Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015

3-4 December 2015, Trento

Editors:

Cristina Bosco
Sara Tonelli
Fabio Massimo Zanzotto

aA Accademia university press

aA

CLiC it

AILC

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

isbn 978-88-99200-62-6
www.aAccademia.it/CLIC_2015

We are glad to introduce CLiC-it 2015 (https://clic2015.fbk.eu/), the second edition of the Italian Conference on Computational Linguistics, organized this year for the first time by the newborn Italian Association for Computational Linguistics (AILC).

AILC (http://www.ai-lc.it/) is born after a long period of discussion within the variegated community linked by the common interest towards Computational Linguistics (CL) in Italy, until now sparse in several research areas and associations. Considering that CL spans over a range of disciplines from Linguistics to Computer Science, AILC proposes the characterization of their members' work in terms of methodologies and approaches, rather than topics. The goal is to collect the different *souls* of CL around the same table, where the future of CL in Italy can be investigated and the initiatives for fostering its development promoted by more coordinated activities, with an emphasis on Italian language.
AILC's main aim is to promote the theoretical and experimental reflection on methodologies, scientific cooperation and development of shared practices, resources and tools, and, last but not least, the transfer of technology and knowledge to the market within the area of CL.
The goals of the Association include the promotion of scientific and educational initiatives for the diffusion of CL, with a special focus on Italian, as well as of the visibility and knowledge diffusion about initiatives and resources, in order to support interdisciplinary projects. AILC also fosters the integration of competences and professional skills from both the humanity and computational area, and the establishment and consolidation of links with other Italian, European or international initiatives around CL, also proposing direct involvement of the Association. AILC also promotes CL within the national policies for university and scientific research.

CLiC-it 2015 is held in Trento on December 3-4 2015, hosted and locally organized by Fondazione Bruno Kessler (FBK), one the most important Italian research centers for what concerns CL. The organization of the conference is the result of a fruitful conjoint effort of different research groups (Università di Torino, Università di Roma Tor Vergata and FBK) showing the nationwide spreading of CL in Italy.
As in the first edition, the main aim of the event is at establishing a reference forum on CL, covering all the aspects needed to describe the multi-faceted and cross-disciplinary reality of the involved research topics and of the Italian community working in this area. Indeed the spirit of CLiC-it is inclusive, in order to build a scenario as much as possible comprehensive of the complexity of language phenomena and approaches to address them, bringing together researchers and scholars with different competences and skills and working on different aspects according to different perspectives.
Relevant topics for CLiC-it 2015 include, but are not limited to, the following thematic areas:
- Information Extraction and Information Retrieval – Area chairs: Roberto Basili (Università di Roma Tor Vergata), Giovanni Semeraro (Università di Bari)
- Linguistic Resources – Area chairs: Maria Simi (Università di Pisa), Tommaso Caselli (Vrije Universiteit Amsterdam), Claudia Soria (ILC - CNR, Pisa)
- Machine Translation – Area chairs: Marco Turchi (FBK, Trento), Johanna Monti (Università di Sassari)
- Morphology, Syntax and Parsing – Area chairs: Felice Dell'Orletta (ILC - CNR, Pisa), Fabio Tamburini (Università di Bologna), Cristiano Chesi (IUSS, Pavia)
- NLP for Digital Humanities – Area chairs: Alessandro Lenci (Università di Pisa), Fabio Ciotti (Università di Roma Tor Vergata)
- NLP for Web and Social Media – Area chair: Francesca Chiusaroli (Università di Macerata), Daniele Pighin (Google Inc.)
- Pragmatics and Creativity – Area chairs: Carlo Strapparava (FBK, Trento), Rossana Damiano (Università di Torino)
- Semantics and Knowledge Acquisition – Area chair: Elena Cabrio (INRIA, Sophia Antipolis), Armando Stellato (Università di Roma Tor Vergata)
- Spoken language processing – Area chairs: Giuseppe Riccardi (Università di Trento), Piero Cosi (ISTC - CNR, Padova)
- Towards EVALITA 2016: challenges, methodologies and tasks – Area chairs: Franco Cutugno (Uni-

versità di Napoli Federico II), Viviana Patti (Università di Torino), Rachele Sprugnoli (FBK, Trento - Università di Trento).

The large number of researchers that have decided to present their work at CLiC-it and the number of directions here investigated are proof of the maturity of our community and a promising indication of its vitality. We received a total of 64 paper submissions, out of which 52 have been accepted to appear in the Conference Proceedings, which are available online and on the OpenEdition platform. Overall, we collected 129 authors from 15 countries.
We are very proud of the scientific program of the conference: it includes two invited speakers, Enrique Alfonseca (Google Research, Zurich) and Paola Merlo (University of Geneva), oral presentations, as well as two poster sessions preceded by booster sessions. Moreover, we organized two panels for discussing the future of CL with the representatives of both Italian associations and industry, and a session for preparing the ground for the next edition of the evaluation campaign for NLP and speech tools for Italian, Evalita (http://www.evalita.org), to be held within CLiC-it 2016.
We are also happy to assign best paper awards to young authors (PhD students and Postdocs) who appear as first author of their paper.

We thank the conference sponsors for their generous support: CELI (Torino), Expert System (Modena), Reveal (Roma), Euregio (Bolzano), Almawave (Roma), ELRA (Parigi).

We also thank the following organizations and institutions for endorsing CLiC-it:
- Società Italiana di Glottologia (SIG)
- Associazione Italiana per l'Intelligenza Artificiale (AI*IA)
- Società di Linguistica Italiana (SLI)
- Associazione Italiana di Linguistica Applicata (AITLA)
- Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD)
- Associazione Italiana Scienze della Voce (AISV)
Last but not least, we thank the area chairs and all the program committee members for their incredible work, the invited speakers for their contribution to make CLIC-it an international event, and all the persons involved in the local organization of the conference in Trento.

November 2015

PROGRAM COMMITTEE

We thank all the members of the Program Committee that helped us in reviewing papers and in improving the scientific quality of the event.

Enrique Alfonseca, Google inc.
Carlo Aliprandi, Synthema Srl
Costanza Asnaghi, Universiteit Gent and Katholieke Universiteit Leuven
Giuseppe Attardi, Università di Pisa
Anabela Barreiro, L^2F-INESC ID Lisboa
Pierpaolo Basile, Università di Bari
Roberto Basili, Università di Roma "Tor Vergata"
Nuria Bel, Universitat Pompeu Fabra
Luisa Bentivogli, FBK Trento
Giacomo Berardi, ISTI-CNR Pisa
Nicola Bertoldi, FBK Trento
Arianna Bisazza, Universiteit van Amsterdam
Bernd Bohnet, Google inc.
Andrea Bolioli, CELI srl.
Francesca Bonin, Trinity College Dublin
Johan Bos, University of Groningen
Federico Boschetti, Università di Pavia and ILC-CNR Pisa
Paul Buitelaar, Insight - National University of Ireland Galway
Harry Bunt, Tilburg University
Jos Guillermo Camargo de Souza, Università di Trento and FBK Trento
Elena Cabrio, INRIA Sophia-Antipolis
Nicoletta Calzolari, ILC-CNR Pisa
Annalina Caputo, Università di Bari
Claudio Carpineto, Fondazione Ugo Bordoni
Vittore Casarosa, ISTI-CNR Pisa
Tommaso Caselli, Vrije Universiteit Amsterdam
Giuseppe Castellucci, Università di Roma "Tor Vergata"
Maria Catricalà, Università di Roma Tre
Mauro Cettolo, FBK Trento
Cristiano Chesi, Istituto Universitario di Studi Superiori di Pavia
Isabella Chiari, Università di Roma "La Sapienza"
Francesca Chiusaroli, Università di Macerata
Fabio Ciotti, Università di Roma "Tor Vergata"
Gianpaolo Coro, ISTI-CNR Pisa
Piero Cosi, ISTC-CNR Padova
Fabio Crestani, Università di Lugano
Danilo Croce, Università di Roma "Tor Vergata"
Franco Cutugno, Università di Napoli "Federico II"
Federica Da Milano, Università di Milano "Bicocca"
Rossana Damiano, Università di Torino
Marco De Gemmis, Università di Bari
Anna De Meo, Università di Napoli "L'Orientale"
Thierry Declerck, DFKI GmbH
Felice Dell'Orletta, ILC-CNR Pisa
Rodolfo Delmonte, Università di Venezia "Ca' Foscari"
Ernesto William De Luca, Leibniz-Institut für internationale Schulbuchforschung
Giorgio Maria Di Nunzio, Università di Padova

Francesca Dovetto, Università di Napoli “Federico II”
Stefano Faralli, Università di Roma “La Sapienza”
Marcello Federico, FBK Trento
Anna Feldman, Montclair State University
Anna Feltracco, FBK Trento
Katja Filippova, Google inc.
Francesca Frontini, ILC-CNR Pisa
Vincenzo Galatà, Free University of Bozen
Aldo Gangemi, Université Paris 13 and CNR-ISTC Roma
Lorenzo Gatti, FBK Trento
Andrea Gesmundo, Google inc.
Alessandro Giuliani, Università di Cagliari
Nicola Grandi, Università di Bologna
Marco Guerini, FBK Trento
Christian Hardmeier, Uppsala Universitet
Diana Inkpen, University of Ottawa
Elisabetta Jezek, Università di Pavia
Mike Kozhevnikov, University of Saarland
John Laudun, University of Louisiana
Alberto Lavelli, FBK Trento
Alessandro Lenci, Università di Pisa
Felicia Logozzo, Università di Roma “Tor Vergata”
Pasquale Lops, Università di Bari
Claudio Lucchese, ISTI-CNR Pisa
Bernardo Magnini, FBK Trento
Simone Magnolini, FBK Trento
Diego Marcheggiani, ISTI-CNR Pisa
Alessandro Mazzei, Università di Torino
John P. Mccrae, Universität Bielefeld
Massimo Melucci, Università di Padova
Stefano Menini, FBK Trento and Università di Trento
Monica Monachini, ILC-CNR Pisa
Massimo Moneglia, Università di Firenze
Simonetta Montemagni, ILC-CNR Pisa
Johanna Monti, Università di Sassari
Roser Morante, Vrije Universiteit Amsterdam
Andrea Moro, Università di Roma “La Sapienza”
Alessandro Moschitti, Qatar Computing Research Institute and Università di Trento
Cataldo Musto, Università di Bari
Franco Maria Nardini, ISTI-CNR Pisa
Fedelucio Narducci, Università di Bari
Costanza Navarretta, University of Copenhagen
Borja Navarro-Colorado, Universidad de Alicante
Roberto Navigli, Università di Roma “La Sapienza”
Matteo Negri, FBK Trento
Vincent Ng, University of Texas at Dallas
Malvina Nissim, University of Groningen
Alessandro Oltramari, Carnegie Mellon University
Antonio Origlia, Università di Napoli “Federico II”
Gözde Özbal, FBK Trento
Alessio Palmero Aprosio, FBK Trento

