the highest agreement is namely the one related to the presence of hate speech (abbreviated to 'hs' in the table), followed by irony ('iro.') and stereotype ('ster.').

	hs	aggr.	off.	iro.	ster.
before merge	0,54	0,18	0,32	0,44	0,43
after merge	0,54	**0,43**	**0,37**	0,44	0,43

Table 1: Agreement (Cohen's k) for each annotation category before and after merging labels for aggressiveness and offensiveness.

Considering that the lowest agreement was found in aggressiveness ('aggr.') and offensiveness ('off.') – the only categories where three labels were used, instead of two – the agreement was recalculated by merging the *weak-strong* labels; it thus increased considerably (especially in aggressiveness), though still remaining far below an acceptable threshold.

The low agreement with regard to the degree of offensiveness can be attributed to the absence of clear indications within the annotation guidelines in this respect.

Finally, among the annotation criteria established in the preliminary stage, one in particular proved to be quite misleading, i.e. whenever a clearly hateful tweet did not actually refer to the target identified by one of the selected keywords, HS and stereotype were assumed not to occur. On the other hand, the remaining categories should be annotated accordingly. This principle was conceived in order to provide annotated data that could be considered a true reflection of HS towards the targets we identified in our study, though still "preserving" the meaning and the intent of the tweet in itself, regardless of the target involved. This, together with other points of the guidelines, will be further discussed and clarified in the next project phase.

5 Conclusion and Future Work

We introduced in this paper the collection and annotation of an Italian Twitter corpus representing HS towards some selected target. Our main aim is at producing a corpus to be used for training and testing sentiment analysis systems, but some effort must still be applied to achieve this goal. The current contribute is mainly in designing and trying a novel schema for HS, but the relatively low agreement shows that modelling this phenomenon is a very challenging task and a further refinement of the guidelines and of the scheme must be applied, together with the application to larger data sets.

Acknowledgments

The work of Cristina Bosco, Viviana Patti and Manuela Sanguinetti was partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, project S1618_L2_BOSC_01) and partially funded by Fondazione CRT (*Hate Speech and Social Media*, project n. 2016.0688).

References

Cristina Bosco, Patti Viviana, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Ruffo, Rossano Schifanella, and Marco Stranisci. 2017. Tools and resources for detecting hate and prejudice against immigrants in social media. In *Proceedings of First Symposium on Social Interactions in Complex Intelligent Systems (SICIS), AISB Convention 2017, AI and Society*, Bath, UK.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, June.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017.*, pages 86–95.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In Marie desJardins and Michael L. Littman, editors, *AAAI*. AAAI Press.

Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. 2016. Modeling community behavior through semantic analysis of social data: The italian hate map experience. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, pages 307–308.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *CoRR*, abs/1701.08118.

Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2016. A web of hate: Tackling hateful speech in online social spaces. In *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016)s*, Portoro, Slovenia.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrcio Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pages 687–690. AAAI Press.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November. Association for Computational Linguistics.

Caleb Yong. 2011. Does freedom of speech include hate speech? *Res Publica*, 17(4):385, Jul.

Giovanni Ziccardi. 2016. *L'odio online. Violenza verbale e ossessioni in rete*. Saggi / [Cortina]. Cortina Raffaello.

PoS Taggers in the Wild: A Case Study with Swiss Italian Student Essays

Daniele Puccinelli, Silvia Demartini, Aris Piatti, Sara Giulivi, Luca Cignetti, Simone Fornara
University of Applied Sciences and Arts of Southern Switzerland (SUPSI)
{daniele.puccinelli, silvia.demartini, aris.piatti,
{sara.giulivi, luca.cignetti, simone.fornara}@supsi.ch

Abstract

English. State-of-the-art Part-of-Speech taggers have been thoroughly evaluated on standard Italian. To understand how Part-of-Speech taggers that have been pre-trained on standard Italian fare with a wide array of language anomalies, we evaluate five Part-of-Speech taggers on a corpus of student essays written throughout the largest Italian-speaking area outside of Italy. Our preliminary results show that there is a significant gap between their performance on non-standard Italian and on standard Italian, and that the performance loss mainly comes from relatively subtle tagging errors within morphological categories as opposed to coarse errors across categories.

Italiano. *Gli strumenti di Part-of-Speech tagging più rappresentativi dello stato dell'arte sono stati analizzati a fondo con l'italiano standard. Per capire come strumenti pre-addestrati sull'italiano standard si comportano in presenza di un'ampia gamma di anomalie linguistiche, analizziamo le prestazioni di cinque strumenti su di un corpus di elaborati redatti da studenti della scuola dell'obbligo nella Svizzera Italiana. I nostri risultati preliminari mostrano che esiste un notevole divario tra le prestazioni sull'italiano non-standard e quelle sull'italiano standard, e che la perdita di prestazioni deriva principalmente da errori di tagging relativamente sottili all'interno delle categorie grammaticali.*

1 Introduction

The goal of this paper is to present the preliminary results of the evaluation of a set of state-of-the-art Part of Speech (PoS) taggers on the DFA-TIscrivo corpus of Italian-language (L1) K-12 student essays from schools in the Italian-speaking part of Switzerland. The DFA-TIscrivo corpus represents an example of non-standard Italian[1] because its contributors are young students with a poor command of the Italian language living in the largest Italian-speaking area outside of Italy, and therefore prone to regionalisms as well as orthographic mistakes.

The key research question at this stage is how well state-of-the-art PoS taggers that were pre-trained on standard Italian cope with a specific flavor of non-standard Italian. It would of course be possible to retrain all these tools on texts with similar properties as the ones in our corpus, but at this stage in our work this is not possible due to the overly small size of the available annotated data. In turn, using pre-trained models gives us a twofold advantage: it allows us to obtain a performance baseline on non-standard Italian, and it makes it possible to directly compare our performance metrics to previously published results (obtained with the same models we use). While our work is still in progress and the results reported herein are preliminary in nature, we can already share several notable observations.

2 Related Work and PoS taggers under test

There have been various recent efforts focused on social media within the scope of EVALITA 2016 (Bosco et al., 2016), whose goal was the domain adaptation of PoS-taggers to Twitter texts. Notable contributions include (Cimino and Dell'Orletta, 2016), whose authors propose a PoS tagging architecture optimized to process Italian-language tweets. While we do acknowledge the need for

[1] http://www.treccani.it/enciclopedia/italiano-standard_(Enciclopedia-dell%27Italiano)/

domain adaptation with non-standard texts, we ask a more basic question: if we perform no domain adaptation and simply deploy general-purpose PoS taggers *in the wild*, how do they fare? We use K-12 student essays as our flavor of non-standard Italian. Although such texts are beset with all sorts of anomalies, they can still be processed them with general purpose taggers, unlike far more unstructured and unconventional texts such as tweets. While similar studies have been conducted for other languages, such as German (Giesbrecht and Evert, 2009), to the best of our knowledge this is the first study of the accuracy of general-purpose PoS taggers *in the wild* for the Italian language. Our selection of state-of-the-art general purpose PoS taggers is based on their popularity with the research community and the availability of ready-to-use software versions.

TreeTagger (1994). The popular TreeTagger (Schmid, 1994) tool uses decision trees to estimate transition probabilities based on context. Decision trees were extremely popular for PoS tagging in the 1990s, when more sophisticated machine learning tools such as neural networks were still too computationally demanding given the relatively limited resources available at the time. TreeTagger actively addresses the issues encountered by earlier probabilistic PoS taggers with rare words with a very low (but non-zero) probability of occurrence. The use of decision trees enables TreeTagger to account for context, whose nature is not restricted to $n-$grams, but also to allowed/disallowed tag sequences.

UD-Pipe (2014). UD-Pipe (Straka et al., 2016) is a language-agnostic natural language processing (NLP) pipeline developed within Universal Dependencies, whose focus is the development of a treebank annotation scheme that can work consistently across multiple languages. UD-Pipe's PoS tagger uses the Morphological Dictionary and Tagger MorphoDiTa (Straková et al., 2014), developed at Charles University in Prague, Czech Republic. MorphoDiTa uses the averaged perceptron PoS tagger described in (Spoustová et al., 2009) and based on (Collins, 2002).

Tint (2016). *The Italian NLP Tool* (Palmero Aprosio and Moretti, 2016) is an NLP pipeline for the Italian language based on Stanford CoreNLP (Manning et al., 2014). Tint's PoS tagger is based on the Stanford Log-linear Tagger (Toutanova et al., 2003), which leverages maximum entropy PoS tagging (Toutanova and Manning, 2000). Given a word and its context (other words in the sentence and their tags), maximum entropy PoS tagging assigns a probability to every tag in a predefined tagset, eventually enabling the estimation of the probability of a tag sequence given a word sequence. Out of all the possible distributions that satisfy a set of constraints, the one with maximum entropy is chosen, as it represents the most non-committal assignment of probabilities that meets the constraints (Ratnaparkhi, 1996).

Syntaxnet (2016). Various recent efforts focus on the application of recurrent neural networks to PoS tagging and dependency parsing (Ling et al., 2015), but it is shown in (Andor et al., 2016) that recurrence-free feed-forward networks can work at least as well as recurrent ones if they are globally normalized; this is the guiding principle behind PoS tagging in Syntaxnet (syn, 2016), a neural network NLP framework that is built on top of Google's popular TensorFlow machine learning framework (Abadi et al., 2016). Syntaxnet employs beam search, which serves to maintain multiple hypotheses, and global normalization with a conditional random field (CRF) objective, which avoids label bias issues (typically reported in locally normalized models). PoS tagging in Syntaxnet is heavily inspired by (Bohnet and Nivre, 2012) and relies on the close integration of PoS tagging and dependency parsing. A pre-trained English language model whimsically called *Parsey McParseface* was released along with Syntaxnet in May 2016 and a pre-trained model for the Italian language was released in August 2016 as one of *Parsey's Cousins*.

DRAGNN (2017). In March 2017 Google released a Syntaxnet upgrade based on Dynamic Recurrent Acyclic Graphical Neural Networks (DRAGNN) (Kong et al., 2017) along with the *Parseysaurus* set of pre-trained models (Alberti et al., 2017) that was developed for the CONLL 2017 shared task. PoS tagging in DRAGNN (Kong et al., 2017) is based on (Zhang and Weiss, 2016), which closely integrates PoS tagging and parsing in a novel fashion (specifically, the continuous hidden layer activations of the window-based tagger network are fed as input to the transition-based parser network). The tagger works token by token, extracting features from a window of tokens around the target token. It has a fairly standard structure with embedding, hidden, and softmax layers.

3 The DFA-TIscrivo corpus

The DFA-TIscrivo corpus has been prepared within the projects TIscrivo (2011-2014) and TIscrivo 2.0 (2014-2017) projects[2], both funded by the Swiss National Science Foundation. The goal of the projects is to paint an accurate picture of the writing skills of primary school and lower secondary school in Southern Switzerland in order to describe the variety of language written at school and to propose new teaching practices to improve writing skills in compulsory education (Cignetti et al., 2016). Other studies with some similarities to the TIscrivo projects include projects focused on texts by L1 or L2 learners such as ISACCO (Brunato and Dell'Orletta, 2015), CItA (Barbagli et al., 2015)(Barbagli et al., 2016), and KoKo (Abel et al., 2016).

The DFA-TIscrivo corpus is a balanced corpus collected in 56 Italian-speaking primary and lower secondary schools from Southern Switzerland. It contains 1735 narrative-reflective essays (742 from primary, 993 from secondary school), transcribed but not normalized, and accompanied by sociolinguistic metadata (age, gender, school and class, linguistic information). It amounts to about 390,000 tokens. Lexical data were initially lemmatized and PoS tagged using TreeTagger (with the Italian parameters by Marco Baroni) and are being manually revised. Furthermore, we are manually annotating orthographic, morphological and lexical main types of error, multi-word expressions, peculiar lexicon of Italian only used in Southern Switzerland and foreign words. A key project goal is to build up a dictionary of the Italian language as it is written in Southern Switzerland (Cignetti and Demartini, 2016)(Fornara et al., 2016) as an online resource useful both to scholars and to teachers.

4 Methodology and Performance Analysis

We run the five taggers on the corpus and compare their output to a manually tagged ground truth. We note that, at the time of writing, the analysis is restricted to a subset of the DFA-TIscrivo corpus that has been manually PoS-tagged and is limited to essays written by fifth graders. We use the ISST-TANL-PoS reference tagset[3] based on Universal Dependencies.

	Accuracy
TreeTagger	0.84
UD-Pipe	0.79
Tint	0.83
Syntaxnet	0.83
DRAGNN	0.84

Table 1: Overall PoS tagging accuracy for each tool on the DFA-TIscrivo corpus.

We begin by assessing the tagging accuracy of the five PoS taggers under test on the DFA-TIscrivo corpus. We compute the tagging accuracy as the ratio of correctly tagged parts of speech with respect to the aforementioned manually tagged ground truth. While the ground truth isolates out multiword expressions, none of the tools are able to do that, so all multiword expressions are considered to be mistagged and every multiword expression counts as one single miss. Verbal enclitics are not considered and the corresponding verbs are expected to be tagged simply as verbs. Our results are shown in Table 1; we see that UD-Pipe trails behind and falls below the 0.8 mark, while the other four taggers under test offer a similar performance, with TreeTagger slightly ahead of the pack. All these taggers reportedly perform above the 95% mark on standard Italian.

Tables 2-6 contain the confusion matrices of the PoS taggers under test based on the ISST-TANL coarse-grained tags. Row i shows the ground truth for tag i and column k shows the frequency with which it is tagged as k. To abstract away from how individual taggers address prepositional article, we merge the tags for prepositions (E) and articles (R) into a super-tag ER. We also merge the tags for adjectives (A) and determiners (D) because determiners may be viewed as a category of adjectives in Italian. We only show the tags that occur most often (which is why some rows/columns do not add up to one). We note that TreeTagger outperforms all other taggers with AD while lagging behind all of them with P (pronouns) and C (conjunctions), often tagged as P or B (adverbs). TreeTaggers also performs remarkably well with verbs (V).

[2] http://dfa-blog.supsi.ch/DFA-TIscrivo~/la-ricerca/

[3] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

	AD	**B**	**C**	**ER**	**P**	**S**	**V**
AD	0.95	0.03	0.01	0	0	0	0
B	0.02	0.88	0	0	0.02	0.04	0.04
C	0.01	0.07	0.76	0.01	0.15	0	0
ER	0.08	0	0	0.92	0	0	0
P	0.05	0	0.01	0	0.79	0.01	0
S	0.04	0	0	0	0	0.93	0.03
V	0	0	0	0	0	0.01	0.99

Table 2: Tree Tagger confusion matrix.

	AD	**B**	**C**	**ER**	**P**	**S**	**V**
AD	0.77	0.04	0	0	0	0.04	0.01
B	0.04	0.86	0	0.01	0	0.07	0.02
C	0	0.06	0.91	0.02	0	0.01	0
ER	0	0	0	1	0	0	0
P	0.01	0.01	0.02	0.02	0.93	0.01	0
S	0.02	0.01	0	0	0	0.96	0.01
V	0.02	0	0	0	0.01	0.03	0.94

Table 3: UD-Pipe confusion matrix.

	AD	**B**	**C**	**ER**	**P**	**S**	**V**
AD	0.75	0	0	0	0.15	0	0
B	0.06	0.88	0	0.01	0	0.03	0.02
C	0	0.09	0.89	0.01	0	0	0
ER	0	0	0	0.99	0	0	0
P	0.02	0	0	0.03	0.88	0.01	0
S	0.03	0	0	0	0	0.94	0.03
V	0.01	0	0	0	0	0.01	0.92

Table 4: Tint confusion matrix.

We have also studied the confusion matrices within the V category (not shown), noting that TreeTagger performs remarkably better than the others with respect to principal verbs (0.97 accuracy while the others are right around the 0.9 mark). and modal verbs (0.94 versus 0.81 for UD-Pipe and TINT and a disappointing 0.75 for both Syntaxnet and DRAGNN). All taggers perform equally poorly with auxiliary verbs (accuracy just above the 0.8 mark in all cases). Aside from Tint, which does not provide morphological information (at least in the version we used), all taggers do well with finite verbs (> 0.97, with UD-Pipe trailing behind at 0.95). While TreeTagger and UD-Pipe perform at the same level of accuracy for both finite and non-finite verbs, Syntaxnet and DRAGNN barely go beyond the 0.9 mark with the latter.

	AD	**B**	**C**	**ER**	**P**	**S**	**V**
AD	0.75	0.01	0	0	0.03	0.05	0.02
B	0.01	0.88	0	0.05	0.01	0.02	0.03
C	0	0.08	0.9	0.01	0	0	0.01
ER	0.04	0	0	0.92	0	0.02	0.02
P	0.01	0.01	0.03	0.03	0.91	0.01	0
S	0.03	0.01	0	0	0	0.94	0.02
V	0.01	0	0	0	0	0.03	0.96

Table 5: Syntaxnet confusion matrix.

	AD	**B**	**C**	**ER**	**P**	**S**	**V**
AD	0.72	0.03	0	0	0.02	0.07	0.01
B	0.03	0.87	0	0.01	0	0.04	0.01
C	0	0.08	0.90	0.01	0.01	0	0
ER	0.06	0	0	0.92	0	0.01	0.01
P	0.01	0	0.02	0.02	0.93	0.01	0
S	0	0	0	0	0	0.97	0.02
V	0.01	0	0	0	0	0.04	0.95

Table 6: DRAGNN confusion matrix.

5 Conclusion

We have presented a comparative performance assessment of five state-of-the-art PoS taggers on the DFA-TIscrivo corpus of K-12 student essays, along with an analysis of the patterns that can be observed in the mistakes made by individual taggers. As this is still a work in progress, the results in the paper are limited to a subset of the corpus containing fifth grade essays. These results provide a valuable baseline that could likely be improved with domain adaptation. On the other hand, it is fair to ask whether the DFA-TIscrivo corpus is different enough from standard Italian to warrant domain adaptation, or whether we would encounter issues with overfitting. In the latter case, an alternative would be the rule-based combination of the output of the five taggers, informed with the knowledge of the observed error patterns.

Acknowledgments

The partial support of SNF through project TIscrivo 2.0 and of SUPSI through project Scripsit is gratefully acknowledged.

References

Abadi et al., 2016 Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467.

Abel et al., 2016 Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2016. An extended version of the koko german L1 learner corpus. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Alberti et al., 2017 Chris Alberti, Daniel Andor, Ivan Bogatyy, Michael Collins, Dan Gillick, Lingpeng Kong, Terry Koo, Ji Ma, Mark Omernick, Slav Petrov, Chayut Thanapirom, Zora Tung, and David Weiss. 2017. Syntaxnet models for the conll 2017 shared task. *CoRR*, abs/1703.04929.

Andor et al., 2016 Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Barbagli et al., 2015 Alessia Barbagli, Piero Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2015. Cita: un corpus di produzioni scritte di apprendenti litaliano l1 annotato con errori. In *BProceedings of the Second Italian Conference on Computational Linguistics, CLiC-it*, pages 31–35.

Barbagli et al., 2016 Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. Cita: an l1 italian learners corpus to study the development of writing competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may.

Bohnet and Nivre, 2012 Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1455–1465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bosco et al., 2016 Cristina Bosco, Fabio Tamburini, Andrea Bolioli, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 part of speech on twitter for italian task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Brunato and Dell'Orletta, 2015 Dominique Brunato and Felice Dell'Orletta. 2015. Isacco: a corpus for investigating spoken and written language development in italian school–age children. *CLiC it*, page 62.

Cignetti and Demartini, 2016 Luca Cignetti and Silvia Demartini. 2016. From data to tools. theoretical and applied problems in the compilation of lissics (the lexicon of written italian in a school context in italian switzerland). *RiCOGNIZIONI. Rivista di Lingue e Letterature straniere e Culture moderne*, 3(6):35–49.

Cignetti et al., 2016 Luca Cignetti, Silvia Demartini, and Simone Fornara. 2016. *Come TIscrivo? La scrittura a scuola tra teoria e didattica*. Aracne.

Cimino and Dell'Orletta, 2016 Andrea Cimino and Felice Dell'Orletta. 2016. Building the state-of-the-art in POS tagging of italian tweets. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Collins, 2002 Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fornara et al., 2016 Simone Fornara, Luca Cignetti, and Silvia Demartini. 2016. Il lessico di tiscrivo. caratterizzazione del vocabolario e osservazioni in prospettiva didattica. In *Sviluppo della competenza lessicale: acquisizione, apprendimento, insegnamento*, pages 43–60. Aracne, Roma.

Giesbrecht and Evert, 2009 Eugenie Giesbrecht and Stefan Evert. 2009. Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. In I. Alegria, I. Leturia, and S. Sharoff, editors, *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain.

Kong et al., 2017 Lingpeng Kong, Chris Alberti, Daniel Andor, Ivan Bogatyy, and David Weiss. 2017. DRAGNN: A transition-based framework for dynamically connected neural networks. *CoRR*, abs/1703.04474.

Ling et al., 2015 Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015.

Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*.

Manning et al., 2014 Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Palmero Aprosio and Moretti, 2016 A. Palmero Aprosio and G. Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints*, September.

Ratnaparkhi, 1996 Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

Schmid, 1994 Helmut Schmid. 1994. Part-of-speech tagging with neural networks. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 172–176. Association for Computational Linguistics.

Spoustová et al., 2009 Drahomíra "johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron pos tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece, March. Association for Computational Linguistics.

Straka et al., 2016 Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *LREC*.

Straková et al., 2014 Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics.

syn, 2016 2016. Syntaxnet. `https://www.tensorflow.org/versions/r0.12/tutorials/syntaxnet`. Accessed: 2017-07-13.

Toutanova and Manning, 2000 Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, pages 63–70. Association for Computational Linguistics.

Toutanova et al., 2003 Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Zhang and Weiss, 2016 Yuan Zhang and David Weiss. 2016. Stack-propagation: Improved representation learning for syntax. *arXiv preprint arXiv:1603.06598*.

Find Problems before They Find You with AnnotatorPro's Monitoring Functionalities

Mohammed R. H. Qwaider, Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini
Fondazione Bruno Kessler, Trento, Italy
{qwaider,minard,manspera,magnini}@fbk.eu

Abstract

English. We present a tool for annotation of linguistic data. ANNOTATORPRO offers both complete monitoring functionalities (e.g. inter-annotator agreement and agreement with respect to a gold standard) and highly flexible task design (e.g. token and document level annotation, adjudication and reconciliation procedures). We teste ANNOTATORPRO in several industrial annotation scenarios, coupled with Active Learning techniques.

Italiano. *Presentiamo uno strumento per l'annotazione di dati linguistici. AnnotatorPro offre sia complete funzionalità di monitoraggio (es. accordo tra annotatori, accordo rispetto ad un gold standard), sia la alta flessibilità nel definire task di annotazione (per esempio, annotazione per parole o per documento, procedure di aggiudicamento e re-conciliazione). AnnotatorPro è stato sperimentato in diversi scenari di annotazione industriali, accoppiato con tecniche di Active Learning.*

1 Introduction

Driven by the popularity of machine learning approaches, there has been in the last years an increasing need to produce human annotated data for a large number of linguistic tasks (e.g. named entity recognition, semantic role labeling, sentiment analysis, word sense disambiguation, and discourse relations, just to mention a few). Datasets (development, training and test data) are being developed for different languages and different domains, both for research and industrial purposes.

A relevant consequence of this is the increasing demand for annotated datasets, both in terms of quantity and quality. This in turn calls for tools with a rich apparatus of functionalities (e.g. annotation, visualization, monitoring and reporting), able to support and monitor a large variety of annotators (i.e. from linguists to mechanical turkers), flexible enough to serve a large spectrum of annotation scenarios (e.g. crowdsourcing and paid professional annotators), and open to the integration of NLP tools (e.g. for automatic pre-annotation and for instance selection based on Active Learning).

Although there is a large supply of annotation tools, such as *brat* (Stenetorp et al., 2012), *GATE* (Cunningham et al., 2011), *CAT* (Bartalesi Lenzi et al., 2012), and *WebAnno* (Yimam et al., 2013), and several functions are included in common crowdsourcing platforms (e.g. *CrowdFlower*[1]), we believe that none of the available tool possesses the full range of functionalities for a real and intensive industrial use. As an example, none of the afore mentioned tools allows one to implement adjudication rules (i.e. under what condition an item annotated by more than one annotator is assigned to a certain category) or to visualize items with disagreement among annotators.

This paper introduces ANNOTATORPRO, a new annotation tool which was mainly conceived to fulfill the above-mentioned needs. We highlight two main aspects of the tool: (i) a high level of flexibility to design the annotation task, including the possibility to define adjudication and reconciliation procedures; (ii) the rich set of functionalities allowing for constant monitoring of the quality of the data being annotated.

The paper is organized as follows. In Section 2 we compare ANNOTATORPRO with some state-of-the-art annotation tools. Section 3 provides a general description of the tool. Sections 4 and 5 focus on the task design and on the monitoring functionalities, while Section 6 provides a brief overview of the tool's application and future extensions.

[1] https://www.crowdflower.com

2 Related Work

Many annotation tools are available to the community. However, some of them are limited by license, e.g. *CAT* (Bartalesi Lenzi et al., 2012) and *GATE* (Cunningham et al., 2011) are available for research use only, while some others have open licenses, e.g. *brat* (Stenetorp et al., 2012), but offer limited features.

The *brat rapid annotation tool* (*brat*) is an open license annotation tool that supports different annotation levels, in particular annotation at the token level and annotation of relations between marked tokens. It supports multiple annotators, in the sense that many annotators can collaborate on annotating the same corpus, but needs an in-house installation. Despite all these advantages, *brat* does not support either annotation monitoring or annotator/task reports.

Other tools (e.g. *CAT*) provide advanced functionalities to perform annotation at different levels (e.g. token and relation level) through a user-friendly interface, although they do not support annotation monitoring.

CrowdFlower is an outsourcing annotation service that provides a platform for annotation (focusing on annotation at the document level) employing non expert contributors. It uses gold standard tests to evaluate the annotators and supports automatic adjudication features, but no inter-annotator agreement metrics are available. In addition an important issue which could limit the use of outsourcing is the non in-house storage of the data, in particular when sensitive data covered by privacy regulations are concerned.

GATE is a powerful tool that implements most of the features to facilitate the annotation production in all its phases (e.g. task creation, annotator assignment, annotation monitoring and multi-layer annotation of the same corpus). However, visualization of disagreement is not available and no automatic adjudication is available.

3 Overall Description

ANNOTATORPRO is a web-based annotation tool built on top of the open source tool MT-EQUAL (Machine Translation Error Quality Alignment), a toolkit for the manual assessment of Machine Translation output that implements three different tasks in an integrated environment: annotation of translation errors, translation quality rating (e.g. adequacy and fluency, relative ranking of alternative translations), and word alignment (Girardi et al., 2014).

ANNOTATORPRO inherits from MT-EQUAL the capability of scaling over big data in an optimized platform that is able to save annotation in real-time. It also makes use of the MT-EQUAL web-based interface which is a multi-user and user-friendly interface.

It performs simple tokenization based on spaces, punctuation, and other language-dependent rules, but the user can also upload directly tokenized files.

We designed new functionalities to fulfill the requirements of high quality corpus annotation performed by multiple annotators. ANNOTATORPRO's main novel features are:

- The interface includes different options to design the annotation task (Section 4.1), which are set by the project manager.
- The tool enables annotation at two levels (Section 4.2): annotation at the token level (e.g. part-of-speech tagging and named entity recognition) and annotation at the document level (e.g. sentiment analysis).
- ANNOTATORPRO's interface offers functionalities for annotation monitoring (Section 5), which include inter-annotator agreement (IAA) monitoring and quality monitoring.

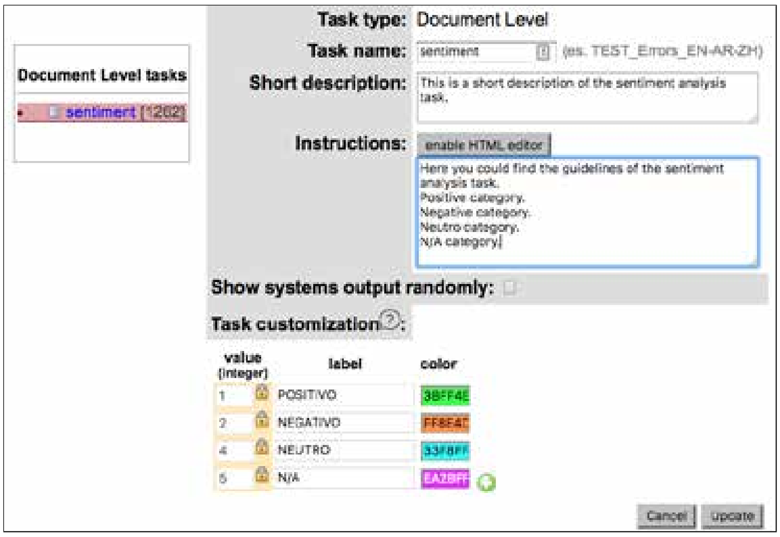
ANNOTATORPRO has been implemented in PHP and JavaScript, and uses MySQL to manage a database. It takes as input several UTF-8 encoded formats: TXT (raw text), IOB2[2] and TSV (tab separated values). It also accepts ZIP archives containing the source files.

As regards data storage, document's annotations are saved in a MySQL database in real time (i.e. while data being annotated). The annotated data can be exported in the following formats: IOB2 and TSV.

4 Annotation Task Design

ANNOTATORPRO distinguishes two types of users, i.e. managers and annotators. Managers

[2] The IOB2 tagging format is a common format for text chunking. B- is used to tag the beginning of a chunk, I- to tag tokens inside the chunk and O to indicate tokens not belonging to a chunk.



Document Level tasks

sentiment [1202]

Task type: Document Level

Task name: sentiment (es. TEST_Errors_EN-AR-ZH)

Short description: This is a short description of the sentiment analysis task.

Instructions:

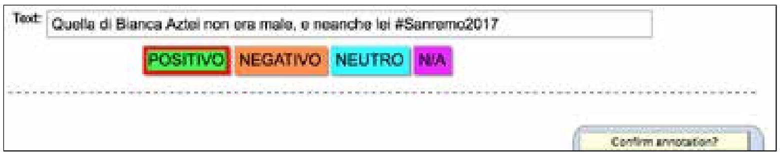
Here you could find the guidelines of the sentiment analysis task.
Positive category.
Negative category.
Neutro category.
N/A category.

Show systems output randomly: □

Task customization:

value (integer)	label	color
1	POSITIVO	3BFF4E
2	NEGATIVO	FF8E4C
4	NEUTRO	33F8FF
5	N/A	EA2BFF

Figure 1: Annotator's task definition: annotation level, task's name, task description, and annotation categories.



Text: Quella di Bianca Aztei non era male, e neanche lei #Sanremo2017

Figure 2: An example annotation interface: sentiment annotation of tweets.

take care of designing the annotation task at hand; in particular, they (i) define the annotation procedure, which depends on the number of annotators, their level of expertise (for example, non-expert annotators might not be allowed to see/modify each other's work) and the use that the dataset is intended for (e.g. evaluation, training, etc.), and (ii) the annotator's task, which includes selecting the most appropriate annotation level and creating the annotation categories/labels (Figure 1). As opposed to managers, annotators are basic users, who only have access to a limited number of (annotation) functionalities (Figure 2).

4.1 Annotation Procedure

One of the main tasks of the manager is to define the annotation procedure, which consists mainly of:

- Defining the number of annotators (one or more) who can collaborate on annotating the same corpus.
- In case of multiple annotators, defining the type of collaboration among them, i.e. whether data are to be annotated only by one or more of them (document level only).
- Defining the automatic adjudication rules in the case where multiple annotations of the same data are collected (document level only). The two basic options are:
  - considering an annotation as solved if the majority of annotators agreed on a certain annotation;
  - considering an annotation as solved if a minimum number of concordant annotations is reached.
- Deciding whether to make the metadata of the documents (e.g. document id, document title) visible to the annotators during the annotation phase.
- Deciding whether to allow for a revision phase after the annotation has been concluded, i.e. give the annotators the possibility to modify their annotations, for example after a reconciliation step has taken place. By default, document metadata will be visible during the revision phase to facilitate the work.
- Decide the modality for the selection of data to be presented to the annotators:

- propose to the annotator preselected ordered documents (default option);
- randomly select documents from a large dataset;
- select documents from a large dataset through an Active Learning process.[3]

4.2 Annotator's Task

ANNOTATORPRO supports two different annotation levels, i.e one where annotation is performed at the document level and one where we have smaller units, typically tokens, being annotated. It is the manager's task to select the most appropriate annotation level for the task at hand; for example, named entity recognition needs data annotated at the token level, whereas for sentiment analysis a corpus is generally annotated at the document level.

Finally, the task manager defines the set of categories or the set of labels to be used by the annotator respectively to classify the documents (in the case of document level annotation) or to mark portion of text.

5 Annotation Monitoring

In ANNOTATORPRO we have implemented several monitoring functionalities aimed at guaranteeing high quality annotation as described below.

5.1 Progress Monitoring

From the manager interface two tabs display information about the annotations already performed. The **Annotation** tab presents the progress of the annotation task, i.e. the annotations done by each annotator. This is real-time information, which means that the manager can follow the progress of the work underway. Moreover the manager can visualize the annotations of each user in read-only mode.

The **Overall stats** panel displays a table which summarizes the overall statistics about the annotation. The following information is given: total number of annotated documents; number of non-annotated documents; number of partially annotated documents (i.e. documents not yet annotated by the required number of annotators); number of completely annotated documents (i.e. documents annotated by the required number of annotators, independently of whether annotators did or did not reach an agreement).

5.2 Inter-Annotator Agreement Monitoring

IAA monitoring, which measures the level of agreement between the annotators at regular intervals, is activated every time two or more annotators annotate the same data.

IAA agreement is computed in terms of Dice coefficient (Lin, 1998) and Cohen's Kappa (Viera and Garrett, 2005); the latter represents the agreement as a continuous value from -1 to 1, where -1 means total disagreement and 1 means total agreement.

The project manager has access to different types of information to constantly monitor the level of agreement between annotators, focusing both on a single annotator and overall:

- the level of agreement each annotator obtains with every other annotator and the average of the IAA values obtained by each annotator;
- the overall average IAA.

ANNOTATORPRO also provides a visualization of the annotations made by each annotator for each document, where a different color is used to present each tag from the tagset (see Figure 3). This enables the manager to have quick and easy access to the cases of disagreement and, if needed, to give feedback to the annotators.

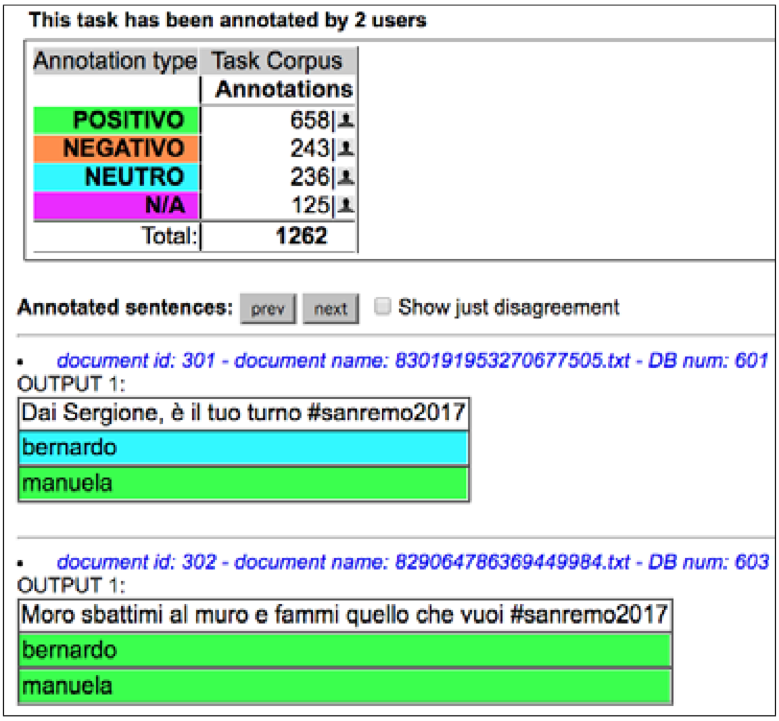
5.3 Quality Monitoring

Quality monitoring makes use of a gold standard dataset previously annotated by an expert. Each annotator is asked to provide an annotation for those samples. The annotators do not know if they are annotating a golden sample or not, which ensures a non-biased evaluation. This enables the project manager to assess the quality of the annotations of each annotator by comparing them against a dataset considered correct. The same quantitative information and visualization as those for IAA monitoring (see Section 5.2) are available.

6 Applications and Further Extensions

We used ANNOTATORPRO for multiple projects, on different tasks, including named entity recognition (Minard et al., 2016a), event detection (Minard et al., 2016b) and sentiment analysis. The

[3] The Active Learning process is not provided in the distribution of ANNOTATORPRO, but the tool can select the data to be annotated if they are associated with a confidence value (in this case the tool can either select those with the highest score or those with the lowest score).



This task has been annotated by 2 users

Annotation type	Task Corpus Annotations
POSITIVO	658
NEGATIVO	243
NEUTRO	236
N/A	125
Total:	1262

- document id: 301 - document name: 830191953270677505.txt - DB num: 601

OUTPUT 1:

Dai Sergione, è il tuo turno #sanremo2017

bernardo

manuela

- document id: 302 - document name: 829064786369449984.txt - DB num: 603

OUTPUT 1:

Moro sbattimi al muro e fammi quello che vuoi #sanremo2017

bernardo

manuela

Figure 3: Visualization of the annotations made for two documents. The first example is a case of disagreement and the second a case of agreement. At the top of the page is given the number of annotations for each tag.

tool has been successfully exploited both in situations with few experienced annotators as well as with more than 20 non-expert annotators (i.e. high school students) working in parallel. ANNOTATORPRO has been fully integrated within an Active Learning platform (Magnini et al., 2016) and successfully employed in two industrial projects, resulting in high quality data.

As for our next steps, we are working to extend ANNOTATORPRO to include relations among annotated entities, such as the relation between a verb and its argument/s in semantic role labeling.

ANNOTATORPRO is distributed as open source software under the terms of Apache License 2.0.[4] from the web page: `http://hlt-nlp.fbk.eu/technologies/annotatorpro`.

Acknowledgments

This work has been partially funded by the Euclip-Res project, under the program *Bando Innovazione 2016* of the autonomous Province of Bolzano.

References

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT annotation tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 333–338, Istanbul, Turkey, May 23-25, 2012.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*. University of Sheffield Department of Computer Science.

Christian Girardi, Luisa Bentivogli, Mohammad Amin Farajian, and Marcello Federico. 2014. MT-EQuAl: A toolkit for human assessment of machine translation output. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 120–123, Dublin, Ireland, August 23-29, 2014. ACL.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, Madison, Wisconsin, USA. Morgan Kaufmann Publishers Inc.

Bernardo Magnini, Anne-Lyse Minard, Mohammed R. H. Qwaider, and Manuela Speranza. 2016.

[4] `https://www.apache.org/licenses/LICENSE-2.0`

TextPro-AL: An active learning platform for flexible and efficient production of training data for NLP tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 131–135, Osaka, Japan, December.

Anne-Lyse Minard, Mohammed R. H. Qwaider, and Bernardo Magnini. 2016a. FBK-NLP at NEEL-IT: Active learning for domain adaptation. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, volume 1749, Napoli, Italy, December 5-7, 2016.

Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini, and Mohammed R. H. Qwaider. 2016b. Semantic interpretation of events in live soccer commentaries. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Avignon, France. Association for Computational Linguistics.

Anthony J. Viera and Joanne M. Garrett. 2005. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363, 5.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.

Evaluating a rule based strategy to map IMAGACT and T-PAS

Andrea Amelio Ravelli
Università di Firenze, Italy
andreaamelio.ravelli
@unifi.it

Lorenzo Gregori
Università di Firenze, Italy
lorenzo.gregori
@unifi.it

Anna Feltracco
Fondazione Bruno Kessler
Università di Pavia, Italy
Università di Bergamo, Italy
feltracco@fbk.eu

Abstract

English. This paper presents the analysis of a mapping between two resources, IMAGACT and T-PAS, made through a rule-based algorithm which converts argument structures in thematic roles. Results are good in terms of Recall, while Precision values are low: an analysis of the causes is proposed.

Italiano. *Questo articolo presenta l'analisi di un mapping tra le risorse IMAGACT e T-PAS, realizzato attraverso un algoritmo basato su regole che converte le strutture argomentali in ruoli tematici. I risultati sono buoni in termini di Recall, mentre sono bassi i valori di Precision per i quali viene proposta un'analisi.*

1 Introduction

The automatic mapping of information between two resources is not a trivial task, but indeed joining information over specific data can benefit the involved resources. This paper describes the analysis of a mapping between two linguistic resources: IMAGACT and T-PAS. The motivation behind this mapping starts with the observation that both resources deal with Italian verbs disambiguation, are corpus-based and contain pieces of information that can be integrated with each other.

IMAGACT is a linguistic ontology of *actions*, that are grouped in *concepts* and related to different verb *Types*. For example, the action "John takes the cup from the shelf" belongs to the concept *"take an object"* and refers to Type 3 of the verb *to take*. Each Type is also associated to one or more thematic structures (e.g. [AGENT-verb-THEME-SOURCE]) and to videos via a set of captions.

T-PAS is a repository of *argument typed structures* for Italian verbs. Each verb is listed with its structures, which correspond to different senses of the verb. For each structure, the specification of the expected *semantic type* in every argument position (e.g. for the *subject*) is provided.

In this paper, we describe the results of a first attempt of mapping information between these resources. Specifically, for each of the 248 verbs analysed in both resources, we aim at matching the IMAGACT Types with the corresponding typed argument structures in T-PAS. We operate this mapping by applying a set of rules which convert the information from the argument structure into a thematic-role combination, and find all the Types that match this combination.

The linking between argument and thematic structures of a predicate is a debated complex task in linguistic theories (Baker, 1997; Pinker, 2009; Bowerman, 1990, among others). The predictability of thematic roles from argument structure (or viceversa) belongs to the *syntax-semantics interface*, and a study in this direction is out of the scope of this paper. Our experiment is focused on an empirical analysis of argument and thematic structures in Italian verbs and our aim is to evaluate whether, and to which extent, a rule-based system is able to produce thematic structures. We also intend to verify how these results can be exploited for a mapping purpose.

The paper is structured as follows: in Section 2 we present the resources; in Section 3 we describe the mapping procedure; in Section 4 we present and discuss the results of the mapping, tested on a gold standard; in Section 5 we provide direction for future work; in Section 6 we report our conclusions.

2 The Resources

In this section we describe IMAGACT and T-PAS. Table 1 shows the total and shared quantitative data of the two resources.

	IMAGACT	T-PAS
Total Verbs	777	1,000
Total Types - *t-pas*s	1,429	4,241
Shared Verbs	248	
Shared Types - *t-pas*s	421	1,153

Table 1: Data of IMAGACT and T-PAS.

2.1 IMAGACT

IMAGACT[1] (Moneglia et al., 2014; Panunzi et al., 2014) is a visual ontology of action that provides a translation and disambiguation framework for action verbs. The resource contains a fine-grained categorization of action concepts, which are represented by one or more visual prototypes, in the form of recorded videos or 3D animations.

Action concepts are derived by a deep analysis of the most frequent action verbs in Italian and English spoken corpora; this ensures the ontology to cover the most relevant actions for our everyday activities. Given that no one-to-one correspondence can be established between an action verb and an action concept (Moneglia, 1993), each verb is divided in Types, which operate a segmentation of the predicate extension by identifying the prominent cores of the verb meaning. Verb Types are connected to action concepts and they are the linkage point between lexical and action levels (Moneglia et al., 2012a). Types in IMAGACT are inter-connected through semantic relations and gather the sentences retrieved in the spoken corpora, which have been classified and linguistically annotated with thematic roles and aktionsart[2].

The resource is growing continuously: by now, it consists of a total of 1010 action concepts, each one with a visual representation (i.e. a scene), and 21 covered languages (9 fully-mapped, 13 underway), with an average of 730 action verbs per language.

2.2 T-PAS

T-PAS[3], Typed Predicate Argument Structures (Jezek et al., 2014), is a repository of verb patterns acquired from corpora by manual clustering distributional information about Italian verbs. For every typed structure (henceforth *t-pas*), the specification of the expected semantic type (ST) for each argument slot is provided. T-PAS accounts for the following argument positions: *subject, object, indirect object, complement, adverbial* and *clausal*. A description of the sense, in the form of an *implicature*, is also linked to the *t-pas*.

Example 1 reports the *t-pas*#2 of the verb *abbattere*: the STs [[Human]] and [[Event]] are specified for the subject position (as alternatives) and [[Building]] for the object position.

(1) [[Human | Event]-subj] *abbattere* [[Building]-obj]
implicature:[[Human | Event]] *distrugge, butta giù* [[Building]]
example: "Il muratore abbatte la parete."
(Eng."The bricklayer knocks the wall.")

The STs aim at generalizing over the set of lexical items observed in a certain position for a particular sense of the verb. For instance, in Example 1, the ST [[Building]] generalizes over the lexical item *parete* (Eng. *wall*). STs are drawn from a list of about 230 types[4] and are also organized in a hierarchy, in which the elements are linked by a "IS-A" relation (Jezek et al., 2016). Table 2 presents a section of the hierarchy in which it is shown that [[Plane]] IS-A [[Vehicle]], [[Vehicle]] IS-A [[Machine]] and so on.[5] If no generalization is possible, the set of lexical items found in the argument position is listed.

...
▷ [[Artifact]]
▷ [[Machine]]
▷ [[Vehicle]]
▷ [[Plane]]
▷ [[Road Vehicle]]
▷ ..

Table 2: Section of the STs hierarchy.

Each *t-pas* corresponds to a distinct sense of the verb and is identified and defined by analysing instances of the verb in a corpus, following the lexicographic procedure called Corpus Pattern Analysis (Hanks, 2004; Hanks and Pustejovsky, 2005).[6] The corpus instances are then associated to the corresponding *t-pas*.

[1] http://www.imagact.it/

[2] See Moneglia et al. (2012b) for details on annotated data and ontology building process.

[3] http://tpas.fbk.eu/

[4] For details on the list creation see (Jezek et al., 2014).

[5] The same list has been used for the English resource PDEV (Hanks and Pustejovsky, 2005), `http://pdev.org.uk`. The hierarchy can be found in `http://pdev.org.uk/#onto`.

[6] According to the CPA procedure, after analysing a random sample of 250 concordances of the verb in the corpus, each *t-pas* is defined by recognizing its relevant structure and identifying the STs for each argument slots.

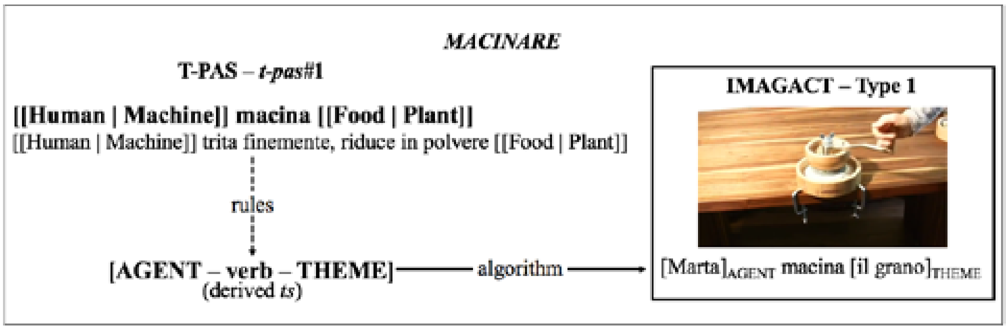

MACINARE
T-PAS – t-pas#1
[[Human | Machine]] macina [[Food | Plant]]
[[Human | Machine]] trita finemente, riduce in polvere [[Food | Plant]]
rules
[AGENT – verb – THEME]
(derived ts)
algorithm
IMAGACT – Type 1
[Marta]AGENT macina [il grano]THEME

Figure 1: An example of the mapping between IMAGACT and T-PAS for the verb *macinare*.

T-PAS currently contains 1000 verbs. The reference corpus is a reduced version of ItWAC (Baroni and Kilgarriff, 2006).

3 The Mapping

We aim at finding the best semantic match between a verb Type in IMAGACT and the *t-pas*s of the same verb in T-PAS, the two referring to the same action concept. Notice that it is possible that a Type in IMAGACT is mapped to more than one *t-pas* due, for instance, to different possible verb alternations that can occur inside the same Type. Figure 1 shows an example of this mapping, in which there is a match between Type 1 and t-pas#1 of the verb *macinare*.

The mapping is done as follows. By observing a sample of verbs in the resources, we first defined a set of simple rules to convert the *t-pas* in a thematic structure. Considering the ST in the argument positions of the *t-pas* (e.g. [Human]-subj, [Food]-obj]), the rules aim at creating a thematic structure for the *t-pas* of the kind AG-v-TH (dotted arrow in Figure 1). Then, we used an algorithm which applies these rules to all the *t-pas*s of a verb, and map the derived thematic structure (*derived-ts*) to the thematic structures (*ts*) of the Types in IMAGACT (horizontal arrow in Figure 1). The system thus compares all the *ts* in IMAGACT with all the *derived-ts* in T-PAS for the same verb, and retrieves the matches.[7] In Figure 1, the *t-pas#1* for the verb *macinare* have been transformed in the structure AG-v-TH and then mapped to the *ts* of the Type.

The mapping between IMAGACT and T-PAS is made for the 248 verbs common to the two resources.

Datasets The rules for the conversion of a *t-pas* in a *derived-ts* have been manually created by observing a sample of 15 verbs shared by the two resources (devset). We evaluated the mapping against a gold standard manually created by pairing the Types of other 14 verbs with the corresponding *t-pas*s. We extracted the 29 verbs from the 248 shared by the two resources. The selection was made preserving the variability of the verbs in the two resources, in terms of their number of Types or *t-pas*. For instance, *prendere* (*to take*) is associated with 17 *t-pas*s in T-PAS and 18 Types in IMAGACT; on the contrary *bussare* (*to knock*) has only 2 *t-pas*s and 1 Type.

Conversion rules Table 3 synthesizes the rules we adopted. The rules consider both the ST in the argument slot and the argument slot itself, and are meant to associate a ST in an argument slot to a thematic role. For example, line 7 of Table 3 has to be interpreted as follows: if for the *subject* position of the *t-pas* the ST [[Animate]] (or a IS-A [[Animate]], according to the hierarchy of ST) is expected, then the AGENT role is selected (line 8). The rules also consider if the verb is in reflexive form (line 13). Moreover, if the *t-pas* registers the ST [[Abstract Entity]] (or a ST that IS-A [[Abstract Entity]]) as unique ST for any argument position (i.e. it is the only ST expected for the position), the *t-pas* was excluded from the mapping, as IMAGACT only accounts for physical actions which do not involve abstract entities.

4 Results and discussion

In order to calculate Precision (P) and Recall (R) of the algorithm, we considered that DESTINATION (DE), SOURCE (SO) and LOCATION (LO) roles can not always be discriminated (for example, *room* is a DE in “John puts a table in the room”, a SO in “John takes the table from

[7] Notice that the mapping is considering just this information of the resources and does not consider e.g. captions in IMAGACT or examples in T-PAS.

```
y = ST in argument slot
for y:
if y = IS or IS-A [Abstract | State | ..]
  do not map
if obj:
  y in obj = Theme TH
  if y in subj IS or IS-A [[Animate]]:
    subj = Agent AG
  else:
    subj = Causer CA
else:
  if y in subj IS or IS-A [[Animate]]
  & verb is reflexive:
    subj = Actor AC
  else:
    subj = Theme TH
for y !=subj and obj:
  x = (ImagAct Role != AG, CA, AC, Instrument IN)
  x = y
```

Table 3: Rules for mapping.

the room", a LO in "John walks in the room"). The same happens for AGENT (AG) and ACTOR (AC): a human can be an agent ("John sweeps the room") or an actor ("John bumps his head"). These limits can not be exceeded by an improvement of the rule definitions, because they are strictly dependent on the verb semantics. When calculating P and R, we grouped these derived structures together.

Precision (P)	Recall (R)	F-measure (F1)
0.283	0.792	0.418

Table 4: Precision, Recall, F1 of the mapping.

We observe good values for R, while the P is very low (Table 4). A deeper analysis shows that in 34.61% of the cases, we have a full match with the gold standard and in 38.46% the results from the mapping include the ones expected by the gold standard. This means that in many cases the system is able to retrieve the correct matches.

Figure 2 shows the distribution of the main thematic structures in the Types of the whole IMAGACT ontology (in orange), in the devset (in red), compared with the *derived-ts* from T-PAS (in green). We verified *a posteriori* that the distribution of *ts*s in the devset is strictly comparable with the one in the whole ontology, meaning that the devset is also well-balanced in terms of the thematic structures coverage (see orange and red bars in Figure 2).

By using the transformational rules we were able to recreate all the structures that are used in IMAGACT; however, there are some discrepancies in the production of AG-v-TH, TH-v (too high) and AG-v-TH-[DE|LO|SO] (too low) (see Figure 2).

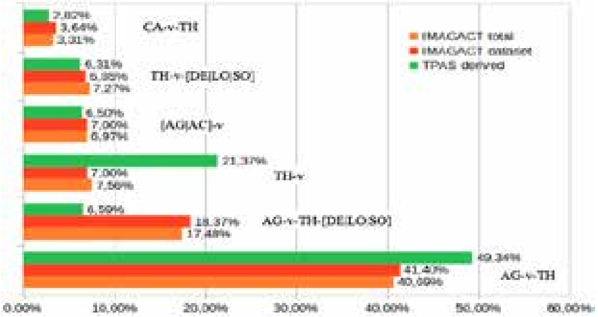


Figure 2: Distribution of the thematic structures.

The critical issue is represented by the AG-v-TH structure: this is the most frequent one among the IMAGACT Types and in our test set (112 over 166 Types). For example, the following sentences belong to 4 different Types of the verb *stringere*, but have the same *ts* AG-v-TH: "Marco stringe la mano a Luca"; "Marco stringe le gambe"; "Marco stringe i pugni"; "Marco stringe la vite". This happens also for the *t-pas* of *stringere*: 3 over the 5 *derived-ts* are AG-v-TH, so the system produces 12 combinations over 3 attested in the gold standard. The high frequency of this structure strongly influences the final P and R results. Moreover, the *ts* AG-v-TH is not distinctive of Types intra-verbs: by taking all the verbs with more than one Type, and for which AG-v-TH is a possible *ts*, we measured that in only 38,22% of them this *ts* is present in only one Type; in the other verbs (61.78%) the AG-v-TH structure appears in more than one Type.

5 Future work

Given the result in terms of Precision we presented in the previous section, we are considering to adopt other strategies that can be useful for the mapping of IMAGACT and T-PAS.

For instance, it would be possible to exploit the examples from the corpus associated with each *t-pas* in T-PAS. In this sense, we hypothesize the processing of these examples through BabelFy (Moro et al., 2014), an online system for word sense disambiguation, based on the BabelNet semantic network (Navigli and Ponzetto, 2012). BabelNet is already linked to IMAGACT (via the scenes). We can use BabelFy in order to perform the disambiguation of a verb in the sentences associated to each *t-pas*. In this way we can obtain a link between the verb under examination and the corresponding BabelNet synset (i.e., a BabelSynset). The application of this method to every example will result in a ranking of the most frequent BabelSynsets for the group of sentences of each *t-pas*. Combining this output ranking with the BabelNet-IMAGACT linking (Gregori et al., 2016), we will obtain the set of IMAGACT Types that most likely match with each *t-pas*.

On the other way round, IMAGACT captions could also be mapped into the corresponding *t-pas*s, by using the output of the algorithm developed in (Feltracco et al., 2016): given a sentence of a *t-pas*, the algorithm identifies the lexical item(s) that are generalized by the ST for each argument position of every *t-pas* (e.g. assigning the ST [[Building]] to "parete" in the sentence "Il muratore abbatte la parete" for the *t-pas* [[Human | Event]] *abbattere* [[Building]]). A measure of semantic similarity between the lexical items of an IMAGACT caption and the set of items associated to the same verb in T-PAS, would provide an approximation of which are *t-pas*s that most likely match the given caption. The application of this method to every caption of an IMAGACT Type will help us in the goal of mapping T-PAS with IMAGACT.

This method added to our rule-based strategy can be particularly useful to solve the ambiguity related to the thematic pattern AG-v-TH, for which the use of lexical information would reduce the number of possible matches.

6 Conclusions

In this paper we presented a first attempt of mapping IMAGACT and T-PAS by using a rule-based algorithm for the automatic conversion of T-PAS semantic types into thematic structures. We took advantage of the strong discriminative power of semantic types in their argument position to reduce the possible set of allowed thematic structures. This approach has an intrinsic limit: thematic roles are determined by verb semantics and their difference is not always reflected in the related semantic type. We also found out that the *ts* AG-v-TH represents the most critical issue, being the most frequent structure, and appearing in more than one Type of the same verb.

The results report a good recall and a low precision, confirming that our algorithm is not able to produce an actual mapping between the two resources, but it provides a reliable set of mapping candidates: we believe that it can be fruitfully exploited for a first step of a mapping process, in order to filter a lot of unwanted matching possibilities. We are confident that by exploiting additional linguistic information from the two resources (e.g. captions and occurrences in IMAGACT, lexical information and examples in T-PAS), the precision of this mapping will improve sensibly.

References

Mark C Baker. 1997. Thematic roles and syntactic structure. In *Elements of grammar*, pages 73–137. Springer.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90. Association for Computational Linguistics.

Melissa Bowerman. 1990. Mapping thematic roles onto syntactic functions: are children helped by innate linking rules? *Linguistics*, 28(6):1253–1290.