Silvia Pareti, University of Edinburgh
Patrick Paroubek, LIMSI-CNRS
Lucia Passaro, Università di Pisa
Marco Passarotti, Università Cattolica del Sacro Cuore di Milano
Viviana Patti, Università di Torino
Marco Pedicini, Università di Roma Tre
Raffaele Perego, ISTI-CNR Pisa
Massimo Pettorino, Università degli studi di Napoli "L'Orientale"
Maria Laura Pierucci, Università di Macerata
Daniele Pighin, Google inc.
Massimo Poesio, University of Essex and Università di Trento
Simone Paolo Ponzetto, University of Mannheim
Bruno Pouliquen, World Intellectual Property Organization
Giuseppe Riccardi, Università di Trento
Giuseppe Rizzo, EURECOM
Paolo Rosso, Technical University of Valencia
Irene Russo, ILC-CNR Pisa
Federico Sangati, FBK Trento
Giorgio Satta, Università di Padova
Giovanni Semeraro, Università di Bari
Fabrizio Silvestri, Yahoo Labs
Maria Simi, Università di Pisa
Claudia Soria, ILC-CNR Pisa
Manuela Speranza, FBK Trento
Rachele Sprugnoli, FBK Trento and Università di Trento
Armando Stellato, Università di Roma "Tor Vergata"
Carlo Strapparava, FBK Trento
Francesca Strik Lievers, Università di Milano Bicocca
Fabio Tamburini, Università di Bologna
Marco Turchi, FBK Trento
Paolo Turriziani, Interactive Media
Kateryna Tymoshenko, Università di Trento
Olga Uryupina, Università di Trento
Eloisa Vargiu, Barcelona Digital Technology Center
Marc Verhagen, Brandeis University
Laure Vieu, Institut de Recherche en Informatique de Toulouse
Enrico Zovato, Loquendo S.p.A.

Contents

Bolzano/Bozen Corpus: Coding Information about the Speaker in IMDI Metadata Structure

Marco Angster
Centro di Competenza Lingue
Libera Università di Bolzano
marco.angster@unibz.it

Abstract

English. The paper introduces a new collection of spoken data (the *Bolzano/Bozen Corpus*) available through The Language Archive of Max Planck Institute of Nijmegen. It shows an example of the issues encountered in accommodating information of an existent corpus into IMDI metadata structure. Finally, it provides preliminary reflections on CMDI: a component-based metadata format.

Italiano. *Questo contributo presenta una nuova raccolta di dati di parlato (il* Bolzano/Bozen Corpus*) che è ora disponibile per la consultazione tramite il Language Archive del Max Planck Institute di Nimega. Vi si mostra un esempio dei problemi che si possono incontrare nell'inserimento all'interno della struttura di metadati IMDI delle informazioni relative a un corpus già esistente. Infine, vi si presentano alcune considerazioni preliminari riguardanti il formato di metadatazione CMDI, basato su componenti.*

1 Introduction

Once a Language Resource (LR) exists it should be used, and this entails several problems. First of all it must be available to the public – which may be the academic community, but also industry or institutions – and, given that producing a LR is an expensive task, it would be ideal that a LR could be exploited beyond the originally intended public. The re-usability of a LR is possible provided that it is conceived following shared standards for formats, tagging and metadata.

In this paper I focus on metadata structures, in particular I introduce a collection of spoken data (the *Bolzano/Bozen Corpus*) and I show the problems encountered in fitting the information available about the speakers sampled in the data in IMDI metadata structure.

The paper aims at providing an example of how flexible are the considered metadata structures in accommodating information of existent collections of data and in adapting to the needs of the researcher in sociolinguistics.

2 *Bolzano/Bozen Corpus*

The *Bolzano/Bozen Corpus* (BBC) collects and organises the language data produced during the years by the researchers of the Competence Centre for Language Studies. The common thread of the BBC is constituted by two main elements: the focus on the speech community in Alto Adige/South Tyrol, the trilingual province in Northern Italy of which Bolzano is the administrative centre; the interest on language variation, both in the social environment and in the educational context.

As a language resource the BBC is mainly destined to scholars interested in sociolinguistics and in the issue of multilingualism. Given that it collects different language varieties of the Romance and the German domain, the corpus has the function of providing original documentation for the local spoken language.

In order to give a better accessibility to the data, the corpus is made available to the public through The Language Archive (TLA), a collection of language resources hosted by the Max Planck Institute of Nijmegen (Nederlands).[1] All projects hosted by TLA must adopt a common metadata scheme on which all the structure of the database is built. The standard adopted by TLA used to be IMDI.[2]

[1] Homepage: https://tla.mpi.nl/
Corpora: https://corpus1.mpi.nl/ds/asv/?1

[2] TLA has recently made available to the users also the new, CLARIN supported CMDI metadata format. See below

The projects included in the BBC were obviously already supplied with rich information which had to fit into the metadata structure available.

3 IMDI and <Actors>

IMDI (ISLE/EAGLES Metadata Initiative) is a standard for metadata developed in the late '90s in the realm of standardisation initiatives ISLE (International Standard for Language Engineering) and EAGLES (Expert Advisory Group on Language Engineering Standards) – see Wittenburg et al. (2000). It provides a very rich structure in which information about a corpus, a session (i.e. a subdivision in a corpus, for example an interview), the relevant media files (the recording of an interview) and written resources (a transcription) are included. The session is the most complex substructure, because it may include a wealth of information about the interview itself: its location, its content (genre, communication context, type of task performed, languages used etc.) and its actors (interviewed, interviewee, but also transcriber, etc.).

Since BBC is a collection of data issued from sociolinguistically oriented projects, it appears clear that information about the speaker is of crucial importance and it is a fundamental concern to fit as much information about the speaker as possible in a metadata structure.

As already mentioned, part of metadata related to a session is devoted to the coding of information about people involved in the interview and in the production of the relevant resources. In this part of metadata structure the available tokens of information about a speaker involved in an interview or a language task are to be found. Some classical social variables are available: <Age>, <Sex>, <Education>, <Ethnic group>. Other useful pieces of information may be coded: <Role> ("The functional role of the person participating in the session" (IMDI, 2003); e.g. interviewer, speaker/signer, annotator, etc.), <Language> ("The language the person participating in the session is familiar with" (IMDI, 2003); more than one language may be added). A further element, <Family Social Role>, is available for coding "[t]he social or family role of the person participating in the session" and may be used "[f]or instance when interviewing part of a family group" where it can "specify the mutual relations within the group" (IMDI, 2003).

It is worth noting about the element <Language> that it is not intended to specify the language used in the session, for which another element is provided at an upper level under the node <Session> of the metadata structure. In this sense <Language> may be considered a good correspondent to the sociolinguistic concept of linguistic repertoire (Gumperz, 1964).

4 Speakers in Komma and Kontatto projects

I turn now back to BBC to show what information available about speakers involved in two different projects may be included in the structure sketched above.

The projects that I take into account are both focussed on South Tyrol, but with quite different perspectives, types of tasks accomplished and homogeneity of speakers involved. KOMMA (SprachKOMpetenzen von MAturandinnen und Maturanden) consists in the analysis of written and oral productions of high school graduands of the German schools of South Tyrol. It aims at studying the competence of the German standard language of young adults in mono- and multilingual settings in order to analyse linguistic phenomena, to find traces of multilingual competence or of a specific sociolinguistic background. At present the data available via TLA involve 41 students, all of German mother tongue: interviews on the language biography of the students and the renarration of a sequence of a Charlie Chaplin film (The Circus) are currently available.

More than a half of the students are female, most of them are 19 years old at the time of the interview. The picture is thus quite homogeneous, while the only variable which differentiates sets of students is the geographic area of the school they attended. This variable is coded as the location where the interaction takes place (<Location>). All students except two have both parents of German mother tongue, but this particular may not be coded in the metadata structure, unless we explicit it in the field <Description>. This is not an excellent solution, but a useful workaround to put a token of information which would be otherwise lost.

The second project considered here is Kontatto (Italiano-tedesco: aree storiche di contatto in Sudtirolo e in Trentino). The aim of the project is

section 5.

to document the present day Italian-German contacts in Bassa Atesina (the area south of Bolzano). The area is highly interesting for sociolinguistics and contact linguistics because there the interaction between German and Romance dialectal varieties dates back to a more remote time than in the rest of South Tyrol. A multilingual and multidialectal corpus of map tasks ((Anderson et al, 1991)) has been created to tackle the objective of documenting the linguistic productions of the speakers in the area.

The speakers involved in Kontatto are less homogeneous: they differ for age, occupation, own linguistic repertoire and linguistic background (parents' mother tongue, variety spoken where they live), place of origin of the parents, place of residence (as opposed to <Location>). This wealth of data – with the exception of the variables already mentioned above for KOMMA – would all be included in a <Description> field if one desires to keep this information available to the user interested in correctly interpreting the relevant data.

As for the case of KOMMA this could be a workaround, but a much more expensive one, from the point of view of future information retrieval. A metadata element is, let's say, a box where information is stored, but it is a box with an own particular tag, which indicates what is in. In addition this tag gives sense to the content and makes possible and easier to find the content itself among all information available. Putting information in a <Description> field corresponds to give up the possibility to exploit its classifying potential at a later time, thus making the information almost unusable.

5 CMDI: a very customisable, but closed structure

The limits of IMDI as a metadata structure are nonetheless well-known as we can read in the User Guide of the CLARIN-D infrastructure (Váradi et al, 2008):

> "Most existing metadata schemas for language resources seemed to be too superficial (e.g. OLAC) or too much tailored towards specific research communities or use cases (e.g. IMDI)." (CLARIN-D User Guide, 2012)

This words express the need of a new, more comprehensive standard for metadata description which could give to the researchers the possibility to tailor metadata profiles on the needs of their sub-disciplines. The new standard should display the following crucial features:

> 1. allow users to define their own components resulting in tailored profiles,
> 2. the components need to make use of categories the definitions of which are registered in ISOcat (see the section called "ISOcat, a Data Category Registry"), and
> 3. semantic interoperability and interpretability [must be] guaranteed by fine-grained semantics.
>
> (CLARIN-D User Guide, 2012)

At present CLARIN-D supports a new standard for metadata: CMDI. It is more flexible in that it allows the researcher to create own components rejecting profiles (for example <Session> or <Actor(s)>) which may be too restrictive or too fine-grained for their specific needs and modifying existing ones by adding or removing elements or by creating brand new profiles.

It is difficult for me to judge how open is CMDI for creating new profiles and how much flexible it is. In fact the possibility of creating new components and profiles is restricted to the accredited users of CLARIN centres.