Anna Feltracco, Lorenzo Gatti, Simone Magnolini, Bernardo Magnini, and Elisabetta Jezek. 2016. Using WordNet to Build Lexical Sets for Italian Verbs. In *Proceedings of the Eighth Global WordNet Conference (GWC '16)*, Bucharest, Romania, January.

Lorenzo Gregori, Alessandro Panunzi, and Andrea Amelio Ravelli. 2016. Linking IMAGACT ontology to BabelNet through action videos. *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-IT 2016)*, pages 162–167.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue française de linguistique appliquée*, 10(2):63–82.

Patrick Hanks. 2004. Corpus pattern analysis. In *Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Universite de Bretagne-Sud.*

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS: a resource of corpus-derived types predicate-argument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Elisabetta Jezek, Anna Feltracco, Lorenzo Gatti, Simone Magnolini, and Bernardo Magnini. 2016. Mapping Semantic Types onto WordNet Synset. In

Proceedings of the Twelfth Joint ACL - ISO Workshop on Interoperable Semantic Annotation (Isa '12), Portorose, Slovenia, May.

Massimo Moneglia, Gloria Gagliardi, Lorenzo Gregori, Alessandro Panunzi, Samuele Paladini, and Andrew Williams. 2012a. La variazione dei verbi generali nei corpora di parlato spontaneo. L'ontologia IMAGACT. In *Proceedings of the VIIth GSCP International Conference: Speech and Corpora*, pages 406–411.

Massimo Moneglia, Gloria Gagliardi, Alessandro Panunzi, Francesca Frontini, Irene Russo, and Monica Monachini. 2012b. Imagact: deriving an action ontology from spoken corpora. In *Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-8)*, pages 42–47.

Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014. The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Massimo Moneglia. 1993. Prototypical vs. non-prototypical predicates: ways of understanding and the semantic partition of lexical meaning. In *International conference" Linguistics at the end of the century" Moscow State University February*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Alessandro Panunzi, Irene De Felice, Lorenzo Gregori, Stefano Jacoviello, Monica Monachini, Massimo Moneglia, Valeria Quochi, and Irene Russo. 2014. Translating Action Verbs using a Dictionary of Images: the IMAGACT Ontology. In *XVI EURALEX International Congress: The User in Focus*, pages 1163–1170, Bolzano / Bozen, 7/2014. EURALEX 2014, EURALEX 2014.

Steven Pinker. 2009. *Language learnability and language development, with new commentary by the author*, volume 7. Harvard University Press.

Domain-specific Named Entity Disambiguation in Historical Memoirs

Marco Rovera[1], Federico Nanni[2], Simone Paolo Ponzetto[2], Anna Goy[1]
[1]Dipartimento di Informatica, Università di Torino, Italy
{rovera,goy}@di.unito.it
[2]Data and Web Science Group, University of Mannheim, Germany
{federico,simone}@informatik.uni-mannheim.de

Abstract

English. This paper presents the results of the extraction of named entities from a collection of historical memoirs about the italian Resistance during the World War II. The methodology followed for the extraction and disambiguation task will be discussed, as well as its evaluation. For the semantic annotations of the dataset, we have developed a pipeline based on established practices for extracting and disambiguating Named Entities. This has been necessary, considering the poor performances of out-of-the-box Named Entity Recognition and Disambiguation (NERD) tools tested in the initial phase of this work.

Italiano. *Questo articolo presenta l'attività di estrazione di entità nominate realizzata su una collezione di memorie relative al periodo della Resistenza italiana nella Seconda Guerra Mondiale. Verrà discussa la metodologia sviluppata per il processo di estrazione e disambiguazione delle entità nominate, nonché la sua valutazione. L'implementazione di una metodologia di estrazione e disambiguazione basata su lookup si è resa necessaria in considerazione delle scarse prestazioni dei sistemi di Named Entity Recognition and Disambiguation (NERD), come si evince dalla discussione nella prima parte di questo lavoro.*

1 Introduction and Motivation

Current NLP techniques allow us to treat some types of historical textual resources provided by, among others, historical archives and libraries, as a source of information (and, in prospect, of knowledge) for automatic systems. Besides encyclopedic resources, libraries and archives provide many different types of texts, often spanning very specific geographical, individual or thematic contexts, for which current knowledge extraction systems may lack the suitable information. Nevertheless, the tasks of extracting, disambiguating and linking information provided by historical textual documents with respect to external knowledge bases is still a crucial step towards automatic access to written resources and for further employ of such knowledge in end-user applications (e.g. navigation, rich semantic search, creation of narrative chains). In order to address longer term tasks, such as event extraction from historical texts (Goy et al., 2015), we first addressed the task of extracting and disambiguating Named Entities (Persons, Locations and Organizations) from a corpus of historical memories of the "Liberation War" in Italy, during the Second World War. Due to the specificity of the domain and of the involved entities, state-of-the-art tools for Named Entity Recognition and Disambiguation show low performances, thus suggesting us to try to achieve our goal using a different approach. In this paper we present a collection of documents created by digitizing historical memoirs, together with an overview of the methodology we followed for the extraction and disambiguation of Persons, Locations and Organizations, as well as the results of the evaluation of its output in comparison with the output of two state-of-the-art systems. The outline of the paper is the following: in Section 2 some related projects are discussed, while in Section 3 the dataset used in the experiment is presented. Section 4 describes the test of two automatic NER tools (4.1) and the methodology devised for our experiment (4.2). In Section 5 the results of the evaluation are discussed, while Section 6 concludes the paper and outlines the next developments of the project.

2 Related Work

The work described in this paper is mainly related to Named Entity Recognition and Disambiguation (NERD) techniques and their application in the field of Digital Humanities (DH), in particular on historical texts. While NER refers to the task of identifying named entities in text and classifying them according to a set of categories, a Named Entity Disambiguation (NED) task is aimed at assigning a correspondence between an ambiguous surface form and the individual entity it refers to. Although analytically they can be considered as two separate tasks, the current availability of large, publicly accessible knowledge bases allowed to merge them into the task of Entity Linking (EL), which aims at linking a surface form from a text to the corresponding entry in a resource like DBpedia or Wikipedia (Barrière, 2016). A recent application of EL techniques in a DH context is presented in Brando et al. (2016), where the authors use a graph-based approach and exploit Linked Data for linking mentions of writers in a corpus of French literary criticism and scientific essays. Discussions and experiments on the use of third-party NER services on historical OCRed texts (typewritten memoirs of Holocaust survivors and old newspapers respectively) are provided by Rodriquez et al. (2012) and by Ehrmann et al. (2016), offering a starting point for our work, since they quantify, showing their limitations, the performances of NER such tools on specific historical texts (as also remarked in Nanni et al. (2017)). Also in the Italian DH research community, the interest for mining historical texts became more evident in the last years and leading to several interesting works. In Boschetti et al. (2014), for example, the authors describe the ongoing work of applying a full Information Extraction pipeline (from OCR digitization to data visualization) to war bulletins in WWI and WWII and discuss the issues they addressed in adapting existing tools to dated and domain-specific language. Another related project with a similar setting is ALCIDE, described in Moretti et al. (2016), a platform that supports the use of text mining techniques for the navigation and visualization of information in historical and literary texts.

3 Dataset

The collection of documents used in this work is composed by 15 printed books, written in Italian, that have been digitized using standard OCR techniques, overall counting over 855,000 words (about 45,000 sentences). The documents are historical memoirs of Italian partisans from the WWII. More specifically, the covered time span goes from the 8th September 1943 to the 25th April 1945, a period known in the Italian historiography as “Resistenza” (Resistance). The geographic area encompassed by the narrated events is the south-western part of the Alps in Piemonte, Italy, with some minor exceptions. The texts have been intentionally selected for digitization for having a partial but significant overlap in terms of narrated events, as well as of places and involved people. None of the 15 documents presents any semantic annotation. Beside the digitization of the documents, three gazetteers have been created: the first one, containing names of persons (1820 entries), has been populated using name indexes provided by 6 of the texts, while the gazetteers containing toponyms and names of organizations (1140 and 190 entries, respectively) have been built manually during the digitization activities. The setting of our work is partly determined by some features of the textual resources under analysis, in particular: 1) due to the specificity of the domain, only 4% of the persons in the gazetteer are available in the italian Wikipedia (according to a manual check carried out on the whole gazetteer); the same problem holds for organizations and, to a smaller extent, for toponyms; 2) while for entities of type Location (LOC) and Organization (ORG) the mining process involves usual problems (abbreviations, upper vs lowercase mention, ambiguity due to the same surface form), with Person (PER) entities the domain at hand presents a further issue as it was quite common, among the partisans, to use aliases, or *nom de guerre*. This feature is showed by 32% of the occurrencies in our PER gazetteer (often the most prominent ones in the narrated events). This means that in text persons are to be found under different combinations of name, surname and nickname. While in some cases this additional information makes the disambiguation process easier, in many other cases it may represent an additional source of ambiguity. The PER gazetteer is structured in three fields, namely Name, Surname and Alias, that are later combined into patterns (see section 4.2); conversely, in the ORG and LOC gazetteers, for each entry all the possible lexical forms are listed (for

	Recognition (%)		
	PER	LOC	ORG
NERD	0.66	0.70	0.51

	Linking (%)		
	PER	LOC	ORG
TagMe	**0.05**	0.45	0.37
NERD	**0.03**	0.47	0.27

Table 1: Evaluation using TagMe and NERD (Percentage of correctly linked occurrencies over a sample of 200 sentences).

the Italian Action Party, for example, we will have: Partito d'Azione, PdA, Pd'A, P.d.A. and so on).

4 Experiment

4.1 Test of existing automatic NERD tools

In order to clarify the need for an ad hoc extraction and disambiguation approach for our texts, we first tried state-of-the-art NERD tools; we randomly selected 200 sentences from the corpus and annotated them with NERD (Rizzo and Troncy, 2012), a framework that aggregates the results from different NER systems (Alchemy API, DBpedia Spotlight, TextRazor, Zemanta among others), and TagMe (Ferragina and Scaiella, 2010), an entity linker to Wikipedia available also for Italian. Table 1 shows the percentage of correctly recognized (i.e. classified) and linked occurrences obtained as result by the two systems. Since TagMe does not separate the two tasks of Recognition and Linking, for this system we only report the Linking results. In the recognition task, NERD performances are quite good for Persons and Locations, while they drop with Organizations. As we turn to the linking task, we observe how the trend in the results is similar in the two systems: performances are very low in the case of Persons, while they improve in the case of Locations and remain quite low for Organizations. This result can partly be explained by the degree of (spatial and social) specificity of the entities that are to be found in the corpus: state-of-the-art tools perform good on prominent entities (for example "Benito Mussolini"), but large-scale knowledge bases lack the suitable knowledge for specific contexts, like those that are more often to be found in the historical memoirs under analysis (and thus NERD systems are not able to link specific entities, such as Chiaffredo Barreri «Tormenta»).

4.2 Methodology

The mining process initially took the form of a simple string matching in text, based on the entries provided by the gazetteers. However, due to the different ways each entity type can appear in text - as discussed in Section 3 - two different strategies have been implemented: string matching with some refinements for LOC and ORG entity types and a slightly more elaborated strategy for PER entities, based on co-occurrence statistics derived directly from the corpus under study.

PER entities. Based on the manual analysis of the documents, 15 lexical patterns have been observed, through which proper names of partisans appear in text; frequent occurring patterns are for example "Name Surname (Alias)", like in "Gustavo Comollo (Pietro)", Name «Alias» Surname, like in "Gustavo «Pietro» Comollo", or "Alias Surname", like in "Pietro Comollo". Each of these 15 patterns have been automatically instantiated for each entry of the gazetteer. This resulted in a dictionary of instantiated patterns that have been used directly for the string matching step in text. Since a certain degree of ambiguity (homonymy) is present in the gazetteer, where many entries share the same name or surname or alias, for each instance of the patterns in the dictionary an ambiguity value has been computed, keeping track, for the ambiguous instances, of all the possible individuals they may actually refer to. For example, the pattern instance "«Renzo»", that in italian can be both a name and an alias, has been connected to all the entries in the gazetteer where "Renzo" appears either as name or as alias, which become candidates for that specific occurrence. Then the string matching in text has been performed. Within the found occurrences, we separated the unambiguous occurrences (those who refer to only one entry in the gazetteer), that have been considered as true positives and did not require further processing, from the ambiguous ones, for which a disambiguation step is needed. Only considering the unambiguous mentions retrieved this way, the system scored a precision measure of .98 (see Section 5), so we used this set of occurrencies as grounding space for the disambiguation step. At this point the system has disambiguated 55.8% (9268) of the PER occurrences in the corpus, while 44.2% (7341) of the occurrences remain ambiguous (for precision and recall scores, see Table 2, "Lookup Search"). In order to disambiguate the remaining occurrences different heuristics have been explored. Based on the literature, we tried to apply to the Named Entity Disambiguation task the "one sense per discourse" hypothesis, as done by the authors in (Barrena et al., 2014). Other two heuristics have been explored, that we can informally designate as *Last Mentioned* and *Most Mentioned.* Given an ambiguous occurrence recognized in text, the former one links the occurrence to the last already disambiguated corresponding candidate. Following from the example above, if we find the pattern "«Renzo»" in text, which is ambiguous and corresponds to more candidates from the gazetteer, the system links the mention to the same candidate as the immediately preceding occurrence of this mention. The *Most Mentioned* rule, conversely, assigns to an ambiguous occurrence the candidate which obtained the highest number of mentions in the document. None of these strategies succeeded in improving the performance of the system and this seems to be at least partly due to the length of the documents and to the high ambiguity degree of some entries (consider that the entry "Renzo" alone has 20 candidates in the dictionary, and there are other more ambiguous entries). A promising strategy for the NED task has been individuated using co-occurrence frequencies (Shen et al., 2015; Hachey et al., 2013). Still based on the unambiguous occurrences, for each entry in the PER gazetteer a co-occurrence score has been computed with all the other entities, including Locations and Organizations, at corpus level. The co-occurrence has been considered with other entities in the span of 10 sentences, in terms of raw frequency. Then, given an ambiguous mention and its local context of 10 sentences, the co-occurrence score has been computed for each of its candidates, and the candidate with the highest score has been assigned to the mention. This strategy allows to further disambiguate 10.6% (1764) of the occurrences, with precision and recall scores as indicated in Table 2 ("Lookup Search and Disambiguation").

LOC and ORG entities. For entities of type Location and Organization only the search step has been implemented, not the disambiguation one. However, a cross cleaning has been performed, eliminating nested mentions belonging to different NE categories (for example the name "Leonardo Cocito" in the ORG entity "Battaglione Leonardo Cocito"). In such cases always the longer string has been chosen.

	Lookup Search		
	Recall	Precision	F1
PER	0.716	0.980	**0.827**
LOC	0.954	0.917	0.935
ORG	0.987	0.991	0.989

	Lookup Search and Disambiguation		
	Recall	Precision	F1
PER	0.751	0.965	**0.845**

Table 2: Evaluation of the presented pipeline.

5 Evaluation

The performances of the system have been evaluated against a manually annotated gold standard made of 1,000 sentences. The gold standard has been built: a) preserving the relative size of each document with respect to the whole corpus size and b) randomly selecting the sentences in a short list that only contains sentences longer than 60 characters and with at least 3 capital letters (which is expected to maximize the probability to have a NE in the sentence). In the resulting gold standard, 1996 entities (belonging to the three mentioned categories) have been annotated as true positives by a single human annotator. The results of the evaluation are presented in Table 2. The co-occurrence approach discussed above allows to gain coverage without losing too much in terms of precision and even if the overall gain is small, the approach shows improvements where other approaches resulted ineffective. The main source of improvement is that, being computed at corpus level, the co-occurrence approach embodies the occurrence information from all the texts, thus going beyond the document level; this proves to be effective when an entity does not appear in unambiguous form in the document at hand but does in other documents of the collection. One limit of the approach emerges when an entity never appears in unambiguous form in the whole corpus, since the grounding space is uniquely based on the set of unambiguous mentions harvested in the search step. Unfortunately this is often the case when memoirs are concerned: many of the authors are non professional writers and do not always provide the full name of the persons they introduce.

6 Conclusions and Future Works

In this paper we presented an ongoing work aimed at performing Named Entity Disambiguation on a digitized historical corpus, along with the results of the evaluation. Further steps will be a) the refinement of the presented method by means of weighting measures on co-occurrence and possibly of feature optimization techniques, b) the application of the tested disambiguation strategy also to LOC and ORG entities, as well as the study of a cross-category disambiguation strategy, and finally c) the extension of the corpus and of the gazetteers in order to obtain a larger coverage of the domain. Furthermore, this work represents the first step for extracting events and their participants from the presented corpus.

References

Ander Barrena, Eneko Agirre, Bernardo Cabaleiro, Anselmo Penas, and Aitor Soroa. 2014. One entity per discourse and one entity per collocation improve named-entity disambiguation. In *COLING*, pages 2260–2269.

Caroline Barrière. 2016. *Natural Language Understanding in a Semantic Web Context*. Springer.

Federico Boschetti, Andrea Cimino, Felice Dell'Orletta, Gianluca E Lebani, Lucia Passaro, Paolo Picchi, Giulia Venturi, Simonetta Montemagni, and Alessandro Lenci. 2014. Computational analysis of historical documents: An application to italian war bulletins in world war I and II. In *Proceedings of LREC 2014 workshop on Language resources and technologies for processing and linking historical documents and archives - deploying linked open data in cultural heritage (LRT4HDA 2014)*.

Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. Reden: named entity linking in digital literary editions using linked data sets. *Complex Systems Informatics and Modeling Quarterly*, (7):60–80.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic evaluation of ner systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016))*, number EPFL-CONF-221391, pages 97–107. Bochumer Linguistische Arbeitsberichte.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

Anna Goy, Diego Magro, and Marco Rovera. 2015. Ontologies and historical archives: a way to tell new stories. *Applied Ontology*, 10(3-4):331–338.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. 2013. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150.

Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. Alcide: Extracting and visualising content from large document collections to support humanities studies. *Knowledge-Based Systems*, 111:100–112.

Federico Nanni, Yang Zhao, Simone Paolo Ponzetto, and Laura Dietz. 2017. Enhancing domain-specific entity linking in DH. *Book of Abstracts of Digital Humanities*, 2:67–88.

Giuseppe Rizzo and Raphaël Troncy. 2012. NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76. Association for Computational Linguistics.

Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw ocr text. In *KONVENS*, pages 410–414.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Toward a Treebank Collecting German Aesthetic Writings of the Late 18th Century

Alessio Salomoni
University of Bergamo-Pavia
Corso Strada Nuova 65
Pavia, Italy, 27100
alessio.salomoni@unibg.it

Abstract

English. In this paper, I will describe the methodology to develop the first sample of a dependency treebank collecting German aesthetic writings of the late 18th century. A gold standard of the target data was annotated in order to evaluate some data-driven tools, trained on contemporary web news. Results are reported and discussed.

Italiano. *In questo articolo descriverò la metodologia adottata nello sviluppo di un sample preliminare di una treebank per il tedesco, che raccoglierà scritti di estetica della fine del XVIII secolo. È stato annotato un campione della varietà target, ed è stata valutata l'accuratezza di alcuni strumenti data-driven addestrati su una varietà giornalistica contemporanea. I risultati sono stati riportati e commentati.*

1 Introduction

A constantly increasing amount of digital texts of the German literary history is freely available online as downloadable raw texts, especially thanks to important ongoing projects, such as deutschestextarchiv.de or zeno.org, to name but a few. In spite of this, we still lack annotated corpora gathering them per author and genre. Indeed, this is a strong bottleneck in exploiting such textual treasure for linguistic analysis through computational methods. At the same time, available training data for data-driven annotation tools mainly come from the domain of contemporary web news. Therefore, models have to be trained on this particular variety of the German language, which could be very different, in terms of linguistic features, from the target unannotated data. Such variation between the training set and the test set could cause tools' performances to drop (Gildea, 2001). Therefore, testing such models on a portion of the target texts is crucial. On the one hand, to show their robustness. On the other hand, more practically, to understand to what extent available tools can actually boost the semi-automatic annotation of new data.

In this paper, I will highlight the methodology behind the development of a first sample of a dependency treebank aiming to collect German aesthetic essays of the late 18th century. By aesthetic essays I mean theoretical writings about art, poetics, beauty and related issues, which were mainly published on literary magazines, chiefly targeting non-academic middle-class readers. [1] In that period, there was a remarkable production of these texts in Germany, and they contributed to popularize the recently born modern 'Hochdeutsch', i.e. the modern variety of the German language. To the best of my knowledge, despite its importance, such textual genre has never been studied in depth at any linguistic level. In a long-term perspective, a dependency treebank will surely provide empirical data to fill the gap, especially concerning syntax and semantics. Indeed, many studies can be done on such resource, ranging from using dependency networks to describe syntactic phenomena (Passarotti, 2014), to extracting a valency lexicon (Passarotti et al., 2016).

In the rest of this paper, some fundamental issues concerning the treebank design are highlighted and preliminary results concerning automatic lemmatization, POS-tagging and dependency parsing are reported and discussed.

[1] Philosophical monographs about aesthetics from the same period are not part of the target data for this resource, belonging to a different genre.

2 Methodology

2.1 Data

Even if we are dealing with texts in prose in a defined domain, style between authors may vary substantially, especially in terms of syntax and lexicon. Therefore, to avoid too much variation in my data, for this first sample I focused on a particular text typology inside the target genre: fragments, i.e. really short texts, sometimes in aphorism-like form. I assumed that such texts could be dealt with as a whole, in spite of their different authorship. [2] For the first sample of the treebank, I selected the following data: F. Schlegel, *Lyceum Fragmente*, fragments from 1 to 90; F. Schlegel and other authors, *Athenaeum Fragmente*, fragments from 1 to 50; Novalis, *Blüthenstaub*, fragments from 1 to 31. All the raw texts in .txt format were obtained from zeno.org. Overall, this initial corpus counts 7337 tokens.

2.2 Annotating a Gold Standard

Such corpus was semi-automatically annotated to build a gold standard. As for the annotation scheme, I adhered to the Universal Dependencies (UD) 2.0 scheme (Nivre et al., 2017). Texts were tokenized and brought into conllu format with UDPipe1.1 (Straka et al., 2016). Then they were brought into conll09 format (Hajič et al., 2009) and processed with Anna 3.6 pipeline (Bohnet, 2010).[3] I had used this suite in previous preliminary experiments on some data from the same period and domain, attaining good initial results for POS-tagging and dependency parsing. I assigned the following metadata: LEMMA, UPOS (the coarse-grained POS-tag, based on the Google tagset (Petrov et al., 2011)), XPOS (the fine-grained POS-tag, based on the STTS tagset (Brants et al., 2002)), HEAD (the regent element of the dependency relation) and DEPREL (the kind of dependency relation). As for LEMMA and XPOS, pre-trained models based on the Tiger Corpus (Brants et al., 2002) were used. As for UPOS, HEAD and DEPREL, I trained a model on the training file of the German treebank in UD 2.0[4]. Then, at each stage of the processing, the automatic output was manually checked. An annotated fragment is shown in Figure 1 in a tree-like form.

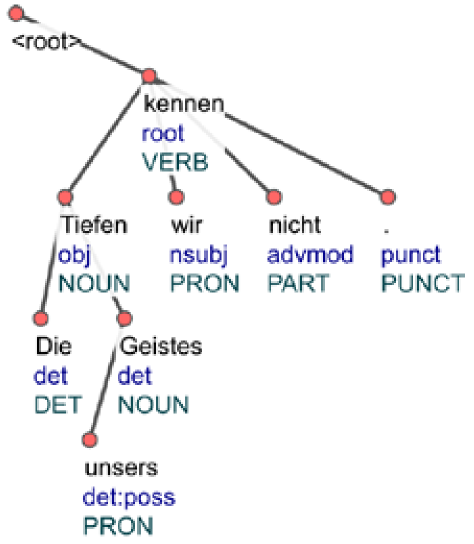


Figure 1: Dependency representation of the simple German sentence 'Die Tiefen unsers Geistes kennen wir nicht.' (We don't know the depths of our soul) by Novalis, according to UD 2.0 scheme.

I briefly describe the formalism in Figure 1. The main node of each sentence usually is the main verb, which is 'kennen' in this case, whose relation is tagged as 'root'. The article 'Die' depends on the common noun 'Tiefen' as determiner, while 'Tiefen' depends on 'kennen' as nominal object. 'Wir' is a personal pronoun playing the role of nominal subject. 'unsers' is a possessive pronoun modifying the common noun 'Geistes', which is in genitive case and modifies the subject. According to the current scheme, such modifier depends on the noun it refers to through 'det' relation.

2.3 Lemmatization

Training	Lemmatizer	Frag
Tiger (pre-trained)	Anna 3.6	97.6

Table 1: Accuracy by Anna 3.6 lemmatizer on the target data. 'Frag' stands for accuracy on fragments.

As for lemmatization, I measured the accuracy by Anna 3.6 lemmatizer (Björkelund et al., 2010) on fragments only. Results are shown in Table 1.[5] Given the high overall accuracy by Anna 3.6, I did

[2] Preliminary clustering and syntactic parsing experiments confirmed this hypothesis.

[3] Double multi-word tokens such as 'der+im' for the determined article 'dem' or 'in+dem' for the preposition 'im' had to be removed to work with this format.

[4] It counts about 287.000 tokens.

[5] All the results in this paper are expressed as percentage.

not test any other system. I briefly report some issues concerning this task: inflected adjectives such as 'andre' or 'unsrer' where 'e' in stem drops after inflection (for instance, the stem of 'unsrer' is 'unser') are lemmatized without 'e'; deadjectival nouns such as 'Langweile' or 'Kürzeste' are lemmatized as nouns with the same form, not as adjectives; the non-finite verb 'seyn' is lemmatized as 'seyn', not with the current spelling 'seien'.

2.4 POS-Tagging

As for POS-tagging, I tested some candidate POS-taggers on fragments first. Once the best-performing one was detected, I tested it also on the source variety to measure the accuracy gap. Before doing that, I had to cope with some issues concerning models and training data. According to the documentation provided with the treebank file, in the UD German treebank UPOS was assigned manually, while XPOS was assigned automatically by using Tree Tagger, trained on Tiger Corpus, with no manual checking. Thus, the UD treebank was not ideal to train a model for XPOS. At the same time, I was interested in testing both tagsets on the target data. Consequently, I followed two different methods. First, I considered the UPOS tagset. I picked up two candidate POS-taggers, I trained them on the whole training file of the German treebank in UD and I tested them on fragments. They were fed with the automatically lemmatized texts by Anna 3.6. Overall accuracy is shown in Table 2. [6]

Training	POS-tagger	Frag
100% de-ud-train	Anna 3.6	93
	UDPipe 1.1	88.5

Table 2: Accuracy by Anna 3.6 and UDPipe 1.1 POS-tagger (Straka et al., 2016) assigning UPOS to fragments.

The best POS-tagger was Anna 3.6, thus I run it on UD, performing a ten-fold validation. I split up the training file of the UD 2.0 German treebank into two partitions with ratio 9:1. I trained the POS-tagger on the 90% and tested it on the remaining 10%. I repeated the experiment ten times, varying each time the two partitions. Overall accuracy concerning these experiments, i.e. the average of the ten measures, is shown in Table 3.

Training	POS-tagger	UD
90% de-ud-train	Anna 3.6	93.6

Table 3: Overall average accuracy by Anna 3.6 in assigning UPOS to the UD test set.

As for Anna 3.6 POS-tagger, I report the accuracy on some specific part-of-speeches on both test sets. The first number in brackets refers to UD [7], while the second one to the fragments: VERB (95.1/94.8), PROPN (proper nouns) (84.01/83.6); NOUN (93.71/94.23); SCONJ (subordinate conjunctions) (89.1/79); ADJ (adjectives) (91.2/94); AUX (auxiliaries) (83.9/77.7) and ADV (adverbs) (90.7/83.02). There is a remarkable gap between the two varieties on adverbs, subordinating conjunctions and auxiliaries. On fragments, a lot of adverbs have been mismatched with adjectives, for instance when they modify adjectives, while many occurrences of the subordinate conjunction 'daß' have been wrongly assigned. As for AUX, the modal verb 'müssen' was frequently assigned a wrong POS. As for VERB, the verb 'sein' was frequently tagged as AUX when it occurs as verbal part of a nominal predicate, while, in this case, it should be tagged as VERB, according to the UD scheme.

Training	POS-tagger	Frag
Tiger (p)	Anna 3.6	97.3
Tiger (p)	RFTagger	88
Negra (p)	Stanford	92.9

Table 4: Accuracy by Anna 3.6, RFTagger (Schmid and Laws, 2008) and Stanford Tagger (Manning et al., 2014) assigning XPOS to fragments. 'p.' stands for pre-trained model.

Second, I considered XPOS. At first, I tested three POS-taggers which are commonly used with the STTS tagset on fragments. I used pre-trained models provided by developers. Overall results are shown in Table 4. Anna 3.6 outperformed other candidates, and its overall accuracy is clearly

[6] In all these POS-tagging experiments, accuracy is the number of correctly assigned POS-tags divided by the total number of POS-tags in the test set.

[7] The reported value is the average of the ten accuracy values attained on each POS in each experiment of the ten-fold validation.

higher than that on UPOS on the same test set. Such a significant improvement could be due to the considerably different size of the training sets.[8]

Following the method adopted in the UPOS session, I performed a ten-fold validation of Anna 3.6 POS-tagger on Tiger Corpus 2.2. Overall average accuracy was 97.7. Results concerning single selected POS on both test sets is shown in Table 5. To remind the difference in granularity between the two tagsets, for each group of XPOS I reported the corresponding UPOS as well. In contrast to UPOS, problems concerning auxiliaries and subordinating conjunctions on fragments seem to be overcome, while there are still issues concerning non-finite modal verbs, such as 'müssen'.

UPOS	XPOS	Tiger	Frag
VERB	VVFIN	93.3	94.5
	VVINF	93.4	96.1
	VVPP	95.8	96
	VVIZU	93	100
AUX	VMFIN	98.6	100
	VMINF	75	88
	VAFIN	98.4	100
	VAINF	94	95.4
ADJ	ADJA	98.3	97.7
	ADJD	94	95.5
ADV	ADV	97.2	88.4
NOUN	NN	98.7	99.2
PROPN	NE	92.1	95.5
SCONJ	KOUS	97.7	100

Table 5: Overall accuracy by Anna 3.6 in assigning XPOS to UD and fragments. As for verbs, VVFIN stands for finite verbs, VVINF for non-finite verbs, VVPP for past participle, VVIZU for non-finite verbs in non-finite clauses. As for auxiliaries, it is the same, with A standing for auxiliary and M standing for modal. For further details, I redirect to STTS online documentation.

2.5 Dependency Parsing

As for dependency parsing, I tested four different candidate parsers. First, I performed a ten-fold validation on the training set of the UD German treebank, using the same partitions from the POS-tagging session. Second, the parsers were trained on the whole training set of the German treebank and tested on fragments. In this case, morphological features were removed from the training set, because they have not been annotated in my test set yet, therefore the parsing model should not include them. All the four parsers were fed with the automatically lemmatized and POS-tagged texts (both with UPOS and XPOS). Such metadata were assigned by Anna 3.6. The candidate parsers and their settings are introduced below, while overall results are reported in Table 6. Parsing accuracy was measured through Malt Eval (Nivre et al., 2010) and it is expressed in terms of *labeled attachment score* (LAS).

- **Malt Parser 1.9.0** (Nivre et al., 2006), a transition-based system. This parser performs better with an optimized configuration obtained through Malt Optimizer, i.e. a software able to suggest the best parsing configuration after reading the training data. First, I run Malt Optimizer on the ten partitions of the training file of the UD German Treebank. Then, for each of them, the suggested configuration was used to parse the corresponding test set. Second, Malt Optimizer (Ballesteros and Nivre, 2012) was run on the whole UD training file, and the suggested configuration [9] was used to parse the target variety.

- **Anna 3.6** (Bohnet, 2010) by Mate Tools, a graph-based system. It was run with 10 training iterations.

- **Joint Parser 1.30** (Bohnet and Nivre, 2012), a transition-based system with beam search, graph completion model and an integrated part-of-speech tagger. It was run with the R6J transition, 25 training iterations and beam search parameter fixed at 40.

- **Parsito**, a transition-based system with a neural network classifier, included in the UDPipe 1.1 suite (Straka et al., 2016). It was run in the standard configuration.

Overall, Anna 3.6 attained the highest accuracy on both test sets. However, there is a 19.2% accuracy gap between the two top scores on the two varieties.

[8] Indeed, Tiger Corpus 2.2 is about three times bigger than the UD German treebank used to train the model for UPOS.

[9] system: liblinear; feature model: addMergPOSTAGS0I0FORMLookahead0; algorithm: stackproj

Training	Parser	UD	Frag
100% de-ud-train	Malt 1.9	81.1	61.3
	Anna 3.6	84. 6	65.4
	Joint	81	64.2
	Parsito	83	60.6

Table 6: Overall accuracy by four different dependency parsers on UD and on fragments.

2.6 Parsing in-depth Evaluation

In order to detect which syntactic relations are more difficult to correctly parse in fragments, I did an in-depth evaluation for all the parsers. Accuracy concerning some of the most problematic relations is reported in Table 7. [10]

Deprel	Parser	F-Score Frag
acl	Malt	52.7
	Anna	69.2
	Joint	61.7
	Parsito	63.8
xcomp	Malt	27.6
	Anna	36.5
	Joint	30.8
	Parsito	33.6
advcl	Malt	39.7
	Anna	62.5
	Joint	51.6
	Parsito	55.6
conj	Malt	67.9
	Anna	77.2
	Joint	61.8
	Parsito	725
root	Malt	68.2
	Anna	73.8
	Joint	74.6
	Parsito	73.2

Table 7: Parsing accuracy on single dependency relations.

I supply a brief description of the dependency relations I reported in Table 7. 'acl' stands for adjectival clause modifier, i.e. it refers to all those finite and non-finite clauses modifying a noun, such as the relative clauses. For instance, it occurs between the noun 'Apfel' in the main clause and the subordinate verb 'liegt' in the sentence 'Die Apfel, die auf dem Tisch liegt' (The apple, that is on the table). It is different from 'advcl', which stands for adverbial clause, i.e. a clause modifying a predicate not as a core argument. It occurs, for instance, between the subordinate verb and the main verb in the sentence 'Ich denke, dass diese Prüfung ganz schwierig ist' (I think that this exam is really difficult). 'xcomp' stands for all those predicative or clausal complements without their own subject. In German, such function matches different syntactic phenomena. For example, it occurs between the main verb and the subordinate verb in non-finite clauses introduced by the particle 'zu', such as in 'Ich habe viel zu tun' (I have a lot to do); or between the predicative part of verbs such as 'lassen' , 'scheinen' or even 'nennen' and the verb, such as in 'Ich lasse dich gehen' (I let you go). 'conj' is the relation occurring between coordinate items, while 'root', as shown in Figure 1, is the dependency relation assigned to the main predicate of each sentence.

In German, the subordinate verb lies at the end of the clause, thus relation length, i.e the number of tokens between the head (in this case the main verb) and the dependent (the subordinate verb), may be really high. This can play a crucial role in parsing accuracy, especially for transition-based systems. Malt parser mostly attained low accuracy on this kind of relations, while performances by this system increases on 'conj' relation. This could be do to the relatively low frequency of coordinate relations occurring between verbs in this test set (23% of all 'conj' relations), which are usually more likely to generate long relations. Anna 3.6 sensibly outperformed the other systems on 'acl', 'advcl' and on 'conj' too. As for the 'root' relation, a part from Malt Parser, performances are almost similar. On 'xcomp', accuracy by all the systems dramatically drops. This could be due to the high relation length between some non-finite verbs and their heads, but also to the wide range of different syntactic constructions in which such relation occurs.

3 Conclusion and Future Work

In this work, I described the methodology behind the development of a first sample of a German treebank collecting a particular kind of aesthetic essays from the late 18th century, called

[10] I have not done an in-depth evaluation of the results on UD yet.

fragments. A gold standard was annotated adhering to UD 2.0. Then some data-driven tools were tested either on the target data and on a test set of the source variety. Some core issues concerning the automatic annotation were highlighted. As for LEMMA and XPOS, overall accuracy on the target data was high and very close to that on the source variety. As for UPOS, the accuracy by the best tagger dropped, especially on the target data. Therefore, to assign POS-tag, the very good results on the STTS tagset may suggest to automatically assign XPOS first and then derive UPOS from XPOS. Furthermore, the influence of POS-tagging granularity on parsing has not been studied yet. As for dependency parsing, the overall gap between the target variety and the source variety was remarkable (19%). An in-depth comparison between the two varieties concerning single relations will surely help to better detect parsing problems on fragments. In addition, parsing manually lemmatized and POS-tagged texts will surely shed light on the error propagation on parsing.

References

Miguel Ballesteros and Joakim Nivre. 2012. Maltoptimizer: an optimization tool for maltparser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–62. Association for Computational Linguistics.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36. Association for Computational Linguistics.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465. Association for Computational Linguistics.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics*, pages 89–97. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, volume 168.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Joakim Nivre, Laura Rimell, Ryan McDonald, and Carlos Gomez-Rodriguez. 2010. Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 833–841. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, et al. 2017. Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Marco Passarotti, Berta González Saavedra, and Christophe Onambélé. 2016. Latin vallex. a treebank-based semantic valency lexicon for latin. In *LREC*.

Marco Passarotti. 2014. The importance of being sum. network analysis of a latin dependency treebank. In *Proceedings of La Prima Conferenza Italiana di Linguistica Computazionale*, pages 291–295.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 777–784. Association for Computational Linguistics.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *LREC*.

Assessing the Use of Terminology in Phrase-Based Statistical Machine Translation for Academic Course Catalogues Translation

Randy Scansani
University of Bologna
Forlì, Italy
randy.scansani@unibo.it

Marcello Federico
Fondazione Bruno Kessler
Trento, Italy
federico@fbk.eu

Luisa Bentivogli
Fondazione Bruno Kessler
Trento, Italy
bentivo@fbk.eu

Abstract

English. In this contribution we describe an approach to evaluate the use of terminology in a phrase-based machine translation system to translate course unit descriptions from Italian into English. The genre is very prominent among those requiring translation by universities in European countries where English is not a native language. Two MT engines are trained on an in-domain bilingual corpus and a subset of the Europarl corpus, and one of them is enhanced adding a bilingual termbase to its training data. Overall systems' performance is assessed through the BLEU score, whereas the f-score is used to focus the evaluation on term translation. Furthermore, a manual analysis of the terms is carried out. Results suggest that in some cases - despite the simplistic approach implemented to inject terms into the MT system - the termbase was able to bias the word choice of the engine.

Italiano. *Nel presente lavoro viene descritto un metodo per valutare l'uso di terminologia in un sistema PBSMT per tradurre descrizioni di unità formative dall'italiano in inglese. La traduzione di questo genere di testi è fondamentale per le università di Paesi europei dove l'inglese non è una lingua ufficiale. Due sistemi di MT vengono addestrati su un corpus in-domain e un sottoinsieme del corpus Europarl. Ad uno dei due sistemi viene aggiunto un glossario bilingue. La valutazione delle prestazioni globali dei sistemi avviene tramite BLEU score, mentre f-score usato per la valutazione specifica della traduzione dei termini. È stata inoltre condotta un'analisi manuale dei termini. I risultati evidenziano che, nonostante il metodo elementare utilizzato per inserire i termini nel sistema di MT, il termbase in alcuni casi in grado di infuenzare la scelta dei termini nell'output.*

1 Introduction

Availability of *course unit descriptions* or *course catalogues* in multiple languages has started to play a key role for universities especially after the Bologna process (European Commission et al., 2015) and the resulting growth in student mobility. These texts aim at providing students with all the relevant information regarding contents, prerequisites, learning outcomes, etc.

Since course unit descriptions have to be drafted in large quantities on a yearly basis, universities would benefit from the use of machine translation (MT). Indeed, the importance of developing MT tools in this domain is further testified by two previous projects funded by the EU Commission, i.e. TraMOOC[1] and Bologna Translation Service[2]. The former differs from the present work since it does not focus on academic courses, while the latter does not seem to have undergone substantial development after 2013 and in addition to that, it does not include the Italian-English language combination.

Automatically producing multilingual versions of course unit descriptions poses a number of challenges. A first major issue for MT systems is the scarcity of high quality human-translated parallel texts of course unit descriptions. Also, descriptions feature not only terms that are typical of institutional academic communication, but also expressions that belong to specific disciplines (Ferraresi, 2017). This makes it cumbersome to

[1] Translation for Massive Open Online Course http://tramooc.eu/

[2] http://www.bologna-translation.eu

choose the right resources and the most effective method to add them to the MT engine.

For this study, we chose to concentrate on course units belonging to the disciplinary domain of exact sciences, since Italian degree programmes whose course units belong to this domain translate their contents into English more often than other programmes.

A phrase-based statistical machine translation system (PBSMT) was used to translate course unit descriptions from Italian into English. We trained one engine on a subset of the Europarl corpus and on a small in-domain corpus including course unit descriptions and degree programs (see sect. 3.1) belonging to the domain of the exact sciences. Then, we enriched the training data set with a bilingual terminology database belonging to the educational domain (see sect. 3.2) and built a new engine. To assess the overall performance of the two systems we automatically evaluated them with the BLEU score. We then focused on the evaluation of terminology translation, by computing the f-score on the list of termbase entries occurring both in the system outputs and in the reference translation (see sect. 4). Finally, to gather more information on term translation, a manual analysis was carried out (see sect. 5).

2 Previous work

A number of approaches have already been developed to use in-domain resources like corpora and terminology in statistical machine translation (SMT), indirectly tackling the domain-adaptation challenge for MT. For example, the WMT 2007 shared task was focused on domain adaptation in a scenario in which a small in-domain corpus is available and has to be integrated with large generic corpora (Koehn and Schroeder, 2007; Civera and Juan, 2007). Recently, the work by Štajner et al. (2016) showed that an English-Portuguese PBSMT system in the IT domain achieved best results when trained on a large generic corpus and in-domain terminology.

For French-English in the military domain, Langlais (2002) reported on improvements of the WER score after using existing terminological resources as constraints to reduce the search space. For the same language combination, Bouamor et al. (2012) used couples of MWEs extracted from the Europarl corpus as one of the training resources, yet only observing a gain of 0.3% BLEU points (Papineni et al., 2002).

Other experiments have focused on how to insert terms in an MT system without having to stop or re-train it. These dynamic methods suit the purpose of the present paper, as they focus (also) on Italian-English. Arcan et al. (2014b) injected bilingual terms into a SMT system dynamically, observing an improvement of up to 15% BLEU points for English-Italian in medical and IT domains. For the same domains and with the same languages (in both directions), Arcan et al. (2014a) developed an architecture to identify terminology in a source text and translate it using Wikipedia as a resource. The terms obtained were then dynamically added to the SMT system. This study resulted in an improvement of up to 13% BLEU score points.

We have seen that results for the languages we are working on are encouraging, but since they are strongly influenced by several factors – i.e. the domain and the injection method – an experiment on academic institutional texts is required in order to test the influence of bilingual terminology resources on the output.

3 Experimental Setup

3.1 Corpora

A subset of 300,000 sentence pairs was extracted from the Europarl Italian-English bilingual corpus (Koehn, 2005). Limiting the number of sentence pairs of the generic corpus was necessary due to limitations of the computational resources available. Then, bilingual corpora belonging to the academic domain were needed as development and evaluation data sets and to enhance the training data set. One course unit description corpus was available thanks to the CODE project[3]. After cleaning of texts not belonging to the exact science domain, we merged the corpus with other two smaller corpora made of course unit descriptions. We then extracted 3,500 sentence pairs to use them as development set.

Relying only on course unit descriptions to train our engines could have led to an over-fitting of the models. Moreover, high quality parallel course unit descriptions are often difficult to be found. To

[3] CODE is a project aimed at building corpora and tools to support translation of course unit descriptions into English and drafting of these texts in English as a lingua franca. http://code.sslmit.unibo.it/doku.php

Data Set	Sent. pairs	It Tokens	En Tokens
Training (Europarl)	300,000	7,848,936	8,046,827
Training (in-domain)	34,800	441,030	399,395
Development	3,500	48,671	43,919
Test	3,465	49,066	45,595

Table 1: Number of sentence pairs and tokens in each of the data sets used.

overcome these two issues we added a small number of degree program descriptions to our in-domain corpus. To conclude, a fourth small course unit descriptions corpus was built to be used as evaluation data set. All the details regarding the sentence pairs and tokens are provided in Table 1.

3.2 Terminology

The terminology database was created merging three different IATE (InterActive Terminology for Europe)[4] termbases for both languages and adding to them the terms extracted from the fifth volume of the Eurydice[5] glossaries. More specifically, the three different IATE termbases were: Education, Teaching, Organization of teaching.

To verify the relevance of our termbase with respect to the training data we measured its coverage. Since the terms in the termbase are in their base form, in order to obtain a more accurate estimate we lemmatised[6] the training sets before calculating the overlap between the two resources.

As we can see in Table 2, the 24.08% of the termbase entries are also in the source side of the two training corpora, and 29.19% are in the target side, meaning that the two resources complement each other well.

	It	En
Europarl lemmas	7,848,936	8,046,827
In-domain lemmas	441,030	399,395
Termbase entries	4,142	4,142
Europarl overlap	23.03%	29.20%
In-domain overlap	27.52%	29.33%
Total overlap	24.08%	29.19%

Table 2: Number of lemmas in the generic and in-domain training sets, termbase entries, and coverage of the termbase wrt. training data.

3.3 Machine Translation System

We tested the performance of a PBSMT system trained on the resources described in sections 3.1 and 3.2. The system used to build the engines for this experiment is the open-source ModernMT (MMT)[7] (Bertoldi et al., 2017). Two engines were built in MMT:

- One engine trained on the subset of Europarl plus our in-domain corpus.
- One engine trained on the subset of Europarl plus our in-domain corpus and the terminology database.

Both engines were tuned on our development set and evaluated on the test set (see sect. 3.1).

4 Experimental results

To provide information on the overall translation quality of our PBSMT engines, we calculated the BLEU scores (Papineni et al., 2002) obtained on the test set. Table 3 shows the results for both engines, where the engine without terminology is referred to as *w/o terms* and the one with terminology is referred to as *w/ terms*.

Furthermore, we evaluated the systems focusing on their performance on terminology translation. To this purpose, we relied on the f-score. More in detail, for both engines we extracted the number of English termbase entries appearing in the system output and in the reference translation. Exploiting these figures, we were able to compute Precision, Recall and f-score. Results are reported in Table 4.

Engine	BLEU
w/o terms	25.92
w/ terms	26.00

Table 3: BLEU score for the two engines.

[4] http://iate.europa.eu/

[5] http://eacea.ec.europa.eu/education/eurydice/

[6] Lemmatisation was performed using the TreeTagger: https://goo.gl/JjHMcZ

[7] http://www.modernmt.eu/

	w/o terms	**w/ terms**
Terms in ref	1,133	1,133
Terms in output	1,061	1,083
Correct terms	633	630
Precision	0.596	0.581
Recall	0.558	0.555
F-score	0.577	0.568

Table 4: Number of occurrences of termbase entries in the reference and in the output texts, number of terms in the reference appearing also in the outputs, Precision, Recall and F-score.

The figures in Tables 3 and 4 show that adding our termbase to the training data set does not affect the output in a substantial way. While according to the BLEU score the w/ terms engine slightly outperforms the w/o terms engine, the f-score – indicating performance on term translation – is marginally higher for the w/o terms system.

Focusing on the usage of terminology, a number of observations can be made. As regards the distribution of termbase entries in the test set - which contains 3,465 sentence pairs - it is interesting to know that the number of output and reference sentences containing at least one term is fairly low, i.e. 945 (27.30%) for the reference text, 866 (24.99%) for the w/o terms output and 870 (25.10%) for the w/ terms output.

Considering the terms found in the two outputs, we observe that their number only differs by 23 units (ca. 2% of the number of terms in the outputs). Also, the number of overlapping terms is very high, i.e. 882 terms (out of 1,061 for the engine w/o terms and out of 1,083 for the engine w/ terms). As a matter of fact, the top-6 frequent terms in the systems' outputs are the same – *course*, *oral*, *ability*, *lecture*, *technology* and *teacher* – and cover approximately a half of the total amount of extracted terms for both outputs.

We then compared the English termbase entries appearing in the target side of the test set to those appearing in the training set. Each of the 78 terms occurring at least one time in the test set (corresponding to 1,133 total occurrences as reported in Table 4), also occur in the training set – out of which 60 in its in-domain component.

However, even though our training data cover the total amount of terms present in the test data, and despite the high overlap between the terms produced by the two engines, still there is a considerable number of terms that are different. We thus cannot exclude an influence of the termbase on the word choice of the w/ terms system. For this reason, an in-depth analysis of the different terms produced by the two engines was carried out.

5 Manual Evaluation

The analysis of the sentences where the termbase entries used by the two engines differed showed that in some cases the termbase forced the system to use its target term even if a different translation - sometimes also correct - was present in the training corpora. Some examples are reported in Table 5. For the source words *prova orale* (Example 1) and *esame scritto* (Example 2), the engine w/ terms used *oral examination* and *written examination*, while the one w/o terms used *written exam* and *oral exam*, but only the occurrences with *examination* are in the termbase. Moreover, Example 2 also includes the termbase word *preparazione*, which is translated with *preparation* by the engine w/ terms, while it is not translated at all by the engine w/o terms.

Another interesting example is the translation of the source word *docente* (Example 3), where the termbase corrected a wrong translation. The Italian term was wrongly translated with *lecture* by the engine w/o terminology, and with *teacher* - which is the right translation for this text - by the engine w/ terminology.

In Example 4, the Italian sentence contained the termbase entry *voto finale*, which was translated with *final vote* by the engine w/o terms and with the termbase MWE *final mark* by the w/ terms engine. Also in this case the termbase corrected a mistake, since *vote* is not the correct translation of *voto* in this context.

The comparison between the two engines' outputs shows that, even though our training data covered the total amount of terms present in the test set, the termbase influenced the MT output of the engine w/ terms biasing the weights assigned to a specific translation.

Such results have to be judged taking into account the preliminary nature of this study, aimed at understanding the practical implications of using terminology in PBSMT, and therefore exploiting a simplistic approach to inject terms. As a matter of fact, we found that also some of the termbase entries occurring in the reference – e.g. *certifica-*

SRC	La **prova orale** si svolgerà sugli argomenti del programma del corso.	
REF	The **oral verification** will be on the topics of the lectures.	
W/O TERMS	The **oral exam** will take place on the program of the course.	✓
W/ TERMS	The **oral examination** will take place on the program of the course.	✓
SRC	La **preparazione** dello studente sarà valutata in un **esame scritto**.	
REF	Student **preparation** shall be evaluated by a 3 hrs **written examination**.	
W/O TERMS	The student will be evaluated in a **written exam**.	✗
W/ TERMS	The **preparation** of the student will be evaluated in a **written examination**.	✓
SRC	Ogni **docente** titolare	
REF	Each **lecturer**.	
W/O TERMS	Every **lecture**.	✗
W/ TERMS	Every **teacher**.	✓
SRC	In tal caso il **voto finale** terrà conto anche della prova orale.	
REF	In this case the **final score** will be based also on the oral part.	
W/O TERMS	In this case the **final vote** will take account the oral test.	✗
W/ TERMS	In this case the **final mark** will be based also the oral test.	✓

Table 5: MT output examples showing the influence of the termbase on the word choice of the w/terms engine. Note that the ✓ and ✗ marks refer to human assessment and not to the correspondence with the reference.

tion, *instructor*, *text book*, *educational material* – were not used in the output of the system w/ terms and this is probably due to the limitations of our method. The terms *instructor*, *text book* and *educational material* did not occur in the w/o terms output neither, while *certification* did.

To sum up, what emerges is that using terminology in PBSMT to translate course catalogues - and more specifically course unit descriptions - can influence the MT output. In our case, since the improvements were measured against the output of the w/o terms engine - which might eventually be correct even if using different terms from those included in the termbase - the metrics results were not informative enough and a manual analysis of the terms had to be carried out.

6 Conclusion and further work

This paper has described a preliminary analysis aimed at assessing the use of in-domain terminology in PBSMT in the institutional academic domain, and more precisely for the translation of course unit descriptions from Italian into English. Following the results of the present experiment and given its preliminary nature, we are planning to carry out further work in this field.

In section 4 we have seen that the institutional academic terms contained in our testing data also appeared in the training data, thus limiting the impact of terminology on the output. However, course catalogues and course unit descriptions include terms belonging to the specific disciplines (see sect. 1) as well. In our future works we are therefore planning to focus not only on academic terminology, but also on the disciplinary one testing its impact on the output of an MT engine translating course unit descriptions.

After this first experiment on the widely-used PBSMT architecture, in future work we are planning to exploit neural machine translation (NMT). In particular, our goal is to develop an NMT engine able to handle terminology correctly in this text domain, in order to investigate its effect on the post-editor's work. For this reason, a termbase focused on the institutional academic domain, e.g. the UCL-K.U.Leuven University Terminology Database[8] or the Innsbrucker Termbank 2.0[9] could be used to select an adequate benchmark for the development and evaluation of an MT engine with a high degree of accuracy in the translation of terms.

Ackowledgements

The authors would like to thank Silvia Bernardini, Marcello Soffritti and Adriano Ferraresi from the University of Bologna for their advice on terminology and institutional academic communication, and Mauro Cettolo from FBK for help with ModernMT. The usual disclaimers apply.

[8] https://goo.gl/huoevR

[9] https://goo.gl/W2GH5h

References

Mihael Arcan, Claudio Giuliano, Marco Turchi, and Paul Buitelaar. 2014b. Identification of bilingual terms from monolingual documents for statistical machine translation. In *Proceedings of the 4th International Workshop on Computational Terminology*. Dublin, Ireland, pages 22–31. http://www.aclweb.org/anthology/W14-4803.

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2014a. Enhancing statistical machine translation with bilingual terminology in a CAT environment. In Yaser Al-Onaizan and Michel Simard, editors, *Proceedings of AMTA 2014*. Vancouver, BC.

Nicola Bertoldi, Roldano Cattoni, Mauro Cettolo, Amin Farajian, Marcello Federico, Davide Caroselli, Luca Mastrostefano, Andrea Rossi, Marco Trombetti, Ulrich Germann, and David Madl. 2017. MMT: New open source MT for the translation industry. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*. Prague, pages 86–91. https://ufal.mff.cuni.cz/eamt2017/user-project-product-papers/papers/user/EAMT2017_paper_88.pdf.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, pages 674–679. ACL Anthology Identifier: L12-1527. http://www.lrec-conf.org/proceedings/lrec2012/pdf/886_Paper.pdf.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180. http://www.aclweb.org/anthology/W/W07/W07-0222.

European Commission, EACEA, and Eurydice. 2015. *The European Higher Education Area in 2015: Bologna Process Implementation Report*. Luxembourg: Publications office of the European Union.

Adriano Ferraresi. 2017. Terminology in European university settings. The case of course unit descriptions. In Paola Faini, editor, *Terminological Approaches in the European Context*. Cambridge Scholars Publishing, Newcastle upon Tyne, pages 20–40.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT, Phuket, Thailand, pages 79–86. http://mt-archive.info/MTS-2005-Koehn.pdf.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, StatMT '07, pages 224–227. http://dl.acm.org/citation.cfm?id=1626355.1626388.

Philippe Langlais. 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology - Volume 14*. Association for Computational Linguistics, Stroudsburg, PA, USA, COMPUTERM '02, pages 1–7. https://doi.org/10.3115/1118771.1118776.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Sanja Štajner, Andreia Querido, Nuno Rendeiro, João António Rodrigues, and António Branco. 2016. Use of domain-specific language resources in machine translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France, pages 592–598.

A little bit of bella pianura: Detecting Code-Mixing in Historical English Travel Writing

Rachele Sprugnoli[1-2], Sara Tonelli[1], Giovanni Moretti[1], Stefano Menini[1-2]
[1]Fondazione Bruno Kessler, Trento
[2] Università di Trento
{sprugnoli,satonelli,moretti,menini}@fbk.eu

Abstract

English. Code-mixing is the alternation between two or more languages in the same text. This phenomenon is very relevant in the travel domain, since it can provide new insight in the way foreign cultures are perceived and described to the readers. In this paper, we analyse English-Italian code-mixing in historical English travel writings about Italy. We retrain and compare two existing systems for the automatic detection of code-mixing, and analyse the semantic categories mostly connected to Italian. Besides, we release the domain corpus used in our experiments and the output of the extraction.

Italiano. *Il code-mixing è l'alternanza di lingue diverse nello stesso testo. Questo fenomeno è particolarmente importante nel dominio dei viaggi, poiché aiuta a comprendere meglio il modo in cui vengono percepite e descritte culture diverse da quella dell'autore. In questo lavoro, analizziamo il code-mixing tra inglese ed italiano nei testi di viaggio scritti in inglese e aventi come soggetto l'Italia. A questo scopo confrontiamo due sistemi esistenti per il riconoscimento automatico del code-mixing dopo averli ri-addestrati e analizziamo le categorie semantiche connesse alle parole/espressioni italiane. Inoltre, rilasciamo il corpus e il risultato dell'estrazione.*

1 Introduction

Code-mixing is the alternation between two or more languages that can occur between sentences (inter-sentential), within the same utterance (intra-sentential), or even inside a single token (mixing of morphemes). This phenomenon has been widely studied from the linguistic, psycholinguistic, and sociolinguistic point of view (Gardner-Chloros, 1995; Grosjean, 1995; Ho, 2007) but there is no consensus on the terminology to be adopted. In this paper code-mixing is used as an umbrella term to indicate a manifestation of language contact subsuming other expressions such as code-switching, languaging, borrowing, language crossing (Muysken, 2000).