In any case I try to imagine how should for instance a new CMDI-compliant component be structured in order to hold all information needed to give a complete description of a student of the KOMMA project. As shown above, the main problem is the impossibility to include information about parents' mother tongue. The solution of this lack would be to attribute to an actor involved in an interview a relation to another person – described as father or mother using the field <Family Social Role> – which is nonetheless not present in the interaction. Another possibility would be to code under the <Language> node one or more <Family Social Role> items pointing at the people with whom the relevant actor has a language in common. However solved, the problem apparently may be overcome.

It is worth noting that CMDI components are still based on the same elements on which IMDI is based. More precisely CMDI elements must point to a trusted data category registry (DCR), among

which ISOcat used to be one of the most used in IMDI structure.[3] In Kontatto, as we have seen, speaker profiles are very complex, but a wealth of information is available to the researcher. To characterise some of the interactions sampled in the project it may be useful to explicit both the "mutual relations within the group" as can be done through the field <Family Social Role> and the social background of the same speaker, for example its occupation, beyond the other social features he or she has. If an actor is the father of another actor, this should be independent from the fact that he is a boss, a doctor, a mayor, a teacher or a shaman/priest – just to cite some of the values of the open vocabulary category <Family Social Role> that are nonetheless suggested in IMDI Guidelines.

This fact highlights two different kinds of problems. The first one is a limit of IMDI: in its structure only one value for <Family Social Role> was allowed leading to the odd conclusion that one cannot be at the same time a father and a doctor. The second problem is more critical and significantly it is inherited by CMDI: <Family Social Role> is a category which is useful only to provide an explanation of the consequences for the interaction of the fact that a boss rather than a shaman/priest or a brother interacts with another actor. The category is instead simply unsatisfactory to accommodate background information, maybe irrelevant for the interaction but crucial to evaluate speaker's choices, such as what is the occupation of an actor, feature which contributes to the definition of the classic sociolinguistic variable of social class (Ash, 2003). However the unsatisfactory category <Family Social Role> appears to have no better alternative in ISOcat DCR, which is quite disappointing, because if I want to create my brand new <Actor> profile within CMDI I need to point to some existent data category and uses which contradict the meaning of a category are rightly deprecated.

As said, adding new data categories implies adding them to a Data Category Registry (DCR). Max Planck Institute for Psycholinguistics ceased in December 2014 to be the Registration Authority for ISOcat DCR. Now the new DCR for CMDI is CCR (CLARIN Concept Registry) which is nonetheless closed to changes. To add or change categories in the CCR the national CCR coordinators must be contacted, because only they are able to input new concepts and edit already existent ones.[4] This means that, in order to include a reasonable field <Occupation> instead of <Family Social Role> I have to operate outside CMDI and propose a new category to CCR national coordinators.

6 Conclusion

In this paper, I have shown an example of the difficulty of using a metadata structure to accommodate information on speaker's linguistic background. I have taken into account the case of *Bolzano Bozen Corpus* and two sociolinguistically oriented projects (KOMMA, Kontatto) hosted on The Language Archive.

IMDI, the former standard of TLA, is now an outdated tool and is too rigid to adapt to specific purposes. The new standard CMDI provides huge possibilities to the research community to define metadata formats tailored on specific needs. However CMDI does not provide until now satisfactory profiles and components for sociolinguistic studies, especially as far as background information about the speaker is concerned. Furthermore, direct contribution to CMDI components is restricted to CLARIN centres and in some crucial cases even categories available in CMDI are unsatisfactory and must be proposed to the relevant (and closed) DCR. The case I have proposed shows on the one hand the possibilities of CMDI. However, on the other hand, the difficulty to contribute to CMDI profiles and components from outside CLARIN may lead to the uncomfortable condition of having huge amounts of data with unsatisfactory metadata, which have low possibilities to be re-used, failing one of the main objectives of a standardisation initiative.

Acknowledgments

I thank the project leaders of KOMMA (Rita Franceschini) and Kontatto (Silvia Dal Negro) and their collaborators for their support during the elaboration of the data which have been loaded on TLA platform. I also thank Roberto Cappuccio for technical support on many occasions. Finally I thank three anonymous reviewers for their useful comments.

[3] The list of data categories of ISOcat is available for consultation at `http://www.isocat.org/`.

[4] I thank an anonymous reviewer for pointing me out this possibility.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The Hcrc Map Task Corpus. *Language and Speech*, 34(4):351-366.

Sharon Ash. 2003. Social Class. In: J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes. *The Handbook of Language Variation and Change*. Malden/Oxford: Blackwell Publishing. 402–422.

CLARIN-D User Guide. 2012. Version: 1.0.1. `http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf`.

John J. Gumperz. 1964. Linguistic and social interaction in two communities. *American Anthropologist*, 66/(6/2): 137–53.

IMDI Metadata Elements for Session Descriptions. 2003. Version 3.0.4. MPI Nijmegen. `https://tla.mpi.nl/?attachment_id=4532`.

Tamás Váradi, Peter Wittenburg, Steven Krauwer, Martin Wynne and Kimmo Koskenniemi. 2008. CLARIN: Common language resources and technology infrastructure. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. 1244-1248.

P. Wittenburg, D. Broeder and B. Sloman. 2000. EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper. LREC 2000 Workshop, Athens. `http://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf`.

Detecting the scope of negations in clinical notes

Giuseppe Attardi, Vittoria Cozza, Daniele Sartiano
Dipartimento di Informatica - Università di Pisa
Largo B. Pontecorvo, 3
{attardi,cozza,sartiano}@di.unipi.it

Abstract

English. We address the problem of automatically detecting the scope of negations and speculations in clinical notes, by proposing a machine-learning algorithm that analyzes the dependency tree of a sentence. Given a negative/speculative cue, the algorithm tries to extend the boundary of the scope towards the left and the right, by navigating through the parse tree. We report on experiments with the algorithm using the Bioscope corpus.

Italiano. *Il lavoro affronta il problema di identificare l'ambito a cui si applica una negazione o un'espressione dubitativa nel testo di un referto medico. Si propone un algoritmo di apprendimento automatico, che analizza l'albero di parsing di ogni frase. Dato un indizio di negazione/ipotesi, l'algoritmo cerca di estendere il confine dell'ambito sia a destra che a sinistra, attraversando l'albero di parsing. Riportiamo infine i risultati di esperimenti con l'algoritmo effettuati usando il corpus Bioscope.*

1 Introduction

Clinical notes are a vast potential source of information for healthcare systems, from whose analysis valuable data can be extracted for clinical data mining tasks, for example confirming or rejecting a diagnosis, predicting drug risks or estimating the effectiveness of treatments. Clinical notes are written in informal natural language, where, besides annotating evidence collected during a patient visit, physician report historical facts about the patient and suggested or discarded hypothesis. Annotations about dismissed hypotheses or evidence about the absence of a phenomenon are particularly abundant in these notes and should be recognized as such in order to avoid misleading conclusions. A standard keyword based search engine might for example return many irrelevant documents where a certain symptom is mentioned but it does not affect the patient.

Medical records are currently analysed by clinical experts, who read and annotate them manually. In some countries like Spain, it has become mandatory by law for all medical records to be annotated with the mentions of any relevant reported fact, associated with their official ICD9 code. To assign the right ICD9 code, it is of critical importance to recognize the kind of context of each mention: assertive, negative or speculative. In the BioScope corpus, a collection of bio-medical text, one out of eight sentences indeed contains negations (Vincze et al. (2008)).

In order to automate the process of annotation of clinical notes, the following steps can be envisaged:

1. recognition of medical entities, by exploiting techniques of named entity (NE);
2. normalization and association to a unique official concept identifier to their key terminology from UMLS metathesaurus (O. Bodenreider, 2004);
3. detection of negative or speculative scope.

NE recognition and normalization steps can be performed by relying on shallow analysis of texts (for an exhaustive and updated overview of the state of the art, see Pradhan et al. (2014)). The identification of negative or speculative scope, instead, cannot just rely on such simple text analysis techniques, and would require identifying relations between parts, by means of a deeper syntactic-semantic analysis of sentences.

This work presents a novel algorithm that learns to determine the boundaries of negative and speculative scopes, by navigating the parse tree of a sentence and by exploiting machine learning techniques that rely on features extracted from the analysis of the parse tree.

2 Related Work

Negation and uncertainty detection are hard issues for NLP techniques and are receiving in-

creasing attention in recent years. For the detection of negative and speculative scope, both rule-based approaches and machine learning approaches have been proposed.

Harkema et al. (2010) propose a rule-based algorithm for identifying trigger terms indicating whether a clinical condition is negated or deemed possible, and for determining which text falls within the scope of those terms. They use an extended cue lexicon of medical conditions (Chapman et al., 2013). They perform their analysis for English as well as for low resources languages, i.e., Swedish. Their experiments show that lexical cues and contextual features are quite relevant for relation extraction i.e., negation and temporal status from clinical reports.

Morante et al. (2008) explored machine-learning techniques for scope detection. Their system consists of two classifiers, one that decides which tokens in a sentence are negation signals, and another that finds the full scope of these negation signals. On the Bioscope corpus, the first classifier achieves an F1 score of 94.40% and the second 80.99%.

Also Díaz et al. (2012) propose a two-stage approach: first, a binary classifier decides whether each token in a sentence is a negation/speculation signal or not. A second classifier is trained to determine, at the sentence level, which tokens are affected by the signals previously identified. The system was trained and evaluated on the clinical texts of the BioScope corpus. In the signal detection task, the classifier achieved an F1 score of 97.3% in negation recognition and 94.9% in speculation recognition. In the scope detection task, a token was correctly classified if it had been properly identified as being inside or outside the scope of all the negation signals present in the sentence. They achieved an F1 score of 93.2% in negation and 80.9% in speculation scope detection.

Sohn et al. (2012) developed hand crafted rules representing subtrees of dependency parsers of negated sentences and showed that they were effective on a dataset from their institution.

Zou et al. (2015) developed a system for detecting negation in clinical narratives, based on dependency parse trees. The process involves a first step of negative cue identification that exploits a binary classifier. The second step instead analyses the parse tree of each sentence and tries to identify possible candidates for a negative scope extracted with a heuristics: starting from a cue, all ancestors of the cue are considered, from which both the full subtree rooted in the ancestor and the list of its children are considered as candidates. A classifier is then trained to recognize whether any of these candidates falls within the scope of the cue. The system was trained on a Chinese corpus manually annotated including scientific literature and financial articles. At prediction time, besides the classifier, also a set of rules based on a suitable lexicon is used to filter the candidates and to assign them to the scope of a cue. Since the classifier operates independently on each candidate, it may happen that a set of discontiguous candidates is selected. A final clean up step is hence applied to combine them. This system achieved an F1 score below 60%.

3 Negation and speculation detection

For the cue negation/speculation detection, we apply a sequence tagger classifier that recognizes phrases annotated with negation and speculation tags. The cui exploits morphological features, attribute and dictionary features.