Code-mixing characterizes communication of post-colonial, migrant and multilingual communities (Papalexakis et al., 2014; Frey et al., 2016) and it emerges in different types of documents, for example parliamentary debates, interviews and social media posts (Carpuat, 2014; Das and Gambäck, 2015; Piergallini et al., 2016). Travel writings (e.g. guidebooks, travelogues, diaries, blogs, travel articles in magazines) are affected as well by this phenomenon that has been studied in particular by analyzing small corpora of contemporary tourism discourse through manual inspection (Dann, 1996). Even if code-mixing occurs in less than 1% of the cases (Cappelli, 2013), it has several important functions in the travel domain: it gives a "linguistic sense of place" (Cortese and Hymes, 2001), it adds authenticity to a narration, it provides translation of cultural-specific words and it is a mean to define social identity ("us" tourists *versus* "they" locals) (Jaworski et al., 2003).

In this work, we investigate the phenomenon of code-mixing in travel writings, but differently from previous works we shift the focus of analysis from contemporary to historical data and from manual to automatic information extraction. As for the first point, we present a corpus of more than 3.5 millions words of English travel writings published between the end of the XIX Century and the beginning of the XX Century, which we have retrieved from freely available sources and we release in a cleaned format. As for automatic information extraction, we retrain two state-of-the-art tools to identify English-Italian code-mixing and evaluate them on a sample of our dataset. We further launch the best system on the whole dataset and then we perform a semi-automatic refinement of the automatic annotation. The corpus, the training and test data and the outcome of the extraction are available online[1].

2 Related Work

Automatic language identification of monolingual documents has a long tradition in Natural Language Processing (Hughes et al., 2006; Lui and Baldwin, 2012). More recently a new hot topic of research has emerged, that is the detection of language at word level in code-mixing texts. Dedicated workshops and evaluation exercises have been organized on this task dealing with different pairs of languages and with social media data (Choudhury et al., 2014; Solorio et al., 2014; Molina et al., 2016). The most common approach of the proposed systems is based on Conditional Random Fields (CRFs) but there are also implementations of Logistic Regression and deep learning algorithms.

To the best our knowledge, there is no previous work on the automatic identification of code-mixing in travel writing. Cappelli (2013) and Gandin (2014) have studied the phenomenon, but they have mainly used standard corpus linguistics tools, i.e. WordSmith (Scott, 2008), to analyse language contact in English guidebooks, travel blogs written by expatriates and travel articles from 2002-2012.

3 Corpus Description

Differently from the works cited in the previous Section, we focus on historical texts. To this end, we collect from Project Gutenberg[2] a corpus of travel writings about Italy written by English native authors and published between the country unification and the beginning of the 30's. We choose this period because in the second half of the XIX Century the tradition of the Grand Tour declined and leisure-oriented travels emerged. This radical transformation was enabled by technological, economic and sociological, factors, such as the development of steam-powered ships and of the railway network, the growth of Anglo-American economy and a greater emancipation of women with more female travelers (Schriber, 1995). Moreover, after unification, new routes to Southern Italy and the islands were opened, so that travelers' attention was no longer limited to the classic destinations in the North and Central Italy, such as Venice, Florence and Rome (Ouditt and Polezzi, 2012).

The corpus is made by 57 texts[3], divided into *travel narratives* (reports, diaries, collections of letters) and *guidebooks*, for a total of 3,630,781 tokens. We distinguish between these two types of text, following a standard classification of documents in the travel domain. However, the distinction was not so clear-cut in the period we take into account as it is now, since reports on personal travel experiences were often mixed with practical recommendations and long disquisitions on art and history. Therefore, we adopt as a rule of thumb the distinction suggested in (Santulli, 2007): travel narratives are those told in the first person, while guidebooks are written in impersonal form.

The authors of the selected texts belong to different nationalities (UK, US, Ireland, Australia) and are both male and female. Some books dwell on specific cities or regions, others cover different parts of Italy or even several countries: in the latter case we extracted only the chapters related to Italy. Although we made an effort to have a diverse, well-balanced corpus in terms of content, author's gender and nationality, this was only partially possible because of the limited availability of online travel books whose text is freely available and cleaned from OCR errors. The distribution of tokens according to the year of publication and type of text is shown in Fig. 1. Details about authors are given in a spreadsheet provided together with the corpus.

4 Code-Mixing Detection

In this Section we describe the experiments on code-mixing, comparing the performance of two available systems in different configurations. We also detail the post-processing step introduced to refine the output of the best performing system.

[1] `https://dh.fbk.eu/technologies/code-mixing`

[2] `https://www.gutenberg.org/`

[3] Thirty of these texts are also available in TEI-XML format on the website `https://sites.google.com/view/travelwritingsonitaly`.

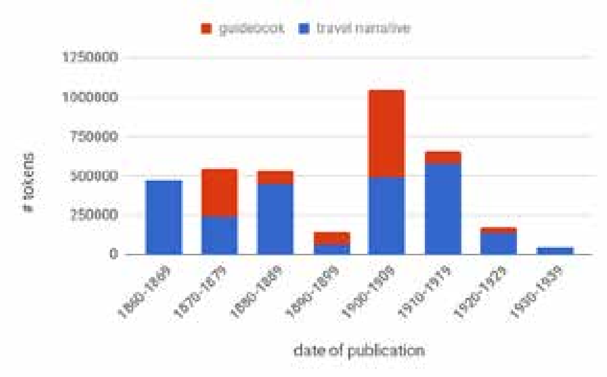


Figure 1: Distribution of tokens per year of publication and sub-genre.

4.1 Experimental Setting

In order to automatically extract Italian words, expressions and sentences from the corpus described in Section 3, we train and test two systems whose source code is available on the web. The first one (henceforth, *langid*) is based on character n-grams (n = 1 to 5) and adopts a weakly supervised approach, i.e. training data are monolingual texts of few thousand tokens (King and Abney, 2013). This system includes four classification algorithms: Conditional Random Field (CRF), Hidden Markov Model (HMM) and Maximum Entropy Model with and without generalized expectation criteria (MaxEnt-GE and MaxEnt). *langid* has been successfully evaluated on documents containing English texts mixed with 30 different minority languages such as Zulu and Chippewa[4].

For our experiments, we retrain *langid* using a collection of about 300,000 tokens taken from monolingual Italian and English books, of different genres, published in the same period of our corpus[5].

The second system (henceforth, *CodeSwitching*), has been developed to detect languages in texts mixing Latin and Middle English (Schulz and Keller, 2016). It implements a CRF classifier with features generated from TreeTagger models and word lists of both languages[6]. Differently from *langid* that classifies words as belonging to one language rather than the other, this latter system performs a fine-grained annotation by distinguishing five classes (see below). Since this system is fully supervised, we create a training set by manually annotating 3,900 tokens from 4 samples extracted from our corpus, a size in line with the training data used in the original paper. The training data were annotated with 5 different classes: Italian tokens (*i*), English tokens (*e*), punctuation (*p*), named entities (NEs) (*n*), and ambiguous tokens that belong to the dictionary of both languages (*a*).

Both *langid* and *CodeSwitching* were evaluated on the same test set, i.e. two samples of texts (one from a travel narrative and one from a guidebook) of 1,623 tokens. The test set was annotated by assigning to each token a label for English or Italian, as required by *langid*, and also marking punctuation, NEs and ambiguous tokens, following *CodeSwitching* scheme. Since the performance of *langid* is sensitive to the length of the input file, we split the test set in batches of 40 sentences, replicating the experimental setting presented in (Schulz and Keller, 2016).

4.2 Evaluation

Table 1 presents the performances of *langid* on the test set: contrary to the results achieved by King and Abney (2013), HMM – not CRF – proved to be the best approach. This is likely due to the greater sparseness of the code-mixing phenomenon in our dataset with respect to what was registered in the original corpus, where languages different from English cover the 56% of the overall number of tokens.

Table 2 reports Precision, Recall and F-measure of the retrained *CodeSwitching* system. Even if the overall performance is slightly better than the one obtained with HMM in *langid*, the scores for the detection of Italian tokens (*i*) are lower (0.82 versus 0.90 in terms of F-measure). Punctuation (*p*) and ambiguous tokens (*a*) are generally detected with a good performance, while NEs (*e*) represent the most challenging class. Given that we are mainly interested in recognising English and Ital-

[4] `http://www-personal.umich.edu/~benking/resources/langid_release.tar.gz`

[5] For Italian: "Le Avventure di Pinocchio" by C. Collodi, "Una donna" by S. Aleramo, "Il Valdarno da Firenze al mare" by G. Carocci, "La vita operosa" by M. Bontempelli, "Dopo il divorzio" by G. Deledda, "Novelle umoristiche" by A. Albertazzi, "Lezioni e Racconti per i bambini" by I. Baccini. For English: "The Adventures of Tom Sawyer" by M. Twain, "Pioneers of the Old Southwest" by C. L. Skinner, "The Happy Prince, and Other Tales" by O. Wilde, "Vanished Arizona" by M. Summerhayes, "The Tale of Peter Rabbit" by B. Potter, "The Strange Case of Dr. Jekyll and Mr. Hyde" by R. L. Stevenson.

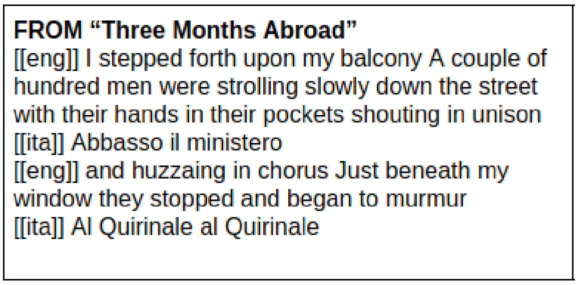
[6] `https://github.com/sarschu/CodeSwitching`

	CRF	HMM	MaxEnt	MaxEnt-GE
P	1	**0.89**	0.59	0.82
R	0.51	**0.92**	0.90	0.47
F	0.67	**0.90**	0.71	0.60

Table 1: Results of the evaluation on the retrained *langid* system in terms of precision (P), recall (R), and F-Measure (F).

	i	*e*	*a*	*n*	*p*	ALL
P	**0.83**	0.98	0.98	0.85	0.98	0.92
R	**0.80**	0.99	0.90	0.85	0.96	0.90
F	**0.82**	0.99	0.94	0.85	0.97	0.91

Table 2: Results of the evaluation on the retrained *CodeSwitching* system in terms of precision (P), recall (R), and F-Measure (F) for each class and the macro-average of all classes.



FROM "Three Months Abroad"
[[eng]] I stepped forth upon my balcony A couple of hundred men were strolling slowly down the street with their hands in their pockets shouting in unison
[[ita]] Abbasso il ministero
[[eng]] and huzzaing in chorus Just beneath my window they stopped and began to murmur
[[ita]] Al Quirinale al Quirinale

Figure 2: Examples of *langid* output.

ian terms, and that on this task *langid* performs better, we run this tool on the whole corpus.

4.3 Post-processing

In order to refine the output of *langid* (see Figure 2), we perform three post-processing steps. First of all, we check whether tokens tagged as Italian are included in Morph-it, an Italian lexicon of inflected forms (Zanchetta and Baroni, 2005): in this way we are able to detect false positives. Then, we run the Polyglot Python module on the corpus to find out if the processed documents contain other languages beside English and Italian[7]. Indeed 27 books result to have a high probability of including text written also in Latin, French, Germany or Greek. These books are likely to be problematic given that *langid* recognizes only English and Italian. Information obtained in these two steps are then used to manually check the outcome of *langid* extraction and correct it semi-automatically. Furthermore, we employ the USAS Italian semantic tagger (Piao et al., 2015) to obtain a categorization of the terms tagged as Italian. Based on the 21 semantic classes recognised by USAS, we are able to understand in which cases and why writers used to switch their narration from English to Italian.

5 Discussion

The classification performed with the USAS tagger shows that Italian is adopted to express concepts covered by 20 semantic classes, both in guidebooks and in travel narratives. Only one USAS class, the one related to "Science and technology", is not found in the corpus. Table 5 shows frequency and examples for each detected class. As in contemporary travel writings (Francesconi, 2007), food is well represented: traditional dishes, drinks and products (e.g. *polenta*, *Chianti*, *mortadella*) appear together with fruits, vegetables (e.g. *mandarini*, *finocchio*) and also eating establishments (e.g. *osteria*, *trattoria*, *locanda*). The attention for Italian art and architecture manifests itself through the use of many specialized terms (*cassettoni*, *gotico*, *giallo antico*). The semantic areas of emotions and psychological processes are not recorded in previous work on contemporary texts but are frequent especially in travel reports (e.g. *addolorata*, *trionfo*, *simpatico*). As for NEs, city names reveal an increasing interest for towns in Central regions (for example, *Perugia* has a high frequency of occurrence in both genres). Moreover, following Italy unification, travellers discovered several locations in the South (e.g. *Ragusa*, *Catanzaro*). Among the most mentioned people, there are representatives of past Italian politics (e.g. *Lorenzo and Cosimo de Medici*), artists (e.g. *Giotto*, *Dante*) and religious figures (e.g. *Madonna*, *San Michele*).

In many cases, the use of Italian is not limited to single words or multi-token expressions (e.g. *appartamento signorile*) but longer utterances are reported. Texts of both genres contain proverbs (e.g. *chi tardi arriva mal alloggia*) and citations, not only from the canon of Italian literature, such as Leopardi's poems, but also from the popular tradition, such as Tuscan songs (*O rosa O rosa O rosa gentillina*). The main difference between travel narratives and guidebooks is the greater presence in the former of dialogues or expressions heard by the author during his/her stay in Italy (*voi siete un*

[7] `http://polyglot.readthedocs.io/en/latest/Installation.html`

GUIDEBOOKS			TRAVEL NARRATIVES		
SEMANTIC CLASS	**#**	**EXAMPLES**	**SEMANTIC CLASS**	**#**	**EXAMPLES**
names & grammar	29,927	*Pisa*	names & grammar	28,694	*Donatello*
architecture	3,070	*villa*	social elements	3,134	*popolo*
movement	2,294	*automobile*	architecture	3,065	*palazzo*
social elements	1,590	*trinità*	environment	1,311	*lago*
materials & objects	717	*fontana*	movement	1,207	*vetturino*
environment	713	*campagna*	materials & objects	965	*rosso*
general/abstract terms	580	*essere*	general/abstract terms	943	*fare*
measurement	340	*alto*	food & farming	665	*trattoria*
arts & crafts	231	*stucco*	life	479	*fiore*
time	225	*nuovo*	measurement	464	*grande*
life	222	*agnello*	time	379	*primavera*
body	211	*cintola*	body	350	*braccio*
public domain	205	*podestà*	psyche	330	*vedere*
psyche	198	*volere*	entertainment	319	*marionetta*
food & farming	162	*maccaroni*	money & commerce	269	*dazio*
entertainment	141	*giuoco*	communication	268	*dire*
emotion	137	*amore*	public domain	260	*carabiniere*
communication	131	*motto*	arts & crafts	206	*arte*
money & commerce	127	*soldo*	emotion	176	*evviva*
education	22	*università*	education	135	*maestro*

Table 3: Italian word frequency for each semantic class

cattivo; e voi siete bella).

6 Conclusions and Future Work

In this work, we presented the first automated analysis of code-mixing in historical travel writings. In particular, we focus on English documents about Italy, and we compare guidebooks and travel narratives, analysing the semantic categories mostly related to code-mixing.

In the future, we plan to investigate how code-mixing phenomena relate to content types in travel writings (Sprugnoli et al., 2017). Besides, we are planning to implement an algorithm to automatically link code-mixing quotations to their original source text. Finally, we would like to extend our experiments to recognise code-mixing in multiple languages, and compare the semantic domains specific to each language.

References

Gloria Cappelli. 2013. Travelling words: Languaging in english tourism discourse. *Travels and translations*, pages 353–374.

Marine Carpuat. 2014. Mixed-language and code-switching in the Canadian Hansard. In *Proceedings of EMNLP 2014*, page 107.

Monojit Choudhury, Gokul Chittaranjan, Parth Gupta, and Amitava Das. 2014. Overview of FIRE 2014 track on transliterated search. In *Proceedings of FIRE*.

Giuseppina Cortese and Dell Hymes. 2001. Languaging in and across human groups. *Perspectives on difference and asymmetry. Textus. English Studies in Italy*, 14(2).

Graham MS Dann. 1996. *The language of tourism: a sociolinguistic perspective.* Cab International.

Amitava Das and Björn Gambäck. 2015. Code-mixing in social media text: the last language identification frontier? *Revue TAL*, pages 41–64.

Sabrina Francesconi. 2007. Italian borrowings from the semantic fields of food and drink in English tourism texts. *The Languages of Tourism: turismo e mediazione, Milano: Unicopli*, page 129.

Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W Stemle. 2016. The DiDi Corpus of South Tyrolean CMC Data: A Multilingual Corpus of Facebook Texts. In *Proceedings of CLIC-it*.

Stefania Gandin. 2014. Investigating loan words and expressions in tourism discourse: a corpus driven analysis on the BBC-travel corpus. *European Scientific Journal*, 10(2).

Penelope Gardner-Chloros. 1995. Code-switching in community, regional and national repertoires: the myth of the discreteness of linguistic systems. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 68–89.

François Grosjean. 1995. A psycholinguistic approach to code-switching: The recognition of guest words by bilinguals. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 259–275.

Judy Woon Yee Ho. 2007. Code-mixing: Linguistic form and socio-cultural meaning. *The International Journal of Language Society and Culture*, 21.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006. Reconsidering language identification for written language resources. In *Proc. International Conference on Language Resources and Evaluation*, pages 485–488.

Adam Jaworski, Crispin Thurlow, Sarah Lawson, and Virpi Ylänne-McEwen. 2003. The uses and representations of local languages in tourist destinations: A view from British TV holiday programmes. *Language Awareness*, 12(1):5–29.

Ben King and Steven P Abney. 2013. Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of HLT-NAACL*, pages 1110–1119.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Giovanni Molina, Nicolas Rey-Villamizar, Thamar Solorio, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, and Mona Diab. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of EMNLP 2016*, pages 40–49.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Sharon Ouditt and Loredana Polezzi. 2012. Introduction: Italy as place and space. *Studies in Travel Writing*, 16(2):97–105.

Evangelos Papalexakis, Dong-Phuong Nguyen, and A Seza Doğruöz. 2014. Predicting code-switching in multilingual communication for immigrant communities. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics.

Scott Piao, Francesca Bianchi, Carmen Dayrell, Angela D'egidio, and Paul Rayson. 2015. Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015)*. Association for Computational Linguistics.

Mario Piergallini, Rouzbeh Shirvani, Gauri S Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in Swahili-English language data. In *Proceedings of EMNLP 2016*.

Francesca Santulli. 2007. Le parole e i luoghi: descrizione e racconto. In Donelli Antelmi, Gudrun Held, and Francesca Santulli, editors, *Pragmatica della comunicazione turistica*, pages 81–153. Editori. Riuniti, Roma.

Mary Suzanne Schriber. 1995. Women's Place in Travel Texts. *Prospects*, 20:161179.

Sarah Schulz and Mareike Keller. 2016. Code-switching ubique est – Language identification and part-of-speech tagging for historical mixed text. In *Proceedings of LaTeCH Workshop*.

Mike Scott. 2008. WordSmith tools version 5. *Liverpool: Lexical Analysis Software*, 122.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Rachele Sprugnoli, Tommaso Caselli, Sara Tonelli, and Giovanni Moretti. 2017. The Content Types Dataset: a new Resource to Explore semantic and functional Characteristics of Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 260–266, Valencia, Spain, April. Association for Computational Linguistics.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! A free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1).

PARAD-it: Eliciting Italian Paradigmatic Relations with Crowdsourcing

Irene Sucameli
University of Pisa
Pisa, Italy
irenesucameli@gmail.com

Alessandro Lenci
University of Pisa
Pisa, Italy
alessandro.lenci@unipi.it

Abstract

English. In this paper, we present a new dataset of semantically related Italian word pairs. The dataset consists of nouns, adjectives and verbs together with their synonyms, antonyms and hypernyms. The data have been collected with crowdsourcing from a pool of Italian native speakers. The dataset, the first of its kind, is useful not only to evaluate computational models of Italian semantic relations, but also for linguistic and psycholinguistic investigations of the mental lexicon.

Italiano. *In questo articolo si presenta un nuovo dataset di parole italiane legate da relazioni semantiche. L'analisi si basa su una raccolta di nomi, verbi e aggettivi a cui sono stati associati sinonimi, antonimi e iperonimi. I dati sono stati raccolti da un gruppo di parlanti nativi di italiano tramite crowdsourcing. Il dataset, primo del suo tipo, è utile per valutare modelli computazionali relativi alle relazioni semantiche dell'italiano, per la ricerca linguistica teorica e psicolinguistica.*

1 Introduction

The present project aims at providing new data about the internal organization of the Italian lexicon. For this purpose, we present PARAD-it[1] a paradigmatic relation dataset elicited from Italian native speakers with crowdsourcing. This dataset consists of a set of target words selected from the Italian section of MultiWordNet paired with relata belonging to different kinds of paradigmatic semantic relations. The data have been collected using the same method adopted by Scheible and Schulte im Walde (2014) for German and by Benotto (2015) for English, thereby making the three datasets fully comparable for crosslingual analyses. PARAD-it is a collection of hypernyms, antonyms, and synonyms for a set of Italian nouns, adjectives and verbs.

2 Related Works

Our contribution is just the latest in a series of recent works aimed at eliciting judgments about semantic relations, to develop testsets for computational models. Besides Scheible and Schulte im Walde (2014) and Benotto (2015), we can mention BLESS, realized by Baroni and Lenci (Baroni and Lenci, 2011). Bless is a dataset created for the evaluation of distributional semantic models. The BLESS dataset includes 200 English nouns, equally divided into animate and inanimate entities. Each noun is associated to multiple relata belonging to five types of relations: hyperonymy, co-hyponymy, meronymy, attributes and events.

Another relevant project is EVALution. This dataset combines data extracted from Concept-Net 5.0 (Liu and Singh, 2004) and WordNet 4.0 (Fellbaum, 1998), and then checked by native speakers. The crowdsourcing task consisted in rating the truthfulness of sentences generated from the selected word pairs, according to templates indicative of various semantic relations and to be used as a proxy for the prototypicality of the relations. PARAD-it extends this line of research to Italian for the first time.

3 Collecting PARAD-it

3.1 Target Selection

The PARAD-it targets were extracted from the Italian section of the MultiWordNet database (Pianta, Bentivogli and Girardi, 2002) .

[1] PARAD-it is freely distributed and it will be available for download from: http://colinglab.humnet.unipi.it/resources/

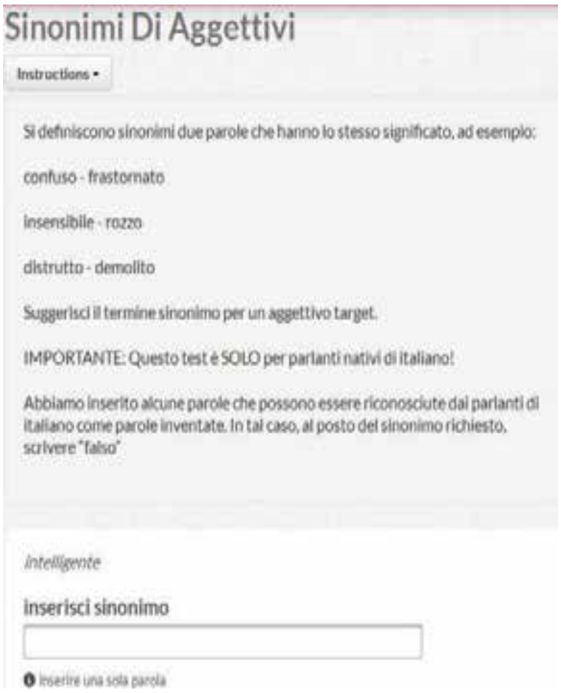
The selection of nouns, adjectives and verbs was balanced for:[2]

- **Frequency** - three frequency classes were identified using the itWaC corpus (Baroni et al. 2009): i.) words with frequency from 200 to 2999, ii.) words with frequency from 3,000 to 9,999, and iii.) words with frequency greater than 10,000.
- **Polysemy -** three polisemy classes were identified, according to the number of synsets in MultiWordNet: i.) words with one synset, ii.) words with two synsets, iii.) words with three or more synsets.

Then, 11 targets were randomly sampled for each class, making a total of 99 targets for each PoS.

3.2 Data Elicitation

Italian native speakers were asked to produce, for each target word, a synonym, an antonym and a hypernym. The data were collected through CrowdFlower,[3] a crowdsourcing web-based platform to design various data collection tasks (i.e., sentiment analysis, data categorization, etc.) thanks to the help of external workers which are paid according to the type of task.



Sinonimi Di Aggettivi

Si definiscono sinonimi due parole che hanno lo stesso significato, ad esempio:

confuso - frastornato

insensibile - rozzo

distrutto - demolito

Suggerisci il termine sinonimo per un aggettivo target.

IMPORTANTE: Questo test è SOLO per parlanti nativi di italiano!

Abbiamo inserito alcune parole che possono essere riconosciute dai parlanti di italiano come parole inventate. In tal caso, al posto del sinonimo richiesto, scrivere "falso"

intelligente

inserisci sinonimo

Inserire una sola parola

Figure 1: Example of CrowdFlower form

In the present project, we collected data from ten subjects, for each target word, and for each semantic relation. In order to guarantee that the tasks would be completed only by Italian native speakers, the CrowdFlower form also included a test to discriminate Italian words from “pseudo words”. The responses produced by subjects that failed to pass the test were excluded. All the elicited data were then manually normalised: Typing errors were corrected and the words written in lower case and capital letters were mapped onto a single standard form.

3.3 Results

The number of responses for each PoS and each relation type is shown in Table 1. The lowest number of responses concerns mainly antonyms and then hypernyms. This is due to the fact that antonyms are characterized by a high degree of canonicity (Paradis and Willners 2011, de Weijer et al. 2012). For this very reason, it may be more difficult for a speaker to provide an antonym for a input word since he can rely only on a small group of possible answers.

Compared to antonyms and hypernyms, synonyms are more easily identified by users. In fact, 2,674 tokens have been provided for this paradigmatic relation. However, if we consider the number of types, instead of the number of tokens, the situation is different. In fact, with 1,528 types, the relation of hypernymy is the one with the highest number of types produced. This result shows that, even if for the users it is simpler to provide a synonym for a given target, words have in general a lower number of distinct synonyms. On the other hand, the users have provided less responses for the hypernyms but more differentiated. This might due to the fact that taxonomies (typical of hypernyms) have different levels of depth (Murphy, 2010). Concerning the target PoS, verbs have elicited the highest number of responses, possibly because of their inherent higher polysemy (Murphy, 2010). These results regarding the identification of verbs and hypernyms by native speakers are in line with those obtained by Scheible and Schulte im Walde for German and with those produced by Benotto for English.

[2] The balancing parameters are the same used by Scheible and Schulte im Walde (2014) and by Benotto (2015).

[3] https://www.crowdflower.com

	ANT		HYP		SYN		all	
	types	tokens	types	tokens	types	tokens	types	tokens
Adj	269	805	435	706	455	853	1159	2364
Noun	306	493	570	843	453	883	1329	2219
Verb	444	849	523	915	466	938	1433	2702
all	1019	2147	1528	2464	1374	2674	3921	7285

Table 1: Number of total responses

	ANT+SYN		HYP+SYN		ANT+HYP		ANT+HYP+SYN	
	types	tokens	types	tokens	types	tokens	types	tokens
Adj	3	15	182	883	3	27	0	0
Noun	48	195	109	541	35	140	21	147
Verb	55	243	214	916	45	208	39	330
all	106	453	505	2340	83	357	60	447

Table 2: Ambiguous responses

As an additional level of analysis, we have identified the ambiguous responses (Table 2). When users have provided the same response for different paradigmatic relation, that response has been considered as ambiguous. Here, the highest number of ambiguity has been recorded in relation to the synonymy-hypernymy pair. Actually, this high number of ambiguity was expected and the result seems to be reasonable since it is similar to the one obtained by Scheible and Schulte im Walde for German (with 470 types recorded as ambiguous within the couple synonymy-hypernymy). This result may depend on the fact that in many cases the distinction between synonymy and hypernymy is blurred or not easily identifiable, especially for more abstract items. For instance, the target *mattino* ('morning') has prompted the word *giorno* ('day') both as synonym and as hypernym.

Concerning the different responses provided by subjects (Figure 2), we saw that a) speakers are mostly in agreement referring to the relation of antonymy, consistently with the trend in the parallel English and German data; b) only in few cases more than 7 different responses have been provided for the same input, while c) in most cases between 3 and 5 different responses have been indicated for target.

This suggests that Italian native speakers do not tend to have one-to-one lexical associations. At the same time, they tend to identify a reduced group of terms that can be used with a certain relation.