For scope detection, we implemented a novel algorithm that explores the parse tree of the sentence, as detailed in the following.

3.1 Scope Detection

For identifying negative/speculative contexts in clinical reports, we exploit information from the

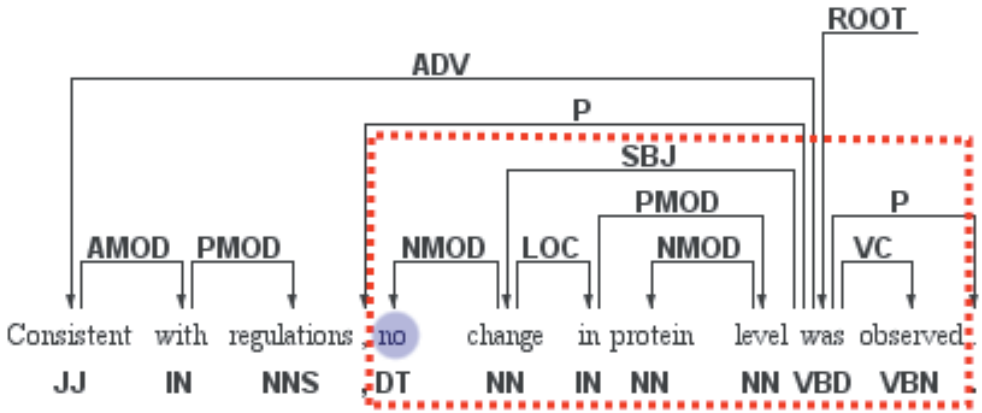


Figure 0. Example of parse tree with a negative scope.

parse tree of sentences. Our approach is however different from the one by Zou et al. (2015), which has the drawback, as mentioned earlier, of operating independently on subtrees and hence it requires an extra filtering step to recombine the candidates and to exclude poor ones according to lexical knowledge.

Our approach assumes that scopes are contiguous and they contain the cue. Hence, instead of assembling candidates independently of each other, our process starts from a cue and tries to expand it as far as possible with contiguous subtrees either towards the left or towards the right.

In the description of the algorithm, we will use the following definitions.

Definition. *Scope adjacency order is a partial order such that, for two nodes x, y of a parse tree, $x < y$ iff x and y are consecutive children of the same parent, or x is the last left child of y or y is the first right child of x.*

Definition. *Right adjacency list. Given a word w_i in a parse tree, the right adjacency list of w_i ($RAL(w_i)$) consists of the union of $RA = \{w_j \mid w_i < w_j\}$ plus $RAL(y)$ where y is the node in RA with the largest index.*

Definition. *Left adjacency list. Symmetrical of Left adjacency list.*

The algorithm for computing the scope S of a cue token at position c in the sentence, exploits the definitions of *RAL* and *LAL* and is described below.

Algorithm.

1. S = $\{w_c\}$
2. for w_i in $LAL(w_c)$ sorted by reverse index
   if w_i belongs to the scope,
   $S = S \cup \{w_k \mid i \leq k < c\}$
   otherwise proceed to next step.
3. for w_i in $RAL(w_c)$ sorted by index
   if w_i belongs to the scope,
   $S = S \cup \{w_k \mid c < k \leq i\}$
   Otherwise stop.

In essence, the algorithm moves first towards the left as far as possible, and whenever it adds a node in step 2, it also adds all its right children, in order to ensure that the scope remains contiguous. It then repeats the same process towards the right.

Lemma. *Assuming that the parse tree of the sentence is non-projective, the algorithm produces a scope S consisting of consecutive tokens of the sentence.*

The proof descends from the properties of non-projective trees.

The decision on whether a candidate belongs to a scope is entrusted to a binary classifier which is trained on the corpus, using features from the nodes in the context of the candidate.

These are nodes selected from the parse tree. In particular, there will be two cases to consider, depending on the current step of the algorithm. For example, in step 2 the nodes considered are illustrated in Figure 1.

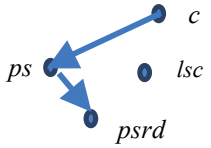


Figure 1. *lsc* is the leftmost child of *c* within the current scope, *ps* is its left sibling, *psrd* is the rightmost descendant of *ps*.

Below we show which nodes are considered for feature extraction in step 3:

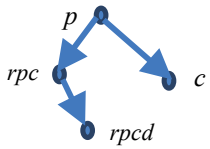


Figure 2. *c* is the leftmost child of *p*, *rpc* is its rightmost child of *p*, *rpcd* is the rightmost descendant of *rpc*

The features extracted from these tokens are: form, lemma, POS, dependency relation type of the candidate node *c*, the cue node, *rpcd* and *psrd*; the distance between node *c* and the cue node; the number of nodes in the current scope; if there are other cues in the subtree of node *c*; the dependency relation types of the children of node *c*; whether the nodes *psrd* and *rpcd* are within the scope; the part of speech, form, lemma and dependency relation types of *lsc* and *rpc*.

We illustrate which nodes the algorithm would visit, on the parse tree of Figure 0. The negative cue is given by the token "`no`", marked in grey in the figure. Initially S = {`no`}, and LAL(no) = {`Consistent`, `,`}, while RAL(no) = {`change`, `in`, `level`, `was`, `observed`, `.`}. The word with largest index in LAL is "," it is not within the scope, hence S stays the same and we proceed to step 3. The token with smallest index in RAL is "`change`", which is part of the scope, hence S = {`no`, `change`}. The next token is "`in`", which also gets added to S, becoming S = {`no`, `change`, `in`}. The next token is "`level`", which is part of the scope: it is added to the scope as well as all

tokens preceding it ("protein"), obtaining {no, change, in, protein, level}. The next two tokens are also added and the algorithm terminates when reaching the final dot, which is not part of the scope, producing S = {no, change, in, protein, level, was, observed}.

Lemma. *The algorithm always terminates with a contiguous sequence of tokens in S that include the cue.*

Notice that differently from (Zou et al. (2015)), our algorithm may produce a scope that is not made of complete subtrees of nodes.

3.2 Experiments

We report an experimental evaluation of our approach on the BioScope corpus, where, according to Szarvas et al. (2008), the speculative or negative cue is always part of the scope.

We pre-processed a subset of the corpus for a total of 17.766 sentences, with the Tanl pipeline (Attardi et al., 2009a), then we splitted it into train, development and test sets of respectively 11.370, 2.842 and 3.554 sentences.

In order to prepare the training corpus, the BioScope corpus was pre-processed as follows. We applied the Tanl linguistic pipeline in order to split the documents into sentences and to perform tokenization according to the Penn Treebank (Taylor et al., 2003) conventions. Then POS tagging was performed and finally dependency parsing with the Desr parser (Attardi, 2006) trained on the GENIA Corpus (Kim et al. 2003).

The annotations from BioScope were integrated back into the pre-processed format using an IOB notation (Speranza, 2009). In particular, two extra columns were added to the CoNLL-X file format. One column for representing negative or speculative cues, using tags NEG and SPEC along with a cue id. One other column for the scope, containing the id of the cue it refers to, or '_' if the token is not within a scope. If a token is part of more then one scope, the id of the cue of each scope is listed, separated by comma.

Here is an example of annotated sentence:

ID	FORM	CUE	SCOPES
1	The	O	_
2	results	O	_
3	indicate	B-SPEC	3
4	that	I-SPEC	3
5	expression	O	3
6	of	O	3
7	these	O	3
8	genes	O	3
9	could	B-SPEC	3, 9
10	contribute	O	3, 9
11	to	O	3, 9
12	nuclear	O	3, 9
13	signaling	O	3, 9
15	mechanisms	O	3, 9

where "could contribute to nuclear signaling mechanisms" is a nested scope within "indicate that expression of these genes could contribute to nuclear signaling mechanisms", whose cues are respectively "could" and "indicate that".

For the cue detection task, we experimented with three classifiers:

1. a linear SVM classifier implemented using the libLinear library (Fan et al. 2008)
2. Tanl NER (Attardi et al., 2009b), a statistical sequence labeller that implements a Conditional Markov Model.
3. deepNL (Attardi, 2015) is a Python library for Natural Language Processing tasks based on a Deep Learning neural network architecture. DeepNL also provides code for creating word embeddings from text using either the Language Model approach by Collobert et al. (2011) or Hellinger PCA, as in (Lebret et al., 2014).

The features provided to classifiers 1) and 2) included morphological features, lexical features (i.e. part of speech, form, lemma of the token and its neighbours), and a gazetteer consisting of all the cue words present in the training set.

The solution based on DeepNL reduces the burden of feature selection since it uses word embeddings as features, which can be learned through unsupervised techniques from plain text; in the experiments, we exploited the word embedding from Collobert et al. (2011). Besides word embeddings, also discrete features are used: suffixes, capitalization, Part of speech and presence in a gazetteer extracted from the training set.

The best results achieved on the test set, with the above mentioned classifier, are reported in Table 1.

	Precision	Recall	F1
LibLinear	88.82%	90.46%	89.63%
Tanl NER	**91.15%**	90.31%	**90.73%**
DeepNL	88.31%	**90.69%**	89.49%

Table 1. Negation/Speculation cue detection results.

The classifier, used in the algorithm of scope detection for deciding whether a candidate be-

longs to a scope or not, is a binary classifier, implemented using libLinear.

The performance of the scope detection algorithm is measured also in terms of Percentage of Correct Scopes (PCS), a measure that considers a predicted scope correct if it matches exactly the correct scope. Precision/Recall are more tolerant measures since they count each correct token individually.

The results achieved on our test set from the BioScope corpus are reported in Table 2.

Precision	Recall	F1	PCS
78.57%	79.16%	78.87%	54.23%

Table 2. Negation/Speculation Scope detection results

We evaluated the performance of our algorithm also on the dataset from the CoNLL 2010 task 2 and we report the results in Table 3, compared with the best results achieved at the challenge (Morante et al. 2010).

	Precision	Recall	F1
Morante et al.	59.62%	55.18%	57.32%
Our system	**61.35%**	**63.68%**	**62.49%**

Table 3. Speculation scope detection

We can note a significant improvement in Recall, that leads also to an relevant improvement in F1.

4 Conclusions

We have described a two-step approach to speculation and negation detection. The scope detection step exploits the structure of sentences as represented by its dependency parse tree. The novelty with respect to previous approaches also exploiting dependency parses is that the tree is used as a guide in the choice of how to extend the current scope. This avoids producing spurious scopes, for example discontiguous ones. The algorithm also may gather partial subtrees of the parse. This provides more resilience and flexibility. The accuracy of the algorithm of course depends on the accuracy of the dependency parser, both in the production of the training corpus and in the analysis. We used a fast transition-based dependency parser trained on the Genia corpus, which turned out to be adequate for the task. Indeed in experiments on the BioScope corpus the algorithm achieved accuracy scores above the state of the art.