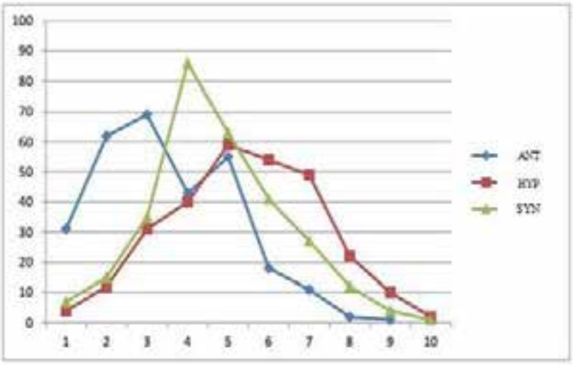


Figure 2: Targets for different responses.
The Y axis reports the number of targets provided by users while the X axis reports the number of different responses per input

Figure 3 and Figure 4 show the production of frequency distribution among classes and relations.

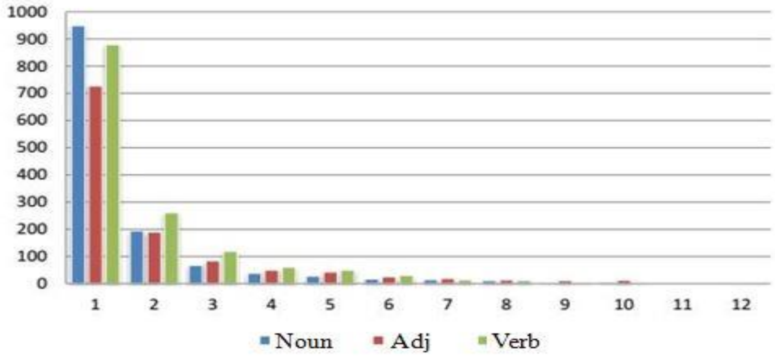


Figure 3: Distribution of production frequency among classes

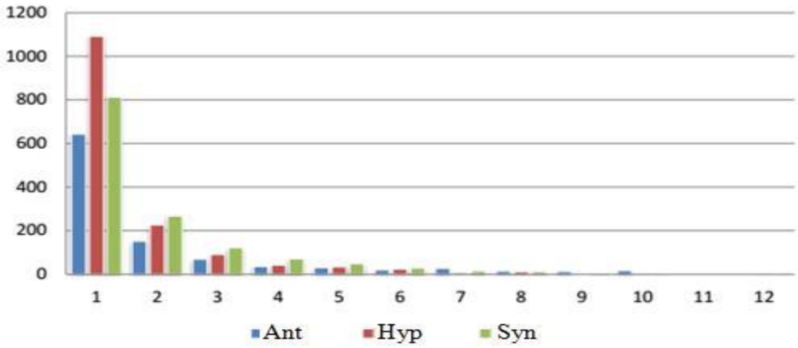


Figure 4: Distribution of production frequency among relations

Concerning the distribution among classes, 949 nouns have been produced by users only once. On the other hand, verbs have 879 hapax responses, and adjectives 727. From Figure 4, it is possible to observe that hypernyms have the highest number of hapax. In fact, for this relation there are 1,090 hapax, while synonymy has 812 hapax and antonymy only 643. This result is due to the existence of canonicity relations for antonymy, and to the notorious paucity of true synonyms.

3.4 Distributional Semantic Analysis of the Elicited Data

A distributional space has been built in order to analyse the synonyms, antonyms and hypernyms produced by subjects. Distributional Semantic Models (DSMs) use corpus co-occurrences to measure the similarity/relatedness between two words: The closer two vectors are in distributional space, the more semantically related the two words are.

We used DISSECT (DIStributional SEmantic Composition Toolkit) to train a standard count-based DSM on the *Repubblica* corpus, a corpus made up of newspaper articles with over 300 million tokens. Our targets and contexts include the PARAD-it data plus all the content words in *Repubblica* with frequency greater than 200. Co-occurrences have been extracted, using a context window of 2 content words to the left and right of each target item. For each PARAD-it relatum, we measured its cosine with the target word, using PPMI (Positive Pointwise Mutual Information) as weighting scheme, and truncated SVD (Singular Value Decomposition) to 300 latent dimensions. Figure 5 and Figure 6 report the boxplot summarizing the cosine distribution by semantic relation and by PoS.

The analysis shows that there are no significant differences in the cosine median neither between different types of relations nor between different grammatical classes. As shown in Figure 5, the highest cosine values have been recorded for antonyms (over 0.90). This is due to the fact that this type of relation is characterized by a high rate of canonicity. On the other side, hypernyms show the greatest median values (0.76).

Concerning the distribution of relata cosine by PoS, nouns have the highest cosine values, while adjectives and verbs show a more reduced variability. These results are coherent with the production data. Indeed, as we saw above, high frequency values were recorded both for nouns and hypernyms while speakers' production show a greater homogeneity in responses for the relation of antonymy.

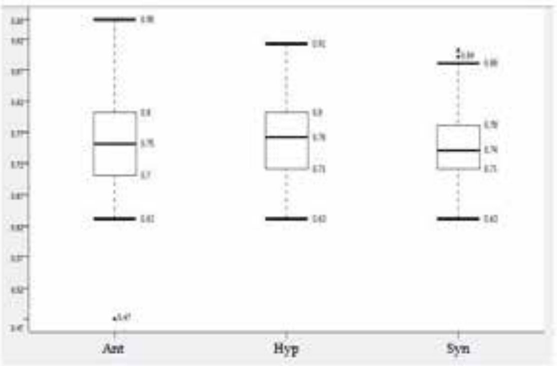


Figure 5: Distribution of relata cosines by relations

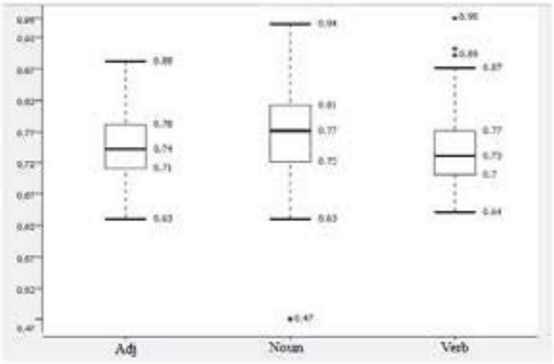


Figure 6: Distribution of relata cosines by target PoS

3 Conclusion

This project presents PARAD-it, a new collection composed by pairs of Italian nouns, verbs and adjectives related by different types of paradigmatic relations, elicited by native speakers with crowdsourcing. Starting from this new resource, a quantitative analysis was carried out to analyze the mechanisms underlying the Italian language. In particular, the analysis has shown that: i) high frequency values tend to be recorded for nouns and hypernyms while ii) Italian speakers tend to use a more uniform vocabulary to describe the relation of antonymy. This analysis has revelead some interesting differences in the response distribution both with respect to the PoS of the target, and with respect to the semantic relation. Moreover, this study confirms the differential salience of the various paradigmatic relations in organizing the mental lexicon.

To the best of our knowledge, PARAD-it is the first, freely available resource of this kind for Italian, paving the way for its use as a test set for computational models of semantic relation identification and classification. For future research, we plan to realize and additional round of crowdsourcing in order to validate the words previously produced, checking also if there is an overlap between these words and the targets from MultiWordNet. Moreover, we plan to carry out a crosslingual comparison with the similar datasets collected for German and for English.

References

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, Marco Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. *LREC, European Language Resources Association*.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. In *Language Resources and Evaluation*, Vol. 43, Issue 3, pp.209-226.

Marco Baroni, Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pp.1–10. Association for Computational Linguistics.

Giulia Benotto. 2015. *Distributional Models for Semantic Relations: A study on Hyponymy and Antonymy*. PhD thesis Pisa, Italia.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge.

Hugo Liu, Push Singh. 2004. ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal* Vol. 22 pp. 211-226. Kluwer Academic Publishers.

M. L. Murphy. 2010. *Lexical Meaning.* Cambridge University Press, Cambridge.

Carita Paradis, Caroline Willners. 2011. Antonymy: From convention to meaning-making. *Review of cognitive linguistics*, pp.367-391.

Emanuele Pianta, Luisa Bentivogli, Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database .In *Proceedings of the First International Conference on Global WordNet* pp.293-302. Mysore, India.

Adriana Roventini, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Christian Girardi, Bernardo Magnini, Rita Marinelli, and Antonio Zampolli. 2003. Italwordnet: building a large semantic database for the automatic treatment of italian. *Computational Linguistics in Pisa*, Special Issue, Vol. 18-19, 2:745–791, IEPI, Pisa-Roma.

Entico Santus, Alessandro Lenci, Frances Young, e Chu-Ren Huang. 2015. EVALution 1.0: an Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics,* pp. 64-69. Beijing.

Silke Scheible, Sabine Schulte im Walde. 2014. A Database of Paradigmatic Semantic Relation Pairs for German Nouns, Verbs, and Adjectives. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing,* pp. 111-119. Dublin, Ireland.

Joost van de Weijer, Carita Paradis, Caroline Willners & Magnus Lindgren. 2012. *As lexical as it gets: the role of co-occurrence of antonyms in a visual lexical decision experiment*. In D. Divjak & St. Th. Gries (Eds.). *Frequency effects in language: linguistic representation*. De Gruyter Mouton. pp. 255-279.

The impact of phrases on Italian lexical simplification

Sara Tonelli, Alessio Palmero Aprosio
Fondazione Bruno Kessler
Trento, Italy
{satonelli,aprosio}@fbk.eu

Marco Mazzon
Dept. of Psychology and Cognitive Science
University of Trento
marco.mazzon@studenti.unitn.it

Abstract

English. Automated lexical simplification has been performed so far focusing only on the replacement of single tokens with single tokens, and this choice has affected both the development of systems and the creation of benchmarks. In this paper, we argue that lexical simplification in real settings should deal both with single and multi-token terms, and present a benchmark created for the task. Besides, we describe how a freely available system can be tuned to cover also the simplification of phrases, and perform an evaluation comparing different experimental settings.

Italiano. *La semplificazione lessicale automatica è stata affrontata fino ad ora dalla comunità di ricerca TAL concentrandosi sulla sostituzione di parole singole con altre parole singole. Questa modalità ha condizionato sia lo sviluppo di sistemi di semplificazione che la creazione di benchmark per la valutazione. In questo articolo, sosteniamo che la semplificazione lessicale in contesti reali debba includere sia parole singole che espressioni composte da più parole, e presentiamo un benchmark creato a questo fine. Inoltre, descriviamo come adattare un sistema disponibile per la semplificazione lessicale in modo che supporti anche la semplificazione di sintagmi, e presentiamo una valutazione confrontando diversi setting sperimentali.*

1 Introduction

Lexical simplification is a well-studied topic within the NLP community, dealing with the automatic replacement of complex terms with simpler ones in a sentence, in order to improve its clarity and readability. Thanks to the development of benchmarks (Paetzold and Specia, 2016a) and freely available tools for lexical simplification (Paetzold and Specia, 2015), a number of works have focused on this challenge, see for example the systems participating in the simplification shared task at SemEval-2012 (Specia et al., 2012). However, the task has been designed as an exercise to replace complex *single tokens* with simpler *single tokens*, and most widely used benchmarks and systems all follow this paradigm. We believe, however, that this setting covers only a limited number of lexical simplifications as they would be performed in a real scenario. In particular, we advocate the need to shift the lexical simplification paradigm from single tokens to phrases, and to develop datasets and tools that deal also with these cases. This is mainly the contribution of this work, which covers four main points:

- We analyse existing corpora of simplified texts, not specifically developed for a shared task or for system evaluation, and we measure the impact of phrases in lexical simplifications
- We modify a state-of-the-art tool for lexical simplification in order to support phrases
- We compare different strategies for phrase extraction and evaluate them over a benchmark
- We perform all the above on Italian, for which there was no lexical simplification system available.

Besides, we make freely available the first benchmark for the evaluation of Italian lexical simplification, with the goal to support research on this task and to foster the development of Italian simplification systems.

2 Corpus analysis and Benchmark creation

We first analyse existing simplification corpora in Italian to study the impact of phrases on lexical simplification. There are only two such manually created corpora, which contain different types of data but have been annotated following the same scheme: the Simpitiki corpus (Tonelli et al., 2016) and the one developed by the ItaNLP Lab in Pisa (Brunato et al., 2015). The former contains 1,163 sentence pairs[1], where one is the original sentence and the other is the simplified one. The pairs were created starting from Wikipedia edits and from documents in the public administration domain. The ItaNLP corpus, instead, contains 1,393 pairs extracted from children's stories and from educational material. Both corpora were annotated following the scheme proposed in (Brunato et al., 2015), in which simplifications were classified as *Split*, *Merge*, *Reordering*, *Insert*, *Delete* and *Transformation* (plus a set of subclasses for the *Insert*, *Delete* and *Transformation* cases). Since our goal was to isolate a benchmark of pairs containing only the lexical cases, we discarded the classes not compatible with lexical simplifications (e.g. *Delete*, *Reordering*) and then manually checked the others to identify the cases of interest. When, as in the majority of cases, a lexical simplification was present together with other simplification types, we re-wrote the target sentence in order to retain only lexical cases. For example, in the examples below, *a)* is the original sentence and *b)* is the simplified one in the Simpitiki corpus, which contains a lexical simplification of 'include' and a shift of position of 'per convenzione'. We created version *c)*, so that only the lexical simplification is present:

a) *Eurasia è il termine con cui per convenzione si definisce la zona geografica che include l'Europa e l'Asia.*
b) *Eurasia è, per convenzione, il termine con cui si definisce la zona geografica che comprende l'Europa e l'Asia.*
c) *Eurasia è il termine con cui per convenzione si definisce la zona geografica che comprende l'Europa e l'Asia.*

This revision process led to the creation of a benchmark with pairs extracted from the two original corpora, where only cases of lexical simplification are present[2]. Some statistics related to the benchmark are reported in Table 1. We identify four possible lexical simplification types: a single token is replaced by a single token (ST→ST), a single token is simplified through a phrase (ST→P), a phrase is simplified through a single token (P→ST), and a phrase is replaced by another phrase (P→P).

	ST→ST	ST→P	P→ST	P→P	Total
ItaNLP	369	112	139	87	707
Simpitiki	112	24	30	28	194
Total	481	136	169	115	901

Table 1: Statistics on lexical simplification benchmark (ST = Single token, P = Phrase)

We observe that the most frequent lexical simplification type is ST→ST, on which most systems and shared tasks are based. However, this simplification type covers only half of the cases included in our benchmark. This confirms the need to include cases of phrase-based simplification in the creation of benchmarks. It corroborates also the importance of developing systems for lexical simplification that support phrase replacement, so as to make them work in real settings and not only on ad-hoc test sets. Another interesting remark is that single tokens are not necessarily simpler than phrases, or vice versa: in our data, there are 136 ST→P and 169 P→ST, showing that no general rule can be applied to favour (or demote) Ps over STs.

We use the final benchmark[3], containing 901 sentence pairs, to evaluate a system for lexical simplification taking into account phrases, as described in the following Section.

3 Automated lexical simplification

In this Section we describe the experiments we carried out to perform automated lexical simplification using the benchmark presented in Section 2. We describe the tool used and how it was mod-

[1] The number is slightly different from what was reported in the original paper because the corpus was revised after the first release.

[2] In Simpitiki we focused only on the pairs in the public administration domain due to project constraints. We plan to include the pairs from Wikipedia in the next benchmark version.

[3] Available at `https://drive.google.com/file/d/0B4QAWZllD-egYS0yNWZ5dTdYQVE/view?usp=sharing`

ified to deal with phrases. We also detail the resources (language model and word embeddings) created for the task.

3.1 The Lexenstein system

We use Lexenstein (Paetzold and Specia, 2015), an open source tool for lexical simplification, to collect a list of candidates that should replace a given word in the text. In particular, the Paetzold generator (Paetzold and Specia, 2016b) is based on an unsupervised approach to produce simplification candidates using a context-aware word embeddings model: features used for the selection include word2vec vectors (Mikolov et al., 2013), language model created by SRILM (Stolcke, 2002), and conditional probability of a candidate given the PoS tag of the target word. So far, no evaluation on Lexestein for Italian is available.

For each complex word, five candidate replacements are first retrieved, ranked according to several features, such as n-gram frequencies and word vector similarity with the target word, and then reranked according to their average rankings (Glavaš and Štajner, 2015).

Since we wanted to test different strategies to create the embeddings (i.e. with and without phrases), we created the word/phrase vectors and the language model starting from freely available corpora (1.3 billion words in total): the Italian Wikipedia,[4] OpenSubtitles2016 (Lison and Tiedemann, 2016),[5] PAISÀ,[6] and the Gazzetta Ufficiale,[7] a collection of Italian laws. Due to the size of the data, both the corpus and the model are available upon request to the authors.

3.2 Experimental Setup

We conduct several experiments to evaluate the quality of lexical simplification when taking into account phrases (or not), and compare different strategies for phrase recognition. We compare different variants to create the embeddings and the language model (LM) that were then used by Lexenstein.

The first *baseline model* relies on the standard Lexenstein setting: word embeddings are created using the word2vec package, and the LM considers each token separately.

The first system variant (*word2phrase*) includes phrase recognition, i.e. before extracting the embeddings and creating the LM, the documents are analysed by the word2phrase module in the word2vec package. This is an implementation of the algorithm presented in (Mikolov et al., 2013), which basically identifies words that appear frequently together, and infrequently in other contexts, and treats them as single tokens (connected by an underscore).

The second system variant (*word2phrase+LemmaPos*) adds another information layer, in that each document is first lemmatized and PoS tagged using the Tint NLP Suite (Aprosio and Moretti, 2016), that works at token level; then word2phrase is run, and then the embeddings and the LM are created. In this way, we obtain so-called 'context-aware' embeddings, which is the recommended setting in (Paetzold and Specia, 2016b).

4 Evaluation

The evaluation of automated simplification is an open issue since, similar to machine translation, there may be different acceptable simplifications for a term, while a benchmark usually presents only one solution. Therefore, we perform two evaluations: the first is based on an automated comparison between Lexenstein output and the gold simplifications in the benchmark. The second is a manual evaluation aimed at scoring fluency, adequacy and simplicity of the output.

For the first evaluation, we compute the Mean Reciprocal Rank (MRR), which is usually adopted to evaluate a list of possible responses ordered by probability of correctness against a gold answer. We use this metrics because Lexenstein returns 5 possible simplifications, ranked by relevance, and with MRR it is possible to weight the response matching with the gold simplification according to its rank. In particular, MRR is computed as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

where Q is the number of simplifications to be performed (901) and $rank_i$ is the position of the correct simplification in the rank returned by Lexenstein.

We run the system in the three configurations described in Section 3.2 on each source sentence

[4] https://it.wikipedia.org/wiki/Pagina_principale

[5] http://www.opensubtitles.org/

[6] http://www.corpusitaliano.it/

[7] http://www.gazzettaufficiale.it/

in the benchmark. The single or multi-token term to be simplified is given. If it is found in the LM, the system suggests 5 ranked simplification candidates. Otherwise, no output is given.

Results show that the baseline model, i.e. the standard Lexenstein configuration replacing only single tokens with single tokens, yields $\mathrm{MRR} = 0.036$. The one using word2phrase achieves $\mathrm{MRR} = 0.042$, while the version including also lemma and PoS information yields $\mathrm{MRR} = 0.050$. A detailed evaluation is reported in Table 2: for each of the three experimental settings, we report the number of cases in which the gold simplification matches the first ranked replacement returned by Lexenstein (*1st*), the second, the third, and so on. In the last column, we report how many times (out of 901) the rank returned by Lexenstein does not contain the gold simplification present in the benchmark.

	1st	2nd	3rd	4th	5th	none
Baseline	23	12	7	3	2	854
word2phrase	30	8	8	4	1	850
+LemmaPos	32	16	11	4	4	834

Table 2: Rank of correct simplifications returned by Lexenstein

This evaluation shows that, although limited, using word2phrase in combination with lemma and PoS information yields an improvement over the baseline. However, the informativeness of this automated simplification is limited because the cases labeled as ‘none’ include both wrong simplifications and correct simplifications that are not present in the benchmark. Besides, they include also cases in which the word to be simplified was not found in the LM.

In order to better understand where the approach fails, we also perform a manual evaluation. Following the standard scheme for human evaluation of automatic text simplification (Saggion and Hirst, 2017), we judge Fluency (grammaticality), Adequacy (meaning preservation) and Simplicity of lexical simplifications using a five-point Likert scale (the higher the score, the better the output). For the setting using lemma and PoS, we do not judge Fluency, since the output is lemmatized and not converted in the original form of the source term (we plan to add this in the near future). Evaluation is performed using a set of 150 sentence pairs randomly extracted from the benchmark. We introduce also this kind of evaluation in order to have a fine-grained analysis of system output. For example, in the original sentence d) (see below), ‘tempestivamente’ was simplified with ‘periodicamente’, which is grammatically correct (high Fluency) but does not preserve the meaning of the original sentence (low Adequacy).

d) Il richiedente dovrà comunicare tempestivamente l’esattezza dei recapiti forniti.

When using word2phrase without lemmatization, the average Fluency is 3.72, Adequacy is 2.60 and Simplicity is 2.95. This shows that, while PoS and form of a simplified term are generally correct also without any processing, the preservation of the meaning is a critical issue. Simplicity achieves better scores than Adequacy, but it still needs improvements. Results obtained using lemma and PoS in combination with word2phrase are slightly better, with 2.64 Adequacy and 3.01 Simplicity. In general, the above evaluations show that using word2phrase with lemma and PoS information is a promising approach to improve the performance of lexical simplification in real settings. The performance of Lexenstein could be further improved by adding other corpora to the LM and post-process the output of the system, so as to discard inconsistent simplifications, for example when a verb is simplified through an adverb. However, some linguistic phenomena like non-local dependencies cannot be addressed using this approach, and a separate strategy to simplify them should be taken into account.

5 Conclusions

In this work, we presented a first analysis of the role of phrases in Italian lexical simplification. We also introduced the adaptation of Lexenstein, an existing lexical simplification system, so as to take phrases into account. In the future, we plan to test other approaches for the extraction of phrases, for example by applying algorithms for recognising multiword expressions. We also plan to integrate our best model for phrase simplification in ERNESTA (Barlacchi and Tonelli, 2013), a system for syntactic simplification of Italian documents. Furthermore, within the H2020 SIMPATICO project, we will integrate our phrase simplification approach in the existing services of Trento Municipality and perform a pilot study with real users.

Acknowledgments

The research leading to this paper was supported by the EU Horizon 2020 Programme via the SIMPATICO Project (H2020-EURO-6-2015, n. 692819).

References

Alessio Palmero Aprosio and Giovanni Moretti. 2016. Italy goes to Stanford: A collection of CoreNLP modules for Italian. *CoRR*, abs/1609.06204.

Gianni Barlacchi and Sara Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, pages 476–487, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and Annotation of the First Italian Corpus for Text Simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA, June. Association for Computational Linguistics.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China, July. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Gustavo Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. In *ACL-IJCNLP 2015 System Demonstrations*, ACL, pages 85–90, Beijing, China.

Gustavo Paetzold and Lucia Specia. 2016a. Benchmarking lexical simplification systems. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Gustavo H. Paetzold and Lucia Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3761–3767. AAAI Press.

H. Saggion and G. Hirst. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 347–355, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. pages 901–904.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. SIMPITIKI: a Simplification corpus for Italian. In *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it)*, volume 1749 of *CEUR Workshop Proceedings*.

Analysis of Italian Word Embeddings

Rocco Tripodi
Ca' Foscari University of Venice
rocco.tripodi@unive.it

Stefano Li Pira
University of Warwick
stefano.li-pira@wbs.ac.uk

Abstract

English. In this work we analyze the performances of two of the most used word embeddings algorithms, skip-gram and continuous bag of words on Italian language. These algorithms have many hyper-parameter that have to be carefully tuned in order to obtain accurate word representation in vectorial space. We provide an extensive analysis and an evaluation, showing what are the best configuration of parameters for specific analogy tasks.

Italiano. *In questo lavoro analizziamo le performances di due tra i più usati algoritmi di word embedding: skip-gram e continuous bag of words. Questi algoritmi hanno diversi iperparametri che devono essere impostati accuratamente per ottenere delle rappresentazioni accurate delle parole all'interno di spazi vettoriali. Presentiamo un'analisi accurata e una valutazione dei due algoritmi mostrando quali sono le configurazioni migliori di parametri su specifiche applicazioni.*

1 Introduction

The distributional hypothesis of language, set forth by Firth (1935) and Harris (1954), states that the meaning of a word can be inferred from the contexts in which it is used. Using the co-occurrence of words in a large corpus, we can observe for example that the contexts in which *client* is used are very similar to those in which *customer* occur, while less similar to those in which *waitress* or *retailer* occur. A wide range of algorithms have been developed to exploit these properties. Recently, one of the most widely used method in many natural language processing (NLP) tasks is word embeddings (Bengio et al., 2003; Mikolov et al., 2010; Mikolov et al., 2013). It is based on neural network techniques and has demonstrated to capture semantic and syntactic properties of words taking as input raw texts without other sources of information. It represents each word as a vector such that words that appear in similar contexts are represented with similar vectors (Collobert and Weston, 2008; Mikolov et al., 2013). The dimensions of the word are not easily interpretable and, with respect to explicit representation, they do not correspond to specific concepts.

In Mikolov et al. (2013), the authors propose two different models that seek to maximize, respectively, the probability of a word given its context (Continuous bag-of-word model), and the probability of the surrounding words (before and after the current word) given the current word (Skip-gram model). In this work we seek to further explore the relationships by generating word embedding for over 40 different parameterizations of the continuous bag-of-words (CBOW) and the skip-gram (SG) architectures, since as shown in Levy et al. (2015) the choice of the hyper-parameters heavily affect the construction of the embedding spaces.

Specifically our contributions include:

- **Word embedding.** The analysis of how different hyper-parameters can achieve different accuracy levels in relation recovery tasks (Mikolov et al., 2013).
- **Morpho-syntactic and semantic analysis.** Word embeddings have demonstrated to capture semantic and syntactic properties, we compare two different objectives to recover relational similarities for semantic and morph-syntactical tasks.
- **Qualitative analysis.** We investigate problematic cases.

2 Related works

The interest that word embedding models have achieved in the NLP international community has recently been confirmed by the increasing number of studies that are adopting these algorithms in languages different from English. One of the first example is the Polyglot project that produced word embedding for 117 languages (Al-Rfou et al., 2013). They demonstrated the utility of word embedding, achieving, in a part of speech tagging task, performances competitive with the state-of-the art methods in English. Attardi et al. (2014) have done the first attempt to introduce word embedding in Italian obtaining similar results. They have shown that, using word embedding, they obtained one of the best accuracy levels in a named entity recognition task.

However, these optimistic results are not confirmed by more recent studies. Indeed the performance of word embedding are not directly comparable in the accuracy test to those obtained in the English language. For example, Attardi and Simi (2014) combining the word embeddings in a dependency parser have not observed improvements over a baseline system not using such features. Berardi et al. (2015) found a 47% accuracy on the Italian versus 60% accuracy on the English. The results may be a sign of a higher complexity of Italian with respect to English as we will see section 4.1.

Similarly, recent work that trained word embeddings on tweets have highlighted some criticalities. One of these aspects is how the morphology of a word is opaque to word embeddings. Indeed, the relatedness of the meaning of a lemma's different word forms, its different string representations, is not systematically encoded. This means that in morphologically rich languages with long-tailed frequency distributions, even some word embedding representations for word forms of common lemmata may become very poor (Kim et al., 2016).

For this reason, some recent contribution on Italian tweets have tried to capture these aspects. Tamburini (2016) trained SG on a set of 200 million tweets. He proposed a PoS-tagging system integrating neural representation models and a morphological analyzer, exhibiting a very good accuracy. Similarly, Stemle (2016) proposes a system that uses word embeddings and augment the WE representations with character-level representations of the word beginnings and endings.

HP	SG	CBOW
dim	200, 300, 400, 500	200, 300, 400, 500
w	3, 5	2, 5
m	1, 5	1, 5
n	1, 5, 10	1, 5, 15

Table 1: Hyper-parameters

We have observed that in these studies the authors used either the most common set-up of parameters gathered from the literature (Tamburini, 2016; Stemle, 2016; Berardi et al., 2015) or an arbitrary number (Attardi and Simi, 2014; Attardi et al., 2016). Despite the relevance given to these parameters in the literature (Goldberg, 2017) we have not seen studies that analyze the different strategies behind the possible parametrization. In the next section, we propose a model to deepen these aspects.

3 Italian word embeddings

Previous results on the word analogy tasks have been reported using vectors obtained with proprietary corpora (Berardi et al., 2015). To make the experiments reproducible, we trained our models on a dump of the Italian Wikipedia (dated 2017.05.01), from which we used only the body text of each articles. The obtained texts have been lowercased and filtered according to the corresponding parameter of each model. The corpus consists of 994.949 sentences that result in 470.400.914 tokens.

The hyper-parameters used to construct the different embeddings for the SG and the CBOW models are: the size of the vectors (dim), the window size of the words contexts (w), the minimum number word occurrences (m) and the number of negative samples (n). The values that these hyper-parameters can take are shown in Table 1.