References

Giuseppe Attardi. 2006. Experiments with a Multilanguage Non-Projective Dependency Parser, *Proc. of the Tenth Conference on Natural Language Learning*, New York, (NY).

Giuseppe Attardi et al.. 2009a. Tanl (Text Analytics and Natural Language Processing). SemaWiki project: http://medialab.di.unipi.it/wiki/SemaWiki

Giuseppe Attardi, et al. 2009b. The Tanl Named Entity Recognizer at Evalita 2009. In *Proc. of Workshop Evalita'09* - Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, ISBN 978-88-903581-1-1.

Giuseppe Attardi. 2015. DeepNL: a Deep Learning NLP pipeline. *Workshop on Vector Space Modeling for NLP, NAACL 2015*, Denver, Colorado (June 5, 2015).

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, vol. 32, no. supplement 1, D267–D270.

Wendy W. Chapman, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E. Chapman, Michael Conway, Melissa Tharp, Danielle L. Mowery, Louise Deleger. 2013. Extending the NegEx Lexicon for Multiple Languages. *Proceedings of the 14th World Congress on Medical & Health Informatics* (MEDINFO 2013).

Ronan Collobert et al. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, 2461–2505.

N. P. Cruz Díaz, et al. 2012. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American society for information science and technology*, 63.7, 1398–1410.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9(2008), 1871–1874.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2010. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, Volume 42, Issue 5, 839–851.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-text mining. *ISMB* (Supplement of *Bioinformatics*), pp. 180–182.

Rémi Lebret and Ronan Collobert. 2014. Word Embeddings through Hellinger PCA. *EACL 2014*: 482.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Proc-*

essing (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 715–724.

Roser Morante, Vincent Van Asch, and Walter Daelemans. 2010. Memory-based resolution of in-sentence scopes of hedge cues. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning Shared Task*. Association for Computational Linguistics, 2010.

Sameer Pradhan, et al. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, August 2014, Dublin, Ireland, pp. 5462.

Sunghwan Sohn, Stephen Wu, Christopher G. Chute. 2012. Dependency parser-based negation detection in clinical narratives. *Proceedings of AMIA Summits on Translational Science*. 2012: 1.

Maria Grazia Speranza. 2009. The named entity recognition task at evalita 2007. *Proceedings of the Workshop Evalita*. Reggio Emilia, Italy.

György Szarvas, Veronika Vincze, Richárd Farkas, János Csirik, The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts, *Proceedings of the Workshop on Current Trends in Biomedical Natural Language* Processing, June 19-19, 2008, Columbus, Ohio.

Ann Taylor, Mitchell Marcus and Beatrice Santorini. 2003. The Penn Treebank: An Overview, chapter from Treebanks, *Text, Speech and Language Technology*, Volume 20, pp 5-22, Springer Netherlands.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, Janos Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(Suppl 11), S9.

Bowei Zou, Guodong Zhou and Qiaoming Zhu. Negation and Speculation Identification in Chinese Language. *Proceeding of the Annual ACL Conference 2015*.

Deep Learning for Social Sensing from Tweets

Giuseppe Attardi, Laura Gorrieri, Alessio Miaschi, Ruggero Petrolito
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
I-56127 Pisa, Italy
attardi@di.unipi.it

Abstract

English. Distributional Semantic Models (DSM) that represent words as vectors of weights over a high dimensional feature space have proved very effective in representing semantic or syntactic word similarity. For certain tasks however it is important to represent contrasting aspects such as polarity, opposite senses or idiomatic use of words. We present a method for computing discriminative word embeddings can be used in sentiment classification or any other task where one needs to discriminate between contrasting semantic aspects. We present an experiment in the identification of reports on natural disasters in tweets by means of these embeddings.

Italiano. *I Distributional Semantic Model (DSM) rappresentano le parole come vettori di pesi in uno spazio di feature ad alte dimensioni, e si sono dimostrati molto efficaci nel rappresentare la similarità semantica o sintattica tra parole. Per certi compiti però è importante rappresentare aspetti contrastanti come la polarità, significati opposti o parole usate con significato idiomatico. Presentiamo un metodo per calcolare dei word embedding discriminativi che possono essere usati nella sentiment classification o per qualunque altro compito dove vi sia necessità di discriminare tra aspetti semantici contrastanti. Presentiamo un esperimento sull'identificazione di tweet relativi a calamità naturali utilizzando questi embedding.*

1 Introduction

Distributional Semantic Models (DSM) that represent words as vectors of weights over a high dimensional feature space (Hinton et al., 1986), have proved very effective in representing semantic or syntactic aspects of lexicon. Incorporating such representations has allowed improving many natural language tasks. They also reduce the burden of feature selection since these models can be learned through unsupervised techniques from plain text.

Deep learning algorithms for NLP tasks exploit distributional representation of words. In tagging applications such as POS tagging, NER tagging and Semantic Role Labeling (SRL), this has proved quite effective in reaching state of art accuracy and reducing reliance on manually engineered feature selection (Collobert & Weston, 2008).

Word embeddings have been exploited also in constituency parsing (Collobert, 2011) and dependency parsing (Chen & Manning, 2014). Blanco et al. (2015) exploit word embeddings for identifying entities in web search queries.

Traditional embeddings are created from large collections of unannotated documents through unsupervised learning, for example building a neural language model (Collobert et al. 2011; Mikolov et al. 2013) or through Hellinger PCA (Lebrét and Collobert, 2013). These embeddings are suitable to represent syntactic similarity, which can be measured through the Euclidean distance in the embeddings space. They are not appropriate though to represent semantic dissimilarity, since for example antonyms end up at close distance in the embeddings space

In this paper we explore a technique for building *discriminative word embeddings*, which incorporate semantic aspects that are not directly

obtainable from textual collocations. In particular, such embedding can be useful in sentiment classification in order to learn vector representations where words of opposite polarity are distant from each other.

2 Building Word Embeddings

Word embeddings provide a low dimensional dense vector space representation for words, where values in each dimension may represent syntactic or semantic properties.

For creating the embeddings, we used DeepNL[1], a library for building NLP applications based on a deep learning architecture. DeepNL provides two methods for building embeddings, one is based on the use of a neural language model, as proposed by Collobert et al. (2011) and one based on a spectral method as proposed by Lebret and Collobert (2013).

The neural language method can be hard to train and the process is often quite time consuming, since several iterations are required over the whole training set. Some researcher provide precomputed embeddings for English[2].

Mikolov et al. (2013) developed an alternative solution for computing word embeddings, which significantly reduces the computational costs and can also exploit concurrency trough the Asynchronous Stochastic Gradient Descent algorithm. An optimistic approach to matrix updates is also exploited to avoid synchronization costs.

The authors published single-machine multi-threaded C++ code for computing the word vectors[3]. A reimplementation of the algorithm in Python, but with core computations in C, is included in the Genism library (Řehůřek and Sojka, 2010)

Lebret and Collobert (2013) have shown that embeddings can be efficiently computed from word co-occurrence counts, applying Principal Component Analysis (PCA) to reduce dimensionality while optimizing the Hellinger similarity distance.

Levy and Goldberg (2014) have shown similarly that the skip-gram model by Mikolov et al. (2013) can be interpreted as implicitly factorizing a word-context matrix, whose values are the pointwise mutual information (PMI) of the respective word and context pairs, shifted by a global constant.

2.1 Discriminative Word Embeddings

For certain tasks, as for example sentiment analysis, semantic similarity is not appropriate, since antonyms end up at close distance in the embeddings space. One needs to learn a vector representation where words of opposite polarity are distant.

Tang et al. (2013) propose an approach for learning sentiment specific word embeddings, by incorporating supervised knowledge of polarity in the loss function of the learning algorithm. The original hinge loss function in the algorithm by Collobert et al. (2011) is:

$$\mathcal{L}_{CW}(x, x^c) = \max(0, 1 - f_\theta(x) + f_\theta(x^c))$$

where x is an ngram and x^c is the same ngram corrupted by changing the target word with a randomly chosen one, $f_\theta(\cdot)$ is the feature function computed by the neural network with parameters θ. The sentiment specific network outputs a vector of two dimensions, one for modeling the generic syntactic/semantic aspects of words and the second for modeling polarity.

A second loss function is introduced as objective for minimization:

$$\mathcal{L}_{SS}(x, x^c) = \max(0, 1 - \delta_s(x) f_\theta(x)_1 + \delta_s(x) f_\theta(x^c)_1)$$

where the subscript in $f_\theta(x)_1$ refers to the second element of the vector and $\delta_s(x)$ is an indicator function reflecting the sentiment polarity of a sentence, whose value is 1 if the sentiment polarity of x is positive and -1 if it is negative.

The overall hinge loss is a linear combination of the two:

$$\mathcal{L}(x, x^c) = \alpha \mathcal{L}_{CW}(x, x^c) + (1 - \alpha) \mathcal{L}_{SS}(x, x^c)$$

Generalizing the approach to discriminative word embeddings entails replacing the loss function $\mathcal{L}_{ss}$ with a one-vs-all hinge loss function:

$$\mathcal{L}_h(x, t) = \max(0, 1 + \max_{y \neq t}(f(x)_t - f(x)_y))$$

where t is the index of the correct class.

The DeepNL library provides a training algorithm for discriminative word embedding that performs gradient descent using an adaptive learning rate according to the AdaGrad method. The algorithm requires a training set consisting of documents annotated with their discriminative value, for example a corpus of tweets with their sentiment polarity, or in general documents with

[1] https://github.com/attardi/deepnl

[2] http://ronan.collobert.com/senna/, http://metaoptimize.com/projects/wordreprs/, http://www.fit.vutbr.cz/~imikolov/rnnlm/, http://ai.stanford.edu/~ehhuang/

[3] https://code.google.com/p/word2vec

multiple class tags. The algorithm builds embeddings for both unigrams and ngrams at the same time, by performing variations on a training sentence replacing not just a single word, but a sequence of words with either another word or another ngram.

3 Deep Learning Architecture

The Deep Learning architecture used for training discriminative word embeddings consists of the following layers:

1. Lookup layer: extracts the embedding vector associated to each token
2. Linear layer
3. Activation layer: using the hardtanh function
4. Linear layer
5. Hinge loss layer

4 Experiments

We tested the use of discriminative word embeddings in the task of social sensing, i.e. of detecting specific signals from social media. In particular we explored the ability to monitor and alert about emergencies caused by natural disasters. We explored the corpus of Social Sensing[4], which consist of 5,642 tweets about natural catastrophic events like earthquakes or floods. To obtain a balanced training set, we combined this corpus with a set of generic tweets, consisting of 23,507 tweets. The combined corpus, consisting of 29,149 tweets, was randomly split into a training, development and test set consisting respectively of 23,850, 2,649 and 2,650 tweets.

4.1 Lexicon

Most sentiment analysis systems exploit a specialized lexicon (Rosenthal et al, 2014; Rosenthal et al, 2015). We built a lexicon of words related or indicative of disasters, by using the Italian Word Embeddings interface[5]. Starting from a seed set of few specialized words we produced a lexicon of 292 words (including words with a hashtag).

4.2 Classifier

For detecting tweets reporting about natural disasters, we exploit an SVM classifier, which uses as continuous features the word embeddings created from the text of the Italian Wikipedia. Additionally a set of discrete features is used, similar to those used in the top scoring system in the task 10 of SemEval 2014 on Sentiment Analysis in Twitter (Mohammad et al., 2014). These features are summarized in the following table:

Type	***Description***
allcaps	feature telling whether a word is all in uppercase
EmoPos	Presence of a positive emoticon
EmoNeg	Presence of a negative emoticon
Elongated	Presence of an elongated word
Lexicon count	Number of word present in a lexicon
Lexicon min	Lowest score of word in lexicon
Lexicon last	Score of the last word present in lexicon
Lexicon sum	Sum of the scores of words present in lexicon
Negation	Count of negative words
Elongated punct	Count of multiple punctuations (e.g. "!!!")
Ngrams	Ngrams of length 2-4

4.3 Results

We created generic word embeddings on the corpus consisting of the plain text extracted from the Italian Wikipedia, for a total of 1,096,243,235 tokens, 4,456,972 distinct.

We selected the 100,000 most frequent words and we created word embeddings for them, with a space dimension of 64.

The table below shows the results obtained with the discriminative word embeddings compared to a baseline obtained with the same classifier using the generic embeddings.

Data	**System**	**Precision**	**Recall**	**F1**
Develop	baseline	85.91	72.66	78.73
Develop	DE	**87.08**	**76.37**	**81.37**
Test	baseline	**86.87**	70.96	78.11
Test	DE	85.94	**75.05**	**80.12**

The results show a significant improvement in recall with respect to the baseline, which leads to over a 2-point improvement in F1.

4.4 Related Work

Social sensing research is a rapidly growing field; however, it is difficult to compare our work with others since the data sets used are different.

The only experiment performed on the same data set, is described in (Cresci et al., 2015), which focuses on distinguishing whether damage

[4] http://socialsensing.it/en/datasets
[5] http://tanl.di.unipi.it/embeddings/

was reported, rather than just reportig a disaster. Sixteen experiments were carried out, using four subsets of the corpus for training, corresponding to four disaster events, and testing on either different events (*cross-event*) or same/different disaster types (*in-domain*, *out-domain*). F1 scores in detecting non relevant tweets ranged between 19% and 28% for *cross-event* and *out-domain* and reached 73% for in-domain in one of the *in-domain* tests.

5 Conclusions

We have presented the notion of discriminative word embeddings that were designed to cope with semantic dissimilarity in tasks like sentiment analysis or multiclass classification.

As an example of the effectiveness of this type of embeddings in other applications, we have explored their use in detecting tweets reporting alerts or notices about natural disasters.

Our approach consisted in using a classifier trained on a corpus of annotated tweets, using discriminative embeddings as features, instead of the typical manually crafted features or dictionnaries employed in tweet classification tasks as sentiment analysis.

In the future, we plan to explore the use a convolutional network classifier, also provided by DeepNL, without any additional features, as Severyn and Moschitti (2015) have done for the SemEval 2015 task on Sentiment Analysis in Twitter.

References

R. Al-Rfou, B. Perozzi, and S. Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. arXiv preprint arXiv:1307.1662.

R. K. Ando, T. Zhang, and P. Bartlett. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Roi Blanco, Giuseppe Ottaviano, Edgar Meij, 2015. Fast and Space-efficient Entity Linking in Queries, ACM WSDM 2015.

D. Chen and C. D. Manning. 2014. Fast and Accurate Dependency Parser using Neural Networks. In: Proc. of EMNLP 2014.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, 2008.

R. Collobert. 2011. Deep Learning for Efficient Discriminative Parsing. In AISTATS, 2011.

R. Collobert et al. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, 2461–2505.

S. Cresci, M. Tesconi, A. Cimino and F. Dell'Orletta. 2015. A Linguistically-driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages. Proceedings of the 24th international conference companion on World Wide Web (WWW'15).

M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, N. Bhamidipati. 2015. Context- and Content-aware Embeddings for Query Rewriting in Sponsored Search. *Proceedings of SIGIR 2015*, Santiago, Chile.

Huang et al. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes, *Proc. of the Association for Computational Linguistics 2012 Conference*.

G.E. Hinton, J.L. McClelland, D.E. Rumelhart. Distributed representations. 1986. In *Parallel distributed processing: Explorations in the microstructure of cognition*. Volume 1: Foundations, MIT Press, 1986.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014. JMLR:W&CP volume 32.

Rémi Lebret and Ronan Collobert. 2013. Word Embeddings through Hellinger PCA. *Proc. of EACL 2013*.

Omer Levy and Yoav Goldberg. 2014. Neural Word Embeddings as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems* (NIPS), 2014.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press. Cambridge, Massachusetts.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets, In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), June 2013, Atlanta, USA.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko. 2014. Nrc-canada-2014: Recent improvements in sentiment analysis of tweets, and the Voted Perceptron. In Eighth International Workshop on Semantic Evaluation Exercises (SemEval-2014).

T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010*, 11th Annual Conference of the International

Speech Communication Association, Makuhari, Chiba, Japanfmikol.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*, 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, 2013.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. Computational Linguistics, 34(4):513-553.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, pp. 45–50.

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 '14, pages 73–80, Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. *Proc. of the ninth International Workshop on Semantic Evaluation* (SemEval-2105), Denver, USA.

Aliaksei Severyn, Alessandro Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation* (SemEval-2015), Denver, USA.

S. Srivastava, E. Hovy. 2014. Vector space semantics with frequency-driven motifs. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 634–643, Baltimore, Maryland, USA.

Tang et al. 2014. Learning Sentiment-SpecificWord Embedding for Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 1555–1565, Baltimore, Maryland, USA, June 23-25 2014.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistic*s, pp. 384-394. Association for Computational Linguistics.2013.

Evolution of Italian Treebank and Dependency Parsing towards Universal Dependencies

Giuseppe Attardi, Simone Saletti, Maria Simi
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo 3
56127 Pisa
{attardi,saletti,simi}@di.unipi.it

Abstract

English. We highlight the main changes recently undergone by the Italian Dependency Treebank in the transition to an extended and revised edition, compliant with the annotation schema of Universal Dependencies. We explore how these changes affect the accuracy of dependency parsers, performing comparative tests on various versions of the treebank. Despite significant changes in the annotation style, statistical parsers seem to cope well and mostly improve.

Italiano. *Illustriamo i principali cambiamenti effettuati sulla treebank a dipendenze per l'italiano nel passaggio a una versione estesa e rivista secondo lo stile di annotazione delle Universal Dependencies. Esploriamo come questi cambiamenti influenzano l'accuratezza dei parser a dipendenze, eseguendo test comparativi su diverse versioni della treebank. Nonostante i cambiamenti rilevanti nello stile di annotazione, i parser statistici sono in grado di adeguarsi e migliorare in accuratezza.*

1 Introduction

Universal Dependencies (UD) is a recent initiative to develop cross-linguistically consistent treebank annotations for several languages that aims to facilitate multilingual parser development and cross-language parsing (Nivre, 2015). An Italian corpus annotated according to the UD annotation scheme was recently released, as part of version 1.1 of the UD guidelines and resources. The UD-it v1.1 Italian treebank is the result of conversion from the ISDT (Italian Stanford Dependency Treebank), released for the shared task on dependency parsing of Evalita-2014 (Bosco et al., 2013 and 2014). ISDT is a resource annotated according to the Stanford dependencies scheme (de Marneffe et al. 2008, 2013a, 2013b), obtained through a semi-automatic conversion process starting from MIDT (the Merged Italian Dependency Treebank) (Bosco, Montemagni, Simi, 2012 and 2014). MIDT in turn was obtained by merging two existing Italian treebanks, differing both in corpus composition and adopted annotation schemes: TUT, the Turin University Treebank (Bosco et al. 2000), and ISST-TANL, first released as ISST-CoNLL for the CoNLL-2007 shared task (Montemagni and Simi, 2007).

UD can be considered as an evolution of the Stanford Dependencies into a multi-language framework and introduce significant annotation style novelties (deMarneffe et al., 2014). The UD schema is still evolving with many critical issues still under discussion, hence it is worthwhile to explore the impact of the proposed standard on parser performance, for example to assess whether alternative annotation choices might make parsing easier for statistically trained parsers.

For Italian we are in the position to compare results obtained in the Evalita 2014 DP parsing tasks with the performance of state-of-the-art parsers on UD, since both treebanks share a large subset of sentences.

Moreover, since UD is a larger resource than ISDT, we can also evaluate the impact of increasing the training set size on parser performance.

Our aim is to verify how differences in annotation schemes and in the corresponding training resources affect the accuracy of individual state-of-the-art parsers. Parser combinations, either

stacking or voting, can be quite effective in improving accuracy of individual parsers, as proved in the Evalita 2014 shared task and confirmed by our own experiments also on the UD. However our focus here lies in exploring the most effective single parser techniques for UD with respect to both accuracy and efficiency.

2 From ISDT to UD-it

In this section we highlight the changes in annotation guidelines and corpus composition between ISDT and UD-it.

2.1 Differences in annotation guidelines

The evolution of the Stanford Dependencies into a multi-language framework introduces two major changes (deMarneffe et al., 2014), concerning: (i) the treatment of copulas and (ii) the treatment of prepositions with case marking.

SD already recommended a treatment of the copula "to be" ("*essere*" in Italian) as dependent of a lexical predicate. In UD this becomes prescriptive and is motivated by the fact that many languages often lack an overt copula. This entails that the predicate complement is linked directly to its subject argument and the copula becomes a dependent of the predicate.

The second major change is the decision to fully adhere to the design principle of directly linking content words, and to abandon treating prepositions as a mediator between a modified word and its object: prepositions (but also other case-marking elements) are treated as dependents of the noun with specific *case* or *mark* labels.

The combined effect of these two decisions leads to parse trees with substantially different structure. Figure 1 and 2 show for instance the different parse trees, in passing from ISDT to UD annotations, for the sentence "È stata la giornata del doppio oro italiano ai Mondiali di atletica." [*It was the day of the Italian double gold at World Athletics Championships*.].

In fact exceptions to the general rule are still being discussed within the UD consortium, since the issue of copula inversion is somewhat controversial. In particular there are cases of prepositional predicates where the analysis with copula inversion leads to apparently counterintuitive situations. UD-it version 1.1 in particular does not implement copula inversion when the copula is followed by a prepositional predicate.