4 Evaluation

The obtained embedding[1] spaces are evaluated on an *word analogy* task, using a enriched version of the Google word analogy test (Mikolov et al., 2013), translated in Italian by (Berardi et al., 2015). It contains 19.791 questions and covers 19 relations types. 6 of them are semantic and 13 morphosyntactic (see Table 2). The proportions of

[1] The trained vectors with the best performances are available at http://roccotripodi.com/ita-we

Morphosyntactic	Semantic
adjective-to-adverb	capital-common-countries
opposite	capital-world
comparative	currency
superlative (assoluto)	city-in-state
present-participle (gerundio)	regione capoluogo
nationality-adjective	
past-tense	
plural	
plural-verbs (3rd person)	
plural-verbs (1st person)	
remote-past-verbs (1st person)	
noun-masculine-feminine-singular	
noun-masculine-feminine-plural	
#10.876	#8.915

Table 2: Relation types

these two types of question is balanced as shown in Table 2.

To recover these relations two different methods are used: 3CosAdd (Eq. 1) (Mikolov et al., 2013) and 3CosMul (Eq. 2) (Levy et al., 2014) to compute vectors analogies:

$$\text{3CosAdd} \ \underset{b^* \in V}{\text{argmax}} \, cos(b^*, b - a + a^*) \quad (1)$$

$$\text{3CosMul} \ \underset{b^* \in V}{\text{argmax}} \, \frac{cos(b^*, b)cos(b^*, a^*)}{cos(b^*, a) + \epsilon} \quad (2)$$

These two measures try to capture different relations between word vectors. The idea behind these measures is to use the cosine similarity to recover the vector of the hidden word (b^*) that has to be the most similar vector given two positive and one negative word. In this way, it is possible to model relations such as *queen* is to *king* what *woman* is to *man*. In this case, the word *queen* (b^*) is represented by a vector that has to be similar to *king* (b) and *woman* (a^*) and different to *man* (a). The two analogy measures slightly differ in how they weight each aspect of the similarity relation. 3CosAdd allows one sufficiently large term to dominate the expression (Levy et al., 2014), 3CosMul achieves a better balance amplifying the small differences between terms and reducing the larger ones (Levy et al., 2014). As explained in Levy et al. (2014), we expect 3CosMul to overperform 3CosAdd in evaluating both the syntactic and the semantic tasks as it tries to normalize the strength of the relationships that the hidden term has both with the attractor terms and with the repellers term.

m=1	m=5	Berardi
3.227.282	847.355	733.392

Table 3: Vocabulary length

4.1 Experimental results

The results of our evaluation are presented in Figure 1. The main trend that it is possible to notice is that accuracy increases as the number of dimensions of the embedded vectors increases. This indicates that Italian language benefits of a rich representation that can account for its rich morphology. Another important trend that emerges is the fact that the parameters have the same effect on both algorithms and that they perform very differently on all the tasks. CBOW has very low accuracy compared to SG. We can also see that the dim hyper-parameter is not correlated with the dimension of the vocabulary (model complexity) as one should expect. In fact, with increasing values of dim the accuracy increases whatever is the value of m. This hyper-parameter heavily affects the vocabulary length (see Table 3). However the dim hyper-parameter seems to be correlated only with the accuracy in the semantic tasks while the performances on the morpho-syntactic tasks seems not to have a big bust increasing the dimensionality.

With respect to the size of the context (w) used to create the words representations we do not observe a clear difference between the 18 pairs both in the SG and in the CBOW. On the contrary a clear trend can be observed varying the n hyper-parameter, with $n = 1$ the accuracy was significantly lower than the one we obtained with $n = 5$ or $n = 10$. Increasing the number of negative samples constantly increases the accuracy.

These results support also the claim put forward by (Levy et al., 2014) that the 3CosMul method is more suited to recover analogy relations. In fact, we can see that on average the right bars of the plots are higher than the left.

4.2 Error analysis

If we restrict the error analysis to the most macroscopic differences in figure 1 we can compare three different parametrizations: SG-200 w5-m5-n1, SG-500 w5-m5-n1, SG-500 w5-m5-n10. In this way we can analyze the results obtained changing the number of dimensions of the vectors and the role played by n. In Table 4 the total num-

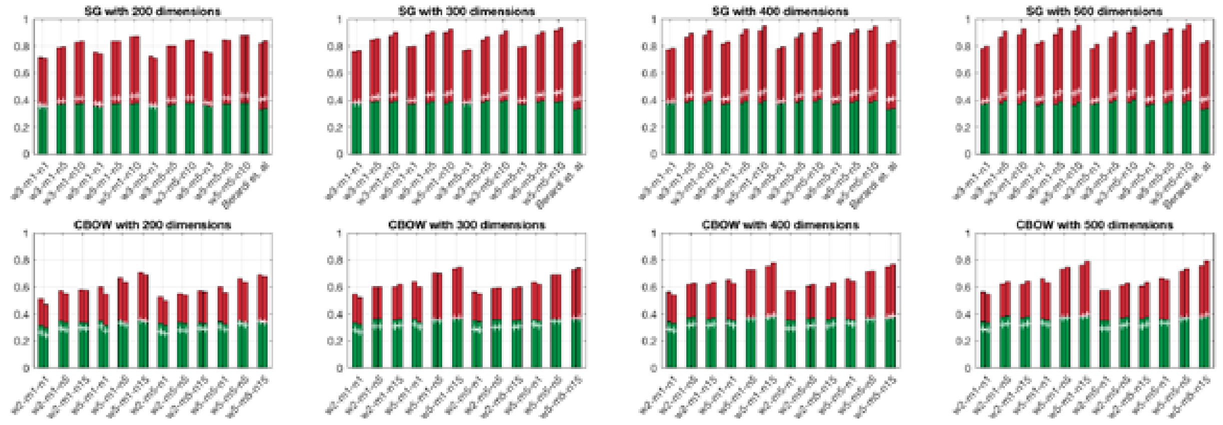


Figure 1: Results as accuracy with different hyper-parameters (y axis) using the 3CosAdd (left bar) and the 3CosMul (right bar) formula. The green part of the bars indicates the accuracy on the morpho-syntactic task whereas the red one the accuracy on the semantic task. The + sign on each bar indicates the accuracy on the entire dataset. The upper row of the figure shows the results of the SG algorithm and the bottom row the results of CBOW. The last two bars of the SG plots indicates the results obtained using the vectors made available by (Berardi et al., 2015)

Parametrization	#errors	#words
SG-200-w5-m5-n10	10.113	543
SG-500 w5-m5-n1	10.506	535
SG-500 w5-m5-n10	9.337	525

Table 4: Total number of errors and number of different words that have not been recovered

ber of errors and the number of different words that have not been recovered by each parametrization are presented. From this table we can see that most of the errors are done one a relatively small set of words. This phenomenon can be studied analyzing the most problematic cases. In Table 5 we can see the list of the most common errors ranked by frequency for each method. As we can see from these lists the errors are done on the same words and this because they are the most common in the dataset (e.g.: in the dataset there are 217 queries that require *Florida* as answer compared to the 55 of *Italia*). However if we compare the frequency of these errors in the analogy test within the three parametrisation we can observe an improvement of approximately 15% in accuracy with SG-500 w5-m5-n10. Indeed, despite many errors are not recovered for any of the parametrisation, we can observe that approximately 21% of the errors are recovered under certain parametrizations (Table 6). To further investigate these improvements related to the aforementioned parametrisation we focused on one of the most frequent errors in the analogy test, the word *California*. As we can see from the list of the analogy test solved (Table 7) different parametrizations are helpful to solve different types of analogies. For example an increase in the dimensionality increases the accuracy, but mainly in analogy test with words that have a representation in the training data related to a wider set of contexts (*Houston*:*Texas*; *Chicago*:*Illinois*). The best parametrisation is obtained increasing the negative sampling. As we can see from the examples provided, the analogies are resolved thanks to a contextual similarity between the two pairs (*Huntsville*:*Alabama*; *Oakland*:*California*). In these cases the negative sampling could help to filter out from each representation those words that are not expected to be relevant for the words embeddings.

SG-200-w5-m5-n10	#	SG-500 w5-m5-n1	#	SG-500 w5-m5-n10	#
california	328	california	349	california	287
texas	223	texas	224	texas	165
arizona	164	arizona	164	arizona	145
florida	144	ohio	142	florida	124
ohio	135	florida	140	ohio	112

Table 5: Most common errors

Similar types of improvement are noticed on analogy tests that contain a challenging word *predire* (*predict*). The results of this analysis are presented in Table 9 where it is possible to see that an higher dimensionality improves the accuracy of analogical tests containing open domain verbs (e.g.: *descrivere*, *vedere*). Similarly to the previous case, an higher dimensionality allows for fine

Parametrization	#errors solved
dim = 500 & n = 10	873
solo dim = 500	645
solo n = 10	927

Table 6: Solved errors

dim = 500 & n = 10	*solo* n = 10	*solo* dim = 500
Milwaukee Wisconsin Oakland California	Huntsville Alabama Oakland California	Houston Texas Oakland California
Shreveport Louisiana Oakland California	Baltimore Maryland Oakland California	Chicago Illinois Oakland California
Irvine California Shreveport Louisiana	Irvine California Phoenix Arizona	Denver Colorado Oakland California
Irvine California Baltimore Maryland	Arlington Texas Irvine California	Philadelphia Pennsylvania Oakland Calif
Sacramento California Henderson Nevada	Phoenix Arizona Sacramento California	Portland Oregon Oakland California
Sacramento California Orlando Florida	Huntsville Alabama Sacramento California	Tulsa Oklahoma Irvine California

Table 7: Examples of analogy tests solved.

grained partitions improving the correct associations between terms. However, also in in this case, the best parametrizations are obtained increasing the negative sampling or both the parameters. As we can see here both the present participle and the past tense pairs are correctly solved. These example provide a preliminary evidence of how negative sampling, filtering out non informative words from the relevant context of each word, is able to build representation by opposition that are beneficial both for semantic and syntactic associations.

Examples of words that almost always are not recovered correctly are presented in Table 10. A selected list of words problematic for all parametrizations is shown in Table 8. It contains plurals, feminine, currencies, superlatives and ambiguous words. The low performances on these cases can be explained by the poor coverage of these categories in the training data. In particular, it would be interesting to study the case of feminine and to analyze if it is due to a gender bias in the Italian Wikipedia, as a preliminary analysis of the most frequent errors that persist in all the parametrization seems to suggest. The words that have been benefited by the increase of n are:

ghana
pakistan
irlandese
migliorano
scrivendo
slovenia
giocando
serbia
implementano
ucraino
zimbabwe
namibia
suonano
maltese
portoghese
contessa
messicano
giordania

the errors that have been introduced increasing this parameter are related to the words in Table 11. It is interesting to notice that given an error in an analogy test, it is possible to find the correct answer in the top five most similar words to the query. Precisely we observed this phenomenon in 26% of the cases for SG-200-w5-m5-n10, in 27% of the cases for SG-500-w5-m5-n1 and in 25% for SG-500-w5-m5-n1. Furthermore, approximately in 50% of these cases the correct answer is the second most similar. Most of the recovery errors are due to vocabulary issues. In fact, many words of the test set have no correspondence in the developed embedding spaces. This is due to the low frequency of many words that are not in the training corpus or that have been removed from the vocabulary because of their (low) frequency. For this reason we kept the m hyper-parameter very low (e.g., 1 and 5), in counter-tendency with recent works that use larger corpora and then remove infrequent words setting m with high values (e.g., 50 or 100). In fact, with increasing value of m the number of not given answers increases rapidly. It passes from 300 ($m = 1$) to 893 ($m = 5$).

Some of the words that are not present in the vocabulary with $m = 1$ include plural verbs (1st person), that probably are not used by a typical Wikipedia editor and remote past verbs (1st person), a tense that in recent years is disappearing from written and spoken Italian. Some of these verbs are:

giochiamo
affiliamo
rallentiamo
zappiamo
implementai
rallentai
mescolai
nuotai

In Berardi et al. (2015) the number of not given answer is 1.220. The accuracy of their embeddings, obtained using a larger corpus and using the hyper-parameters that perform well on English language, is always lower than those obtained with our setting, in both the morphosyntactic and the semantic tasks. This confirms that the regularization of the parameters is crucial for good representation of the embeddings, since the Berardi et al. (2015)'s model has been trained on a much larger corpus and for this should outperform ours. Furthermore, this model seems to have some tokenization problem.

5 Conclusions

We have tested two word representation methods: SG and CBOW training them only on a dump of the Italian Wikipedia. We compared the results of the two models using 12 combinations of hyper-parameters.

We have adopted a simple word analogy test to evaluate the generated word embeddings. The results have shown that increasing the number of

pilotesse	migliore	colori	meloni
pere	matrigna	figliastra	sua
real	lev	yen	mamma
kwanza	vantaggiosissimo	urlano	stimano
aquila	eroina	programmato	impossibilmente

Table 8: Always wrong

dim = 500 & n = 10	***solo* n = 10**	***solo* dim = 500**
dire detto predire predetto	cantare cantato predire predetto	descrivere descritto predire predetto
mescolare mescolando predire predicendo	correre correndo predire predicendo	vedere visto predire predetto
predire predicendo generare generando	generare generando predire predicendo	
rallentare rallentando predire predicendo	predire predicendo programmare programmando	
scoprire scoprendo predire predicendo	scrivere scrivendo predire predicendo	

Table 9: Examples of analogy tests solved.

SG-200-w5-m5-n10	#	SG-500 w5-m5-n1	#	SG-500 w5-m5-n10	#
capre	26	groenlandia	27	ratti	26
rapidamente	26	silenziosamente	27	ovviamente	25
dolcissimo	26	caldissimo	27	incredibilmente	25
apparentemente	26	occhi	27	grandissimo	25
andato	26	greco	27	malvolentieri	25

Table 10: Almost always wrong

irlanda	afghanistan	albania	egiziano
olandese	provvedono	francese	svizzero

Table 11: New errors

dimensions and the number of negative examples improve the performance of both the models.

These types of improvement seems to be beneficial only for the semantic relationships. On the contrary the syntactical relationship are negatively affected by the low frequency of many of its terms. This should be related to the morphological complexity of Italian. In the future it would be helpful to represent the spatial relationship regarding specific syntactical domain in order to evaluate the contribution of hyper-parametrization to syntactical relationship accuracy. Moreover future work will include the testing of these word embedding parametrizations in practical applications (e.g. analysis of patents'descriptions and books' corpora).

Acknowledgments

Part of this work has been conducted during a collaboration of the first author with DocFlow Italia. All the experiments in this paper have been conducted on the SCSCF multiprocessor cluster system at University Ca' Foscari of Venice.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *CoNLL-2013*, page 183.

Giuseppe Attardi and Maria Simi. 2014. Dependency parsing techniques for information extraction.

Giuseppe Attardi, Vittoria Cozza, and Daniele Sartiano. 2014. Adapting linguistic tools for the analysis of italian medical records.

Giuseppe Attardi, Daniele Sartiano, Chiara Alzetta, Federica Semplici, and Largo B Pontecorvo. 2016. Convolutional neural networks for sentiment analysis on italian tweets. In *CLiC-it/EVALITA*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

John Rupert Firth. 1935. The technique of semantics. *Transactions of the philological society*, 34(1):36–73.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Zellig S Harris. 1954. Distributional structure. word, 10 (2-3): 146–162. reprinted in fodor, j. a and katz, jj (eds.), readings in the philosophy of language.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *AAAI*, pages 2741–2749.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Egon W Stemle. 2016. bot. zen@ evalita 2016-a minimally-deep learning pos-tagger (trained for italian tweets). In *CLiC-it/EVALITA*.

Fabio Tamburini. 2016. A bilstm-crf pos-tagger for italian tweets using morphological information. In *CLiC-it/EVALITA*.

Applicazione di un metodo attribuzionistico quantitativo alla monodia liturgica medievale

Francesco Unguendoli
Dipartimento di Scienze Fisiche,
Informatiche e Matematiche
Università di Modena
e Reggio Emilia
francesco.unguendoli@
unimore.it

Giampaolo Cristadoro
Dipartimento di Matematica
Università di Bologna
giampaolo.cristadoro@
unibo.it

Marco Beghelli
Dipartimento delle Arti
visive, performative, medievali
Università di Bologna
marco.beghelli@
unibo.it

Abstract

Italiano. L'articolo mostra come tecniche di analisi stilometriche comunemente usate in ambito letterario (basate sulla distanza tra vettori delle frequenze di n-grammi di lettere) possano essere adattate con successo allo studio di repertori musicali "unidimensionali" (ovvero melodie prive di ritmo e di accompagnamento). I buoni risultati ottenuti su un corpus di monodie liturgie di origine medievale (Canto Gregoriano e Canto Romano Antico) sono un primo passo verso l'adozione e la creazione di tecniche automatiche a supporto di studi stilometrici a carattere e interesse strettamente musicologico.

English. *We adapt a technique commonly used in the stylometric attribution of literary texts (based on a pseudo-distance between frequency-vectors of n-grams of letters) to the analysis of "unidimensional" musical repertoires (rhythm-free melody without accompaniment). We successfully apply the method to a corpus of liturgical monodies of medieval origin (the so-called Gregorian Chant, in comparison with the Old Roman Chant). Our results give a first indication that automatic stylometric techniques can be fruitfully adopted to support the study of refined problems in musicology.*

1 Motivazioni della ricerca

Il problema dell'attribuzione in arte, vale a dire l'identificazione dell'autore di un'opera dell'ingegno adespota, è comunemente noto per le arti visive e letterarie (attribuzione di quadri e testi non firmati). Come problema filologico non è meno sentito fra gli storici della musica, spesso alle prese con composizioni più o meno antiche d'incerta paternità. Mancano tuttavia al musicologo utili strumenti analitici che consentano di andar oltre la semplice impressione soggettiva d'ascolto, mentre le metodologie d'indagine stilistica fino ad oggi applicate alla musica hanno perlopiù lavorato a livello di macro-generi compositivi. Prima di affrontare veri problemi di attribuzione in ambito musicale è dunque necessario individuare metodologie analitiche adeguate.

L'applicazione a repertori musicali semplici di metodi d'indagine stilistica computazionale (stilometria) già verificati su testi verbali offre ora i primi buoni risultati, da testare poi su composizioni più complesse, con le dovute modifiche. Lo scopo ultimo non è la costruzione di algoritmi efficienti per l'attribuzione di testi musicali, confrontando l'efficacia assoluta dei diversi metodi, né di sostituire la macchina all'orecchio e al discernimento del musicologo, ma piuttosto offrire a questo uno strumento d'indagine filologica in più che faccia emergere ulteriori tratti distintivi (features) delle varie musiche, dei vari autori, permettendogli così di valutare aspetti stilistici che da solo non percepirebbe.

2 Gli n-grammi in ambito letterario

Sin dall'avvento dei primi computer si è tentato di processare caratteristiche stilometriche per affrontare problemi attribuzionistici. Inizialmente gli indicatori quantitativi utilizzati erano perlopiù legati a caratteristiche lessicali o sintattico-semantiche dei testi analizzati; in Kešelj et al. (2003) gli autori si rivolsero a indicatori di livello più basso, individuando come features stilistici i cosiddetti n-grammi, ossia sequenze di n simboli (lettere, spazi, interpunzioni) consecutivi.

Tale metodo è stato raffinato da Basile et al. (2008) per adattarlo a uno specifico problema: attribuire a Gramsci oppure a suoi collaboratori una serie di articoli giornalistici pubblicati adespoti (un problema difficile in quanto testi brevi ed estremamente simili per tematiche e linguaggio (Lana 2010)). Gli n-grammi sono stati dunque utilizzati per costruire distanze non più fra il singolo testo adespoto e il profilo medio di un singolo autore, come fatto da Kešelj, ma rispetto ad ogni testo disponibile, prendendo inoltre in considerazione tutti gli n-grammi (e non solo i più frequenti) per contrastare la brevità: al testo adespoto veniva così assegnato un "voto" rispetto a tutti i testi del corpus di riferimento, basato sulla sua posizione in una classifica costruita sulle distanze, e da tali voti veniva ricavato un indice riassuntivo sull'appartenenza all'uno o all'altro gruppo, insieme a una stima sulla validità di tale attribuzione.

3 Verifiche sui testi musicali

In campo musicale è opportuno notare che ad oggi gran parte della ricerca è stata finalizzata alle tecniche per la gestione, l'organizzazione e l'accesso ai grandi database musicali, principalmente quelli della rete, piuttosto che a una fine comparazione di testi nell'ambito della cosiddetta "musica d'arte", cui il musicologo è maggiormente interessato. Il punto di vista e le tecniche coinvolte sono ovviamente differenti, là dove alla richiesta estetica di distinguere con precisione gli autori di musiche estremamente simili fra loro si contrappone nel Music Information Retrieval la necessità di automatizzare e velocizzare procedure che trattano grandi quantità di dati, rinunciando a discriminare fra brani di uno stesso genere o di autori stilisticamente vicini.

I metodi di attribuzione basati sugli n-grammi sono stati già testati più volte, ad esempio da Doraisamy e Ruger (2003) e da Hillewaere et al. (2010), oltre che dallo stesso Kešelj et al. (2008, 2013), sia nel campo già citato della ricerca e categorizzazione in grandi database, sia in problemi attribuzionistici più prettamente musicologici. Passando dalla linearità del linguaggio letterario alla multi-dimensionalità di quello musicale, i problemi maggiori sono, per metodi basati sugli n-grammi, la definizione stessa di unigramma e il trattamento delle "voci" parallele, e per metodi più generali la difficoltà di trovare un insieme di style-markers effettivamente rappresentativo.

Backer e Van Kranenburg (2005) sono tra i primi ad affrontare problematiche di attribuzione, utilizzando un corpus di brani di Bach, Händel, Telemann, Haydn e Mozart e venti style-markers differenti, utilizzati anche singolarmente o a sottogruppi: i risultati sono molto buoni nella maggior parte delle prove effettuate, con un'accuratezza sopra il 90%, tranne che nel confronto tra Mozart e Haydn, stilisticamente assai più impegnativo, in cui l'accuratezza nelle attribuzioni scende a circa il 75%. Metodologie simili vengono usate più di recente anche da Brinkman et al. (2016) per affrontare le problematiche autoriali relative all'opera di Josquin, messo a confronto con Ockeghem, Dufay, De Orto e La Rue; i risultati tuttavia confermano la difficoltà del problema in quanto solo il 60% circa dei pezzi di Josquin vengono attribuiti correttamente, mentre parecchi vengono confusi con quelli di La Rue.

Wołkowicz et al. (2008), e Hillewaere et al. (2010) hanno comparato musiche pianistiche di Bach, Mozart, Beethoven, Schubert e Chopin, e confrontato in particolare i quartetti per archi di Mozart e di Haydn; in quest'ultimo caso, non facile anche per il musicologo, i risultati dei vari metodi, siano essi basati sugli n-grammi o sul riconoscimento di patterns, hanno fornito valori di accuratezza simili, con percentuali massime intorno al 70-75%. Globalmente si può notare che se le varie metodologie hanno dato ottimi risultati per un utente medio nella ricerca e gestione globale, raramente possono raggiungere un livello di affidabilità sufficiente per i sottili problemi attribuzionistici della musicologia storico-estetica, con spiccate velleità filologiche. È dunque nella speranza di poter offrire un giorno risposte a questi ultimi che abbiamo fatto in un certo senso un passo indietro, testando in ambito musicale un metodo già noto in ambito letterario: quello di Kešelj et al. (2003), modificato da Basile et al. (2008)[1].

Per cominciare l'indagine si sono scelti repertori monodici e non mensurali, caratterizzati cioè da una sola e semplice successione di note ad altezze diverse (stringhe di suoni), evitando così tutta una serie di ulteriori parametri che costituiscono la maggiore difficoltà d'indagine per la musica d'arte occidentale (durate, ritmi, dinamiche, agogiche,

[1] Utilizzare il metodo nella sua forma originale col solo scopo di comparare la sua maggiore o minor efficacia sul linguaggio musicale rispetto ad altri metodi non rientra fra i compiti circoscritti di questa ricerca.

intrecci contrappuntistici, agglomerati armonici, ecc.). Un esempio:

In prospettiva, l'intenzione è di estendere il metodo d'indagine - opportunamente adattato - a repertori musicali più complessi (polifonici, armonici, ecc.), nei quali i problemi di attribuzionismo tuttora irrisolti rivestono ben maggior interesse per la musicologia, sul piano storico come su quello filologico.

4 Ambito d'indagine e obiettivo

La presente applicazione alla musica del metodo computazionale fondato sul concetto di n-grammi è cominciata con il confronto di due repertori liturgici d'origine medievale: il cosiddetto Canto Gregoriano (sviluppatosi in area francese per diffondersi poi in tutta l'Europa cristiana) e il meno noto Canto Romano Antico (rimasto limitato alle chiese romane non pontificali).

In tali repertori, alla semplicità lineare della musica si contrappone, ai fini computazionali, la difficoltà prospettata da lunghezze assai limitate se confrontate a quelle dei comuni testi letterari (solo poche centinaia di note musicali per ogni brano) e dalla difficoltà di enucleare efficacemente in quelle melodie elementi sintattici analoghi a parole, frasi e periodi. Quanto poi alla natura stilistica di tale musica, va segnalata la notevole somiglianza melodica non solo fra un testo e l'altro del medesimo corpus, ma anche fra i due repertori in esame: una conseguenza della loro genesi, frutto di una autorialità collettiva estesa su un abito temporale e geografico assai vasto, nonché di ripetute contaminazioni.

Date queste premesse che hanno reso la ricerca ancor più stimolante, l'obiettivo era di attribuire brani dell'uno o dell'altro repertorio al corpus di appartenenza con metodi computazionali, là dove l'orecchio anche esperto non si dimostra sempre in grado di distinguerli con certezza[2].

5 Percorsi e metodi

I 280 brani musicali utilizzati sono di varia natura liturgica e formale: per ciascuno dei due repertori sono stati presi in considerazione 60 Offertori e 50 Graduali, più ulteriori 30 brani con varie e diversificate funzioni liturgiche, destinati ad un test più impegnativo di cui si dirà. Le fonti: per il Gregoriano, le edizioni critiche del *Graduale Triplex* (1979) e dell'*Offertoriale Triplex* (1985) prodotte dal centro di Solesmes; per il Romano, l'edizione diplomatica del *Graduale Vat. lat. 5319* edito nei *Monumenta Monodica Medii Aevi* (1970).

Senza addentrarci in problematiche filologiche, la scelta di tali edizioni è stata dettata dalla loro ampiezza, che ha permesso di avere facilmente a disposizione un vasto assortimento di brani musicali su cui lavorare, offerti in trascrizioni moderne riconosciute come attendibili (a parte una manciata di evidenti refusi che sono stati tacitamente corretti). Si sono ignorati i testi verbali intonati dai singoli brani, l'interesse dell'indagine essendo rivolto esclusivamente alla dimensione musicale. Si è evitata ogni possibile interpretazione ritmica delle melodie, assegnando a ogni nota lo stesso valore di durata standardizzato. Nel gioco dei ritornelli fra le varie antifone si è provveduto a una normalizzazione formale, per evitare eccessive e ingiustificate difformità di lunghezza fra i vari brani.

Quattro le prove effettuate, a difficoltà crescente. Nelle prime due ogni brano dei due insiemi di riferimento A e B è stato trattato come testo incognito e attribuito all'uno o all'altro insieme. Nella prima prova gli insiemi di riferimento erano rappresentati dai soli Offertori (Gregoriani per l'insieme A, Romani per il B); nella seconda ognuno dei due insiemi A e B è stato esteso a comprendere anche i Graduali (Gregoriani e Romani rispettivamente), rendendolo così più vasto e meno omogeneo. Rispetto agli stessi gruppi A e B della seconda prova, nella terza prova si è poi valutata l'attribuzione dei 60 brani di differente indirizzo liturgico.

Siamo partiti dagli Offertori per tre ragioni significative: 1) il loro numero elevato a disposizione, sia nel Gregoriano sia nel Romano; 2) una apprezzabile lunghezza dei singoli brani, tra i più estesi in entrambi i repertori; 3) la quasi totale corrispondenza fra i due repertori dei testi verbali intonati, cosa che sposta tutto il peso delle differenze sulla sola componente melodica. Era così possibile avviare un primo lavoro di confronto su un gruppo di brani omogeneo, senza introdurre potenziali variabili dettate dalle diverse funzioni liturgiche. Con motivazioni simili è stato poi aggiunto agli

[2] Una analisi quantitativa del problema, tramite prove di riconoscimento auditivo, è in corso. I primi risultati stanno confermando tale difficoltà.

Offertori il gruppo dei Graduali, più brevi e con caratteristiche musicali differenti.

Ragioni opposte regolano invece il terzo gruppo di musiche, destinato a testare il metodo attributivo con brani attinenti a differenti funzioni liturgiche (di volta in volta: Introitus, Alleluia, Tractus, Sequentia, Offertorium, Communio, Antiphona, Inno, Canticum). Ne consegue una minore omogeneità melodica e una maggiore varietà di lunghezze (i brani sono tendenzialmente più brevi), difficoltà cui si aggiunge in alcuni casi, specie fra gli Alleluia, la presenza di stesse melodie o di loro parti fra i due repertori, cosa che rende ovviamente molto più complessa una precisa attribuzione all'una o all'altra famiglia.