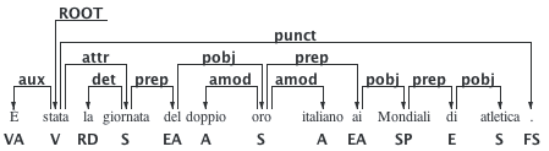


Figure 1. Example parse tree in ISDT

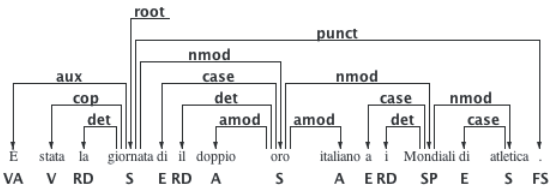


Figure 2. Example parse tree in UD1.1

Figure 3 illustrates the treatment advocated by strictly adhering to the UD guidelines, which is being considered for adoption in UD-it version 1.2. Notice that a quite different structure would be obtained for a very similar sentence like "La scultura appartiene al pachistano Hamad Butt" [*The sculpture belongs to the Pakistan Hamad Butt*].

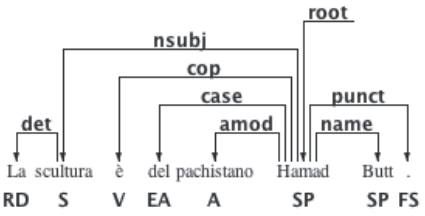


Figure 3. Example parse tree contemplated in UD 1.2

For the purpose of this presentation, we will call this version of the resource UD-it 1.2.[1]

Other changes in the annotation guidelines moving from ISDT and UD are less relevant for this discussion and involve the renaming of dependency labels, the introduction of special constructs for dealing with texts of a conversational nature (*discourse*, *vocative*) and the standardization of part-of-speech and morphological features.

2.2 Change of format

UD 1.1 also introduces an extension of the classical CoNLL-X tab separated format, called CoNNL-U. The main difference is the introduction of a notation for representing aggregated words (e.g. verbs with clitics or articulated prepositions): these can be split into their constituents and given as ID the range of the ID's of the constituents. An example from the guidelines is the following: "vámonos al mar" [*let's go to the sea*]:

1 By this we do not mean to imply that version 1.2 of UD-it, due in November 2015, will match exactly this conventions.

```
1-2     vámonos     _
1       vamos       ir
2       nos         nosotros
3-4     al          _
3       a           a
4       el          el
5       mar         mar
```

2.3 Corpus extension

The ISDT corpus released for Evalita 2014 consists of 97,500 tokens derived from the TUT and 81,000 tokens derived from the ISST-TANL. Moreover a gold test dataset of 9,442 tokens was produced for the shared task. UD-it is a larger resource including the previous texts (with converted annotations), a new corpus of questions, and data obtained from ParTUT[2] (the Multilingual Turin University Treebank) for a total of 324,406 tokens (13,079 sentences). For release 1.1, UD-it was randomly split into train, development and test data sets. Both development and test include 500 sentences each (~13,300 tokens).

3 Dependency parsers

We provide a short description of the state-of-the-art parsers chosen for our experiments.

DeSR was chosen as a representative of transition-based parsers for two main reasons, besides our own interest in developing this technology: given its declarative configuration mechanism it allows to experiment with different feature sets; other parsers in this category, in particular Maltparser (Nivre et al.), were consistently reported to provide inferior results in all Evalita evaluation campaigns for Italian.

3.1 DeSR

DESR MLP is a transition-based parser that uses a Multi-Layer Perceptron (Attardi 2006, Attardi et al., 2009a and 2009b). We trained it on 300 hidden variables, with a learning rate of 0.01, and early stopping when validation accuracy reaches 99.5%. The basic feature model used in the experiments on the Evalita training set is reported in Table 1.

The match expression indicates a feature to be extracted when a value matches a regular expression. Conditional features are used for representing linguistic constraints that apply to long distance dependencies. The feature used in the model takes into account a prepositional phrase (indicated by a dependent token with coarse POS of "E"), and it extracts a feature consisting of the pair: $b_0.l$ and the lemma of last preceding verb (a token whose POS is "V").

Single word features
$s_0.f\ b_0.f\ b_1.f$
$s_0.l\ b_0.l\ b_1.l\ b_0^{-1}.l\ lc(s_0).l\ rc(b_0).l$
$s_0.p\ b_0.p\ b_1.p\ rc(s_0).p\ rc(rc(b_0)).p$
$s_0.c\ s_0.c\ b_0.c\ b_1.c\ b_2.c\ b_3.c\ b_0^{-1}.c\ lc(s_0).c\ rc(b_0).c$
$s_0.m\ b_0.m\ b_1.m$
$lc(s_0).d\ lc(b_0).d\ rc(s_0).d$
$match(lc(b_0).m$, "Number=.")
$match(lc(b_0).m$, "Number=.")
Word pair features
$s_0.c\ b_0.c$
$b_0.c\ b_1.c$
$s_0.c\ b_1.c$
$s_0.c\ b_2.c$
$s_0.c\ b_3.c$
$rc(s_0).c\ b_0.c$
Conditional features
$if(lc(b_0).p$ = "E", $b_0.l)\ last$(POSTAG, "V")).l

Table 1. Feature templates: s_i represents tokens on the stack, b_i tokens on the input buffer. $lc(t)$ and $rc(t)$ denote the leftmost and rightmost child of token t, f denotes the form, l denotes the lemma, p and c the POS and coarse POS tag, m the morphology, d the dependency label. An exponent indicates a relative position in the input sentence.

Furthermore, an experimental feature was introduced, for adding a contribution from the score of the graph to the function of the MLP network. Besides the score computed by multiplying the probabilities of the transitions leading to a certain state, the score for the state reached for sentence x, after the sequence of transitions t, given the model parameters θ, is given by:

$$s(x, t, \theta) = \prod_{i=1}^{n} f_\theta(t_i) + E\left(x, t_1^i\right)$$

where $f_\theta(t)$ is the output computed by the neural network with parameters θ, and $E(x, t)$ is the score for the graph obtained after applying the sequence of transitions t to x. The graph score is computed from the following features:

Graph features
$b_0.l\ rc(b_0).p$
$b_0.l\ lc(b_0).p$
$b_0.l\ rc(b_0).p\ lc(rc(b_0)).p$
$b_0.l\ rc(b_0).p\ rc(rc(b_0)).p$
$b_0.l\ rc(b_0).p\ ls(rc(b_0)).p$
$lc(b_0).p\ b_0.l\ rc(b_0).p$
$b_0.l\ lc(b_0).p\ rc(lc(b_0)).p$
$b_0.l\ rc(b_0).p\ lc(lc(b_0)).p$
$b_0.l\ rc(b_0).p\ rs(lc(b_0)).p$
$rc(b_0).p\ b_0.l\ lc(b_0).p$

Table 2. A graph score is computed from these features. *ls* denotes the left sibling, *rs* the right sibling.

[2] http://www.di.unito.it/~tutreeb/partut.html

For the experiments on the UD corpus, the base feature model was used with 28 additional 3rd order features, of which we show a few in Table 3.

3rd order features
$s_0^{+1}.f\ b_0^{+2}.f\ b_0.p$
$s_0^{+2}.f\ b_0^{+3}.f\ b_0.p$
$s_0^{+2}.f\ b_0.f\ b_0.p$
$s_0^{+3}.f\ b_0^{+2}.f\ s_0.p$ ...

Table 3. Sample of 3rd order features used for UD corpus.

3.2 Turbo Parser

TurboParser (Martins et al., 2013) is a graph-based parser that uses third-order feature models and a specialized accelerated dual decomposition algorithm for making non-projective parsing computationally feasible (cite). TurboParser was used in configuration "full", enabling all third-order features.

3.3 MATE Parser

The Mate parser is a graph-based parser that uses passive aggressive perceptron and exploits reach features (Bohnet, 2010). The only configurable parameter is the number of iterations (set to 25).

The Mate tools also include a variant that is a combination of transition-based and graph-based dependency parsing (Bohnet and Kuhn, 2012). We tested also this version, which achieved, as expected, accuracies that are half way between a pure graph-based and a transition-based parser and therefore they are not reported in the following sections.

4 Experiments

4.1 Evalita results on ISDT

The table below lists the best results obtained by the three parsers considered, on the Evalita 2014 treebank. Training was done on the train plus development data set and testing on the official test data set.

Parser	LAS	UAS
DeSR	84.79	87.37
Turbo Parser	86.45	88.98
Mate	86.82	89.18

Table 4. Evalita 2014 ISDT dataset

The best official results were obtained using a preprocessing step of tree restructuring and performing parser combination: 87.89 LAS, 90.16 UAS (Attardi and Simi, 2014).

4.2 Evalita dataset in UD 1.1

Our first experiment is performed on the same dataset from Evalita 2014, present also in the official UD-it 1.1 resource. We report in Table 5 the performance of the same parsers.

Parser	LAS	UAS	Diff
DeSR	85.57	88.68	+0.78
Turbo Parser	87.07	90.06	+0.62
Mate	88.01	90.43	+1.19

Table 5. Evalita 2014 dataset, UD-it 1.1 conventions

Using the resource converted in UD, the LAS of all the three parsers improved, as shown in the Diff column. This was somehow not expected since the tree structure is characterized by longer distance dependencies.

In fact a basic tree combination of these three parsers achieves 89.18 LAS and 91.28 UAS, an improvement of +1.29 LAS over the best Evalita results on ISDT.

5 Training with additional data

As a next step we repeated the experiment using the additional data available in UD-it 1.1 for training (about 71,000 additional tokens).

Parser	LAS	UAS	Diff
DeSR	85.19	88.18	-0.38
Turbo Parser	87.42	90.25	0.35
Mate	88.25	90.54	0.24

Table 6. Evalita 2014 dataset with additional training data, UD-it 1.1 conventions

The added training data do not appear to produce a significant improvement (Table 6). This may be due to the fact that the new data were not fully compliant with the resource at the time of release of UD1.1. Column Diff shows the difference with respect to the LAS scores reported in 4.2.

5.1 Evalita dataset in UD 1.2

The experiment in section 4.2 was repeated with UD-it 1.2, the version where copula inversion is performed also in the case of prepositional arguments. Table 7 also reports the difference with the LAS scores in 4.2.

Parser	LAS	UAS	Diff
DeSR	85.97	88.52	0.40
Turbo Parser	87.93	90.64	0.86
Mate	88.55	90.66	0.54

Table 7. Evalita 2014 dataset, UD-it 1.2 conventions

5.2 UD-it 1.1 dataset

The next set of experiments was performed with official release of the UD-it 1.1. Tuning of DeSR was done on the development data and the best parser was used to obtain the following results on the test data (Table 8).

Parser	Devel		Test	
	LAS	UAS	LAS	UAS
DeSR	88.28	91.13	87.93	90.78
Turbo Parser	89.99	92.48	89.77	92.46
Mate	91.24	93.05	90.53	92.59

Table 8. UD-it 1.1 dataset, partial copula inversion

5.3 UD-it 1.2 dataset

For completeness, we repeated the experiments with the UD-it 1.2 dataset (same data of UD-it 1.1, but complete copula inversion), obtaining even better results (Table 9).

Parser	Devel		Test	
	LAS	UAS	LAS	UAS
DeSR	89.09	91.40	89.02	90.39
Turbo Parser	89.54	92.10	89.40	92.17
Mate	90.81	92.70	90.22	92.47

Table 9. UD-it 1.1 dataset, complete copula inversion

5.4 Parser efficiency

Concerning parser efficiency, we measured the average parsing time to analyze the test set (500 sentences), employed by the three parsers under the same conditions. This also means that for MATE we deactivated the multicore option and used only one core. The results are as follows:

- DeSR: 18 seconds
- TurboParser: 47 seconds
- Mate: 2 minutes and 53 seconds

6 Conclusions

We have analyzed the effects on parsing accuracy throughout the evolution of the Italian treebank, from the version used in Evalita 2014 to the new extended and revised version released according to the UD framework.

General improvements have been noted with all parsers we tested: all of them seem to cope well with the inversion of direction of prepositional complements and copulas in the UD annotation. Improvements may be due as well to the harmonization effort at the level of PoS and morpho-features carried out in the process.

Graph based parsers still achieve higher accuracy, but the difference with respect to a transition based parser drops when third order features are used. A transition-based parser still has an advantage in raw parsing speed (i.e. disregarding speed-ups due to multithreading) and is competitive for large scale applications.

References

Giuseppe Attardi. 2006. Experiments with a Multilanguage Non-Projective Dependency Parser, Proc. of the Tenth Conference on Natural Language Learning, New York, (NY).

Giuseppe Attardi, Felice Dell'Orletta. 2009. Reverse Revision and Linear Tree Combination for Dependency Parsing. In: *Proc. of Human Language Technologies: The 2009 Annual Conference of the NAACL, Companion Volume: Short Papers*, 261–264. ACL, Stroudsburg, PA, USA.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Joseph Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In: *Proc. of Workshop Evalita 2009*, ISBN 978-88-903581-1-1.

Giuseppe Attardi, Maria Simi, 2014. Dependency Parsing Techniques for Information Extraction, Proceedings of Evalita 2014.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proc. of Coling 2010*, pp. 89–97, Beijing, China. Coling 2010 Organizing Committee.

Bernd Bohnet and Jonas Kuhn. 2012. The Best of BothWorlds -- A Graph-based Completion Model for Transition-based Parsers. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pages 77–87.

Cristina Bosco, Vincenzo Lombardo, Leonardo Lesmo, Daniela Vassallo. 2000. Building a treebank for Italian: a data-driven annotation schema. In Proceedings of LREC 2000, Athens, Greece.

Cristina Bosco, Simonetta Montemagni, Maria Simi. 2012. Harmonization and Merging of two Italian Dependency Treebanks, Workshop on Merging of Language Resources, in Proceedings of LREC 2012, Workshop on Language Resource Merging, Instanbul, May 2012, ELRA, pp. 23–30.

Cristina Bosco, Simonetta Montemagni, Maria Simi. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In: *ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, Maria Simi. 2014. The Evalita 2014 Dependency Parsing task, CLiC-it 2014 and EVALITA 2014 Proceedings, Pisa University Press, ISBN/EAN: 978-886741-472-7, 1–8.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In COLING Workshop on Cross-framework and Cross-domain Parser Evaluation.

Marie-Catherine de Marneffe, Miriam Connor, Natalia Silveira, Bowman S. R., Timothy Dozat, Christopher D. Manning. 2013. More constructions, more genres: Extending Stanford Dependencies, Proc. of the Second International Conference on Dependency Linguistics (DepLing 2013), Prague, August 27–30, Charles University in Prague, Matfyzpress, Prague, pp. 187–196.

Marie-Catherine de Marneffe and Christopher D. Manning. 2013. Stanford typed dependencies manual, September 2008, Revised for the Stanford Parser v. 3.3 in December 2013.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning. 2014. Universal Stanford Dependencies: a Cross-Linguistic Typology. In: *Proc. LREC 2014*, Reykjavik, Iceland, ELRA.

Andre Martins, Miguel Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In: *Proc. of the 51st Annual Meeting of the ACL* (Volume 2: Short Papers), 617–622, Sofia, Bulgaria. ACL.

Simonetta Montemagni, Maria Simi. 2007. The Italian dependency annotated corpus developed for the CoNLL–2007 shared task. Technical report, ILC–CNR.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: a data-driven parser-generator for dependency parsing. In Proceedings of LREC-2006, volume 2216–2219.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing, CICLing (1) 2015: 3–16

Maria Simi, Cristina Bosco, Simonetta Montemagni. 2008. Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In: *Proc. LREC 2014*, 26–31, May, Reykjavik, Iceland, ELRA.

CItA: un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1 Annotato con Errori

A. Barbagli•, P. Lucisano•, F. Dell'Orletta◇, S. Montemagni◇, G. Venturi◇
•Dipartimento di Psicologia dei processi di Sviluppo e socializzazione,
Università di Roma "La Sapienza"
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
alessia.barbagli@gmail.com, pietro.lucisano@uniroma1.it,
{felice.dellorletta,simonetta.montemagni,giulia.venturi}@ilc.cnr.it

Abstract

English. In this paper we present CItA the first corpus of written essays by Italian L1 learners in the first and second year of lower secondary school. CItA was annotated with grammatical, orthographic and lexical errors. The corpus peculiarities and its diachronic nature make it particularly suitable for computational linguistics applications and socio–pedagogical studies.

Italiano. *In questo articolo presentiamo CItA il primo corpus di produzioni scritte di apprendenti l'italiano L1 del primo e del secondo anno della scuola secondaria di primo grado annotato con errori grammaticali, ortografici e lessicali. Le specificità del corpus e la sua natura diacronica lo rendono particolarmente utile sia per applicazioni linguistico-computazionali sia per studi socio-pedagogici.*

1 Introduzione

La costruzione di corpora di produzioni di apprendenti è da sempre al centro delle attività di ricerca della comunità di linguistica computazionale. Un'attenzione particolare è dedicata all'annotazione e classificazione degli errori commessi dagli apprendenti. Corpora annotati con questo tipo di informazione vengono usati tipicamente per lo studio e la creazione di modelli di sviluppo delle abilità di scrittura (Deane and Quinlan, 2010) e per lo sviluppo di sistemi a supporto dell'insegnamento (i cosiddetti *Intelligent Computer–Assisted Language Learning systems*) (Granger, 2003). In questo scenario, un interesse particolare è dedicato alla raccolta e annotazione di corpora di produzioni di apprendenti L2 impiegati come punto di partenza per studi sullo sviluppo dell'interlingua, per attività di riflessione sull'eventuale modifica e/o personalizzazione dell'azione didattica dell'insegnante e per lo sviluppo di sistemi di correzione automatica degli errori. La maggior parte delle attività ha riguardato la costruzione di corpora di apprendenti l'inglese L2, tra cui il più recente e il più ampio è il *NUS Corpus of Learner English (NUCLE)* (Dahlmeier et al., 2013), utilizzato come risorsa di riferimento nel 2013 e 2014 dello "Shared Task on Grammatical Error Correction" (Ng et al., 2013; Ng et al., 2014). Tuttavia, in questi ultimi anni, l'attenzione è stata anche rivolta a L2 diverse dall'inglese, quali ad esempio l'arabo (Zaghouani et al., 2015), il tedesco (Ludeling et al., 2005), l'ungherese (Dickinson and Ledbetter, 2012), il basco (Aldabe et al., 2005) e il ceco e l'italiano (Andorno and Rastelli, 2009; Boyd et al., 2014). Minore attenzione è stata invece dedicata alla costruzione di risorse costituite da produzioni di apprendenti L1. Un'eccezione è rappresentata dal *KoKo* corpus (Abel et al., 2014), collezione di produzioni di apprendenti tedesco L1 dell'ultimo anno della scuola secondaria di secondo grado arricchite con informazioni di sfondo degli apprendenti (es. età, genere, situazione socio–economica), annotazione manuale di errori ortografici e grammaticali, e informazione linguistica annotata in maniera automatica.

Collocandoci in quest'ultimo scenario, in questo articolo presentiamo CItA (*Corpus Italiano di Apprendenti L1*) il primo corpus di produzioni scritte di apprendenti l'italiano L1 annotato manualmente con diverse tipologie di errori e con la relativa correzione. Il corpus, composto da produzioni dei primi due anni della scuola secondaria di primo grado, è a nostra conoscenza non solo il primo corpus italiano di questo tipo ma contiene delle caratteristiche di novità che lo rendono unico anche all'interno del panorama internazionale di ricerca.

2 Corpus

Il punto di partenza per la creazione di CItA è rappresentato dalle trascrizioni delle produzioni scritte di studenti di sette diverse scuole secondarie di primo grado di Roma descritte da Barbagli et al. (2014). Le scuole considerate sono rappresentative di un ambiente socio–culturale che può essere definito medio–alto (il centro) e di uno medio–basso (la periferia). Per ogni scuola è stata individuata una classe, per un totale di 77 studenti in centro e 79 in periferia e per ogni studente sono state raccolte due tipologie di produzioni scritte: le tracce assegnate indipendentemente da ogni docente durante l'anno e due prove comuni a tutte le scuole. Il corpus, composto da 1.352 testi (366.335 tokens), comprende i testi prodotti da ogni studente durante il suo primo e secondo anno scolastico. Il corpus è accompagnato da un questionario che raccoglie alcune variabili di sfondo di ogni studente come ad esempio il background familiare (es. il lavoro e il titolo di studio dei genitori), territoriale (zona della scuola), personale (es. numero di libri letti).

Le principale novità di CItA riguardano il tipo di produzioni considerate (quelle di apprendenti di italiano L1), l'annotazione degli errori e l'ordinamento temporale delle produzioni all'interno di due anni scolastici consecutivi. Queste caratteristiche permettono di condurre uno studio sulle variazioni delle frequenze e delle tipologie di errori commessi al mutamento delle competenze linguistiche di ogni studente sia all'interno di uno stesso anno scolastico sia al passaggio dal primo e al secondo anno della scuola secondaria di primo grado. CItA rende inoltre possibile studiare le relazioni tra le variazioni di errori e le variabili di sfondo contenute nel questionario. L'attenzione posta sulla scuola secondaria di primo grado rappresenta un'ulteriore aspetto innovativo. Il primo biennio della scuola media è stato sino ad oggi poco indagato dalle ricerche empiriche nonostante sia un momento cardine nello sviluppo delle abilità linguistiche di uno studente.

3 Schema di Annotazione

La definizione dello schema di annotazione degli errori qui presentato si inserisce nel