Nella quarta prova, divisa in due parti, gli insiemi A e B erano formati rispettivamente da Offertori e Graduali dello stesso repertorio (Gregoriano o Romano); si è inteso così valutare se l'analisi quantitativa sia in grado di confermare le differenze stilistiche osservate dai musicologi fra Graduali e Offertori, sia all'interno del Gregoriano, sia del Romano: entrambi i generi liturgici sono infatti ben caratterizzati sul piano stilistico, al punto da formare sottogruppi musicali omogenei all'interno dei due repertori.

6 Risultati

Sull'esempio di Basile et al. (2008), non ci si è avvalsi di un profilo medio dei gruppi di raffronto: per ogni brano si sono calcolate le distanze da tutti gli altri brani di riferimento; quindi, tramite una procedura di voto, si è ottenuto un indice riassuntivo i cui valori, tra $[-1, 1]$, indicassero -oltre alla attribuzione all'uno o all'altro repertorio- anche una stima della validità di tale attribuzione.

Come unigramma di base è stata scelta la differenza di altezza fra due note consecutive (e non fra ogni nota e la finalis del brano, per evitare la dipendenza dal modo gregoriano di appartenenza). Inoltre, causa la brevità dei brani, si è scelto di valutare tutti gli n-grammi disponibili (e non solo i più frequenti). Il parametro fondamentale n della lunghezza degli n-grammi è stato testato in un range di valori compresi tra $n = 2$ e $n = 10$ (corrispondenti quindi a frammenti melodici da 3 a 11 note di lunghezza).

Utilizzando le seguenti notazioni: ω per il generico n-gramma, $D_n(x)$ per il dizionario degli n-grammi del testo x, $f_x(\omega)$ per la frequenza relativa dell' n-gramma ω nel testo x; la distanza $d_n(x, y)$ tra i testi x e y, calcolata per un valore fissato n della lunghezza degli n-grammi, è definita come (Basile et al. , 2008):

$$d_n(x, y) = C \sum_{\omega \in D_n(x) \cup D_n(y)} \left(\frac{f_x(\omega) - f_y(\omega)}{f_x(\omega) + f_y(\omega)} \right)^2 \tag{1}$$

con $C = \frac{1}{|D_n(x)|+|D_n(y)|}$.

L'assegnazione dei singoli brani all'uno o all'altro repertorio è stata quindi effettuata tramite una procedura di "voto" che utilizza tutte le distanze intertestuali. Le distanze del testo incognito x da tutti i testi di riferimento dei due gruppi sono ordinate in maniera crescente. Per il j-esimo testo del gruppo A nella lista è stato calcolato $(k(j)/j) - 1$ dove $k(j)$ è la sua posizione nella lista. Sommando infine tali valori per tutti i testi del gruppo A si ottiene un indice di appartenenza $g(x)$; similmente è stato costruito l'indice $ng(x)$ attraverso un'analoga somma sui testi del gruppo B. L'indice $g(x)$ sarà dunque tanto più piccolo quanto più i testi del gruppo A si troveranno in alto nella classifica, ossia quanto più le loro distanze dal testo incognito saranno piccole, e lo stesso varrà per $ng(x)$ relativamente al gruppo B.

L'uso degli indici, sintetizzato in un unico valore

$$v(x) = \frac{ng(x) - g(x)}{ng(x) + g(x)}, \tag{2}$$

permette di offrire anche una stima naturale dell'affidabilità dell'attribuzione: il valore $v(x) \in [-1, 1]$ indicherà infatti testi fortemente Gregoriani per valori vicini a 1 e fortemente Romani per valori prossimi a -1, mentre per valori prossimi a 0 indicherà una valutazione più incerta.

Come riportato nei grafici sottostanti, il metodo di attribuzione utilizzato ha fornito in tutti i casi esaminati ottimi risultati.

Nelle prove più semplici (i primi due test) la percentuale di riconoscimento ha superato il 90% per ogni valore della lunghezza degli n-grammi, con punte del 100% per $n = 3$ e una tendenza ad accuratezze inferiori nelle lunghezze medie o elevate, da $n = 5$ in su (Fig. 1).

Nella terza prova, resa difficile - come detto - dalla brevità dei brani e dalle frequenti commistioni melodiche fra i due repertori, la percentuale si è comunque mantenuta buona per le minori lunghezze degli n-grammi ($n = 2$ e $n = 3$) e per quelle maggiori, con un sensibile calo di accuratezza per le lunghezze intermedie (Fig. 2)

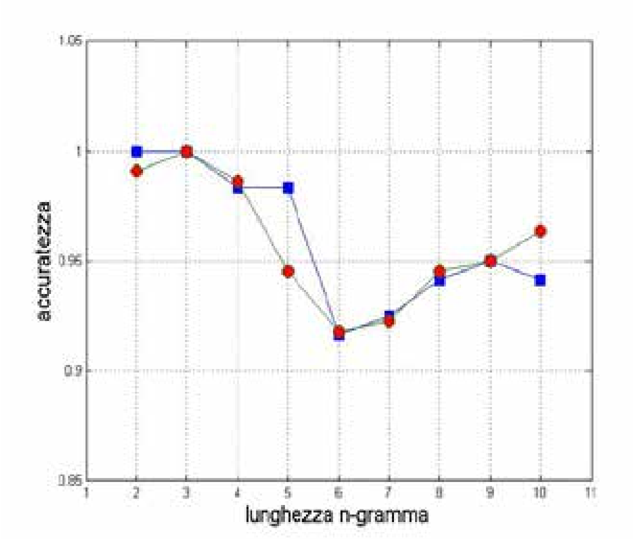


Figura 1: Accuratezza per la I prova (quadrati) e per la II prova (cerchi).

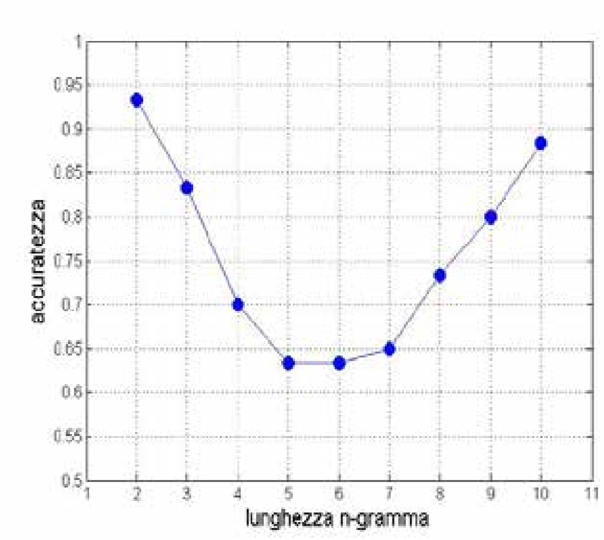


Figura 2: Accuratezza per la III prova.

È interessante notare che i risultati migliori sono ottenuti per n piccolo, in accordo con il fatto che tali repertori sono fortemente caratterizzati da cellule melodiche piuttosto brevi. Nei principali casi letterari studiati, invece, le lunghezze che fornivano i risultati migliori erano intorno a $n = 7, 8$, giustificabili con una sorta di "lunghezza media" di un'unità che, pur non essendo "semantica" in senso stretto, non è troppo lontana dalla lunghezza media delle parole. Infine anche la quarta prova, mirata a distinguere tra Offertori e Graduali, in cui l'indice finale segnala l'appartenenza all'uno o all'altro ambito liturgico, ha dato risultati superiori all'85% (tranne il caso $n = 2$ per il repertorio Romano), con un miglioramento per lunghezze medie ed elevate degli n-grammi, per le quali l'accuratezza supera il 90% (Fig. 3).

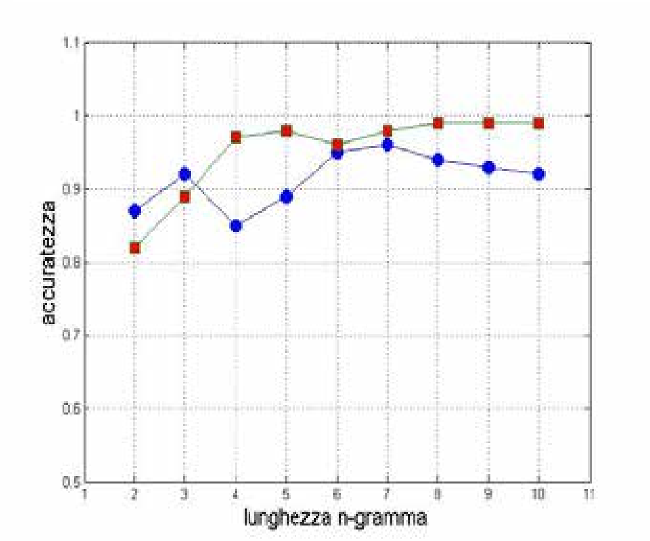


Figura 3: Accuratezza per la IV prova: repertorio Romano (quadrati) e Gregoriano (cerchi).

Il risultato conferma pertanto che le brevi cellule melodiche caratterizzanti rispettivamente il Gregoriano e il Romano sono comuni a Offertori e Graduali, mentre la distinzione fra i due generi liturgici può avvenire solo sulla base di "frasi" di maggior ampiezza. Riteniamo dunque interessante constatare che il metodo quantitativo qui utilizzato possa addentrarsi nelle caratteristiche di questi repertori sufficientemente a fondo da cogliere differenze che solo studi filologico-musicali approfonditi riescono a evidenziare.

7 Conclusioni e prospettive

I risultati presentati, con percentuali di riconoscimento esatto intorno e oltre al 90%, fanno sperare che il metodo quantitativo degli n-grammi possa validamente applicarsi anche a composizioni musicali più complesse, caratterizzate da un maggior numero di "parti" sovrapposte, e che potenzialmente sia pure in grado di contribuire a risolvere problemi di attribuzione ancora aperti fra i musicologi: distinguere ad esempio i diversi autori in partiture frutto di collaborazioni, o valutare il grado di attendibilità dell'attribuzione di una composizione d'incerta paternità. Sarebbe poi interessante appurare quale risultato si ottiene se gli n-grammi calcolati vengono processati come features da un classificatore supervisionato (ad es. una SVM con 10-fold cross-validation), o quali features vengono selezionate da altri algoritmi. Recentemente l'utilizzo di character-level embeddings e convolutional neural networks ha pure mostrato buone potenzialità in problemi di attribuzione (Kim et al. (2016), Ruder et al. (2016)): resta da verificare l'efficacia sui testi musicali. Per mantenere alte le percentuali di riconoscimento si dovrà comunque semplificare sempre il più possibile la musica, mantenendo nell'analisi quantitativa solo quelle componenti che possano risultare effettivamente discriminanti per il problema esaminato, valutate di volta in volta. La stretta collaborazione fra il matematico-informatico e il filologo musicale è dunque indispensabile ad ogni passaggio.

References

Chiara Basile, Dario Benedetto, Emanuele Caglioti, Mirko Degli Esposti. 2003. An example of mathematical authorship attribution. *Journal of Mathematical Physics*, 49(12):125211.

Eric Backer, Peter van Kranenburg. 2005. On musical stylometry - A pattern recognition approach. *Pattern Recognition Letters*, 26:299–309.

Andrew Brinkman, Daniel Shanahan, Craig Sapp. 2016. Musical stylometry, machine learning, and attribution studies: A semi-supervised approach to the works of Josquin. In Vokalek G. (ed.), *Proceedings of the 14th Biennial International Conference on Music Perception and Cognition*, ICMPC14, 593-598.

Jan Buys. 2011. Generative models of music for style imitation and composer recognition. *Honours Project in Computer Science, University of Stellenbosch.*

Shyamala Doraisamy, Stefan Rüger. 2005. Robust polyphonic music retrieval with N-grams. *Journal of Intelligent Information Systems*, 21(1):53-70.

Graduale 1979. *Graduale Triplex seu Graduale Romanum Pauli PP. VI cura recognitum & rhythmicis signis a Solesmensibus monachis ornatum neumis laudunensibus (cod. 239) et Sangallensibus (codicum Sangallensis 359 et Einsidlensis 121) nunc auctum.* Solesmis, Abbaye Saint-Pierre de Solesmes.

Ruben Hillewaere, Bernard Manderick, Darrell Conklin. 2010. String quartet classification with monophonic models. In Downie J. S., Veltkamp R. C. (eds.), *Proceedings of the 11th International Society for Music Information Retrieval Conference*, ISMIR, 537-542.

Vlado Kešelj, Fuchun Peng, Nick Cercone, Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. in PACLING '03, *Proceedings of the Pacific Association for Computational Linguistics Conference, Halifax, Computer Science Dept. at Dalhousie University*, 255-264.

Yoon Kim, Yacine Jernite, David Sontag, Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2741-2749.

Maurizio Lana. 2010. Come scriveva Gramsci? Metodi matematici per riconoscere scritti gramsciani anonimi. *Informatica Umanistica*, 3:31-56.

Bruno Stablein. 1970. Gesänge des altrömischen Graduale, Vat. lat. 5319. *Monumenta Monodica Medii Aevi*,2, Kassel, Bärenreiter.

Offertoriale 1985. *Offertoriale triplex cum versiculis.* Solesmis, Abbaye Saint-Pierre de Solesmes.

Sebastian Ruder, Parsa Ghaffari, John G. Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. preprint *arXiv:1609.06686.*

Gissel Velarde, Tillman Weyde, Carlos C. Chacon, David Meredith, Maarten Grachten. 2016. Composer recognition based on 2d-filtered piano-rolls. In *Proceedings of the 17th International Society for Music Information Retrieval Conference* ISMIR, 115-121.

Jacek Wołkowicz, Vlado Kešelj. 2013. Evaluation of n-gram-based classification approaches on classical music corpora. In Yust J., Wild J., Burgoyne J.A. (eds.) *Mathematics and Computation in Music - MCM 2013 Proceedings - Lecture Notes in Computer Science*, 7937:213-225.

Jacek Wołkowicz, Zbigniew Kulka, Vlado Kešelj. 2008. N-gram based approach to composer recognition. *Archives of Acoustics*, 33(1):43-55.

Commercial Applications through Community Question Answering Technology

Antonio Uva[‡], Valerio Storch[†], Casimiro Carrino[†],
Ugo Di Iorio[†], Alessandro Moschitti[‡◇]
[‡]Department of Computer Science and Information Engineering, University of Trento, Italy
[◇]Qatar Computing Research Institute, HBKU, Qatar
[†]RGI Group, Italy
antonio.uva@unitn.it, amoschitti@gmail.com
{valerio.storch,casimiro.carrino,ugo.diiorio}@rgigroup.com

Abstract

English. In this paper, we describe our experience on using current methods developed for Community Question Answering (cQA) for a commercial application focused on an Italian help desk. Our approach is based on (i) a search engine to retrieve previously answered question candidates and (ii) kernel methods applied to advanced linguistic structures to rerank the most promising candidates. We show that methods developed for cQA work well also when applied to data generated in customer service scenarios, where the user seeks for explanation about products and a database of previously answered questions is available. The experiments with our system demonstrate its suitability for an industrial scenario.

Italiano. *In questo articolo, descriviamo la nostra esperienza nell'usare i metodi attualmente disponibili per il Community Question Answering (cQA) in un'applicazione commerciale riguardante il servizio clienti in lingua italiana. Il nostro approccio si basa su (i) un motore di ricerca per recuperare le domande candidate precedentemente risposte e (ii) metodi kernel applicati a strutture linguistiche avanzate per riordinare i candidati più promettenti. Mostriamo che i metodi sviluppati per il cQA funzionano bene anche quando applicati ai dati generati nell'ambito dell'assistenza clienti, dove l'utente cerca informazioni riguardo a dei prodotti e una base di dati di domande precedentemente risposte è disponibile. Gli esperimenti sul nostro sistema dimostrano l'appropriatezza del suo utilizzo in uno scenario industriale.*

1 Introduction

In recent years, open-domain Question Answering (QA) has been more and more used by large companies, e.g., IBM, Google, Facebook, Microsoft, etc., for their commercial applications. However, medium and smaller enterprises typically cannot invest billions of dollars in achieving the desired QA accuracy: this limits the use of this technology, especially, in case of less supported languages, e.g., Italian. One viable alternative for smaller companies is the design of close-domain systems looking for answers in specific data. For example, most companies require to quickly and accurately search their own documentation or the one of their customers, which are often available in terms of unstructured text. However, even this scenario is complicated as reaching the a satisfactory accuracy may require a lot of resources.

An interesting alternative is provided by cQA technology, which uses techniques tailored for answering questions in specific forums. In addition, to the intuitive observation that the forum topics are rather restricted, making the retrieval task easier, cQA offers an even more interesting property: when a new question is asked in a forum, instead of searching for an answer, the system tries to look for a similar question. Indeed, similar questions were asked before and may have received answers, thus the system can provide the users with such responses. The main advantage of this approach is that searching for similar questions is much easier than searching for text answering a given question. Due to this, challenges such as SemEval-2017 Task 3 (Nakov et al., 2017) and QA4FAQ (Caputo et al., 2016), aimed at testing current cQA available technology, have been organized.

In this paper, we show that help desk applications, generally required by most companies, can adopt the cQA model to automatize the answering process. In particular, we describe our QA system developed for RGI, which is a software vendor specialized in the insurance businesses. One important task carried out by their help desk software regards answering customers' questions using a ticket system. Already answered tickets are stored in specialized databases but manually finding and routing them to the users is time consuming. We show that our approach, using standard search engines and advanced reranker based on machine learning and NLP technology, can achieve answer recall of almost 85% when considering the top three retrieved tickets. This is particularly interesting because the experimented data and models are completely in Italian, demonstrating the maturity of this technology also for this language.

2 Related Work

The first step for any system that aims at automatically answering questions on cQA sites is to retrieve a set of questions similar to the user's input. Over time, different approaches have been proposed. Early methods used statistical machine translation to retrieve similar questions from large question archivies (Zhou et al., 2011). Other approaches (Cao et al., 2009; Duan et al., 2008) use language models with smoothing to compute semantic similarity between two questions. A different approach that exploits syntactic information was proposed in (Wang et al., 2009). The authors find similar questions by computing similarity between the syntactic trees of the two questions. In this work, we use pairs of similar questions to train our relational model, which detects if two questions have similar semantics.

From an industrial viewpoint, NLP (and especially QA) is one of the hot topics of recent years, although it is still mostly unexplored. Many platforms are emerging in the wide area of chatbot development, e.g., Wit.ai and Api.ai (proposed by Facebook and Google, respectively), which enable intent classification and entity extraction and Meya.ai, which can be used to develop rule-based chatbot systems. However, most of them do not integrate QA models, with the notable exception of Expert Systems' Cogito Answer, recently adopted by Ing. Direct and Responsa.

3 The RGI application scenario

The scope of the experiments for this research is the evaluation of state-of-the-art QA models to automatize help desk (HD) processes of RGI. RGI is an Independent Software Vendor specialized in the Insurance Industry, counting 800 professionals and 12 offices spread across the EMEA region (Italy, Ireland, France, Germany, Tunisia and Luxembourg). Its main product, PASS, is a modular Policy Administration System that enables the end-to-end management of Policies, Claims and Insurance Products configuration across all the insurance channels and business lines. With 103 installations for the insurance companies and other 300 for the brokers, RGI is a leader of its sector in the European market.

The Application Scenario described in this paper focuses on the HD services for PASS offered by RGI during the *roll-out phase* (delivery of the new system to the clients). The use of effective and robust QA models is indeed considered by RGI a crucial aspect for the improvement of the quality of its HD process, in terms of (i) reduction of the response time, (ii) enhancement of the coverage of the services etc., and (iii) general customer satisfaction.

3.1 Task description

During the roll-out phase, new users from a client company start to interact with the PASS system and, in case of a problem, contact the HD provided by RGI. This is structured as a hierarchical organization of operators with different skill levels, which provide answers to the user requests, e.g., HD1 involves operators of Level 1 and regards basic knowledge; HD2 (Level 2) is managed by functional analysts with higher domain knowledge and so on. When a request is sent to an HD operator, a ticket is generated and stored in a trouble ticketing system along with all the relevant information of that request: this includes a description of the problem and the detected solution. Such ticket will be then managed, passed and eventually scaled by all the operators involved in the solution of the problem.

In order to search and provide the right answer to the customer, each HD operator may use the following sources of information: **tickets** opened in the past; **Frequently Asked Questions** (FAQ) and their solutions, stored in a shared repository; a **forum**, where HD operators share their knowledge; **user manuals** of the PASS system released for the client; and **domain knowledge** and expertise of the operator itself.

The objective of this paper is studying the impact of advanced QA systems for the automatization of HD1, using FAQ and tickets data stored in

Table 1: An example of two similar tickets: the one used as query on the left and one retrieved by a search engine (only using question words) on the right.

Question$_{org}$	**Answer**$_{org}$	**Question**$_{rel}$	**Answer**$_{rel}$
Abbiamo bisogno delle credenziali di accesso al sistema. Grazie	Buongiorno, questo l'indirizzo mail al quale scrivere per avere le credenziali di accesso al sistema: xxx@xxx.xx Cordiali saluti	Buongiorno, non troviamo credenziali per accesso sistema. Potete aiutarci? Grazie	Buongiorno, questo l'indirizzo mail al quale scrivere per avere le credenziali di accesso al sistema: xxx@xxx.xx Cordiali saluti

the related repositories.

3.2 Data description

Data was gathered from the HD support system, where technical issues are tracked and fixed. Basically, we have tickets organized in Question/Answer (Q/A) pairs, along with fields related to specific information, such as ticket ID and the domain problem. The original data size was around 40,000 tickets but most of them do not provide useful information. Thus, we designed a preprocessing phase both to clean and prepare a valid data set: first, we detected and filtered out spurious Question-Answer pairs, concerning unanswered problems, using basic heuristics. Second, we extracted a subset of general-knowledge problems by selecting only tickets belonging to HD1 with a resolution time less than two days. In addition, our data was also reviewed by an expert team to further filter out invalid tickets. As a result, the preprocessing ended with a dataset of 656 Q/A pairs spread over 10 question domains. Examples of our data are shown in Table 1.

4 Our QA System

Our system is constituted by (i) a search engine to retrieve questions (along with their associated tickets) similar to the new input question and (ii) a reranker built with state-of-the-art NLP and machine learning technology.

4.1 Question and Ticket Retrieval

We used a standard keyword-based Search Engine (SE) to retrieve a list of questions from our dataset similar to the input one. The score produced by SE is the standard cosine similarity between the vectors of the new and the candidate questions. In particular, we built our SE using Lucene TF-IDF based indexing, available in the open-source ElasticSearch platform.

In order to improve the retrieval quality, we merged user request description (the question) and solution fields in a single joint text to build the ticket index. It should be noted that we only used the question text to build the query for SE as in a real scenario, the asked question is not associated with any answer yet.

For each question, in the filtered data mentioned above, we created a list of Question_original - Question_related pairs, by querying each ticket and collecting the first 10 relevant results. The obtained clustered data set resulted in a list $\langle q_{original}, q_{related}\rangle$ of 656 (tickets) x 10 (retrieved questions). These pairs were annotated by a team of experts with relevant vs. irrelevant labels to create the training and test sets. For example, Table 1 shows a question pair: an original ticket with question and answer on the left, and a similar retrieved ticket on the right.

4.2 Reranking Pipeline

Given the initial rank provided by SE, we apply an advanced NLP pipeline to rerank the questions such that those having the highest probability to be similar to the query are ranked on the top.

NLP pipeline. We used various Italian NLP processors of TextPro (Pianta et al., 2008) and embedded them in a UIMA pipeline, to analyze each ticket question as well as the questions of the tickets in the rank. The NLP components includes, part-of-speech tagging, chunking, named entity recognition, constituency and dependency parsing, etc. The result of the processing is used to produce syntactic representations of the ticket questions, which are then enhanced by relational links, e.g., between matching words, between two questions of a pair. The resulting tree pairs are then used to train a kernel-based reranker.

Kernel-based reranker. A kernel reranker is a function $r : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$, where $\mathcal{Q}$ is a set of questions. Such function tells if questions are similar or not and can be used to sort a set of questions q_r with respect to an original one q_o. These functions can be implemented in many ways, but in this work we used (i) a kernel function applied to the syntactic structure of the pair questions, together

Table 2: Results of the reranker obtained by combining Sim features with TKs.

	5-folds cv				
Model	MRR	MAP	P@1	P@2	P@3
IR baseline	70.85 ± 4.54	63.18 ± 3.37	57.67 ± 6.99	71.79 ± 3.98	77.86 ± 4.69
Sim	71.56 ± 4.16	63.90 ± 2.19	58.39 ± 8.04	72.44 ± 2.45	80.77 ± 3.31
TK	72.45 ± 2.19	67.09 ± 2.33	58.31 ± 3.42	75.34 ± 2.32	80.71 ± 3.36
TK + Sim	75.07 ± 1.67	68.51 ± 1.41	61.54 ± 1.86	77.87 ± 3.27	84.57 ± 2.57

with (ii) some features capturing text similarity between two questions.

Feature Vector model. This feature vector embeds a set of *text similarity* features that capture the relationship between two questions. More specifically, we compute a total of 20 similarities such as *n*-grams, greedy string tiling, longest common subsequences, Jaccard coefficient, word containment, cosine similarity and many others.

Tree Kernel model. This model takes in input two tickets and measures the similarity between their syntactic trees. In particular, we build two macro-trees, one for each ticket in the pair, containing the syntactic trees of sentences in each ticket question. In addition, we link two macro-trees by connecting the phrases of two questions, as done in (Da San Martino et al., 2016). Then, we applied Partial Tree Kernel (Moschitti, 2006) and obtain the following kernel:

$$K(\langle q_o, q_r\rangle^i, \langle q_o, q_r\rangle^j) = TK(t(q_o, q_r)^i, t(q_o, q_r)^j),$$

where q_o is the original ticket question and q_r are the questions of similar tickets. In contrast, the function $t(x, y)$ extract the syntactic tree from the text x, enriching it with REL tags.

5 Experiments

To evaluate our approach, we performed experiments on a dataset composed of $6,650$ pairs of ticket questions annotated with similarity judgment, i.e., Relevant and Irrelevant. We selected only questions having at least one answer in the first 10 retrieved tickets. We performed 5-fold cross-validation and used SVM-Light-TK[1] software to train 5 different reranking models. SVM-Light-TK allows us to learn a reranking model that combines both feature vectors and Tree Kernels. The latter are especially useful because avoid the burden of manually engineering feature for this task. A more detailed description of the Tree Kernel models and Text Similarity features employed by the model is reported in (Da San Martino et al., 2016). Then, we used the learned model to predict similarities for all pairs of questions present in each test fold.

5.1 Results

We conducted three experiments to assess the effectiveness of the different feature sets, similarity features (Sim), TK and TK+Sim in the reranking model. The baseline is computed by means of the rank given by Lucene. Following previous work of the SemEval challenge, we evaluated our ranking with Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Precision at k (P@k).

The results are reported in Tab. 2. As it can be seen, the best results are obtained by combining Sim and TK in the reranker, which improved the MRR and MAP of the IR baseline by 4.22 and 5.33 absolute points, respectively. In addition, P@1, @2 and @3 improved by 3.87, 6.08 and 6.71 absolute points, respectively. This shows the effectiveness of using syntactic structures in powerful algorithms such as TK.

We analyzed some selected errors of our system, focusing on the cases where the search engine performs better than our reranking model. We note that for each cluster of question_original-question_related pairs, when the P@1 is high, our model does not perform better than the search engine, or performs even worse. However, our reranking model always tends to push relevant results on the top.

6 Conclusions

In this paper, we have described our experience in building a QA model for an Italian help desk in the field of insurance policies. Our main findings are: (i) the Italian NLP technology seems enough accurate to support advanced cQA technology based on syntactic structures; (ii) cQA model can boost the retrieval systems targeting text in Italian; and (iii) the achieved accuracy seems appropriate to create business at least in the filed of help desk applications, although it should be considered that our results refer to only questions having an answer in our database.

[1] http://disi.unitn.it/moschitti/Tree-Kernel.htm

References

Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. 2009. The use of categorization information in language models for question retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 265–274. ACM.

Annalina Caputo, Marco de Gemmis, Pasquale Lops, Francesco Lovecchio, Vito Manzari, and Acquedotto Pugliese AQP Spa. 2016. Overview of the evalita 2016 question answering for frequently asked questions (qa4faq) task. In *CLiC-it/EVALITA*.

Giovanni Da San Martino, Alberto Barrón Cedeño, Salvatore Romeo, Antonio Uva, and Alessandro Moschitti. 2016. Learning to re-rank questions in community question answering using advanced features. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1997–2000. ACM.

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *ACL*, volume 8, pages 156–164.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *ECML*, volume 4212, pages 318–329. Springer.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. Semeval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In *LREC*.

Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 187–194. ACM.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 653–662. Association for Computational Linguistics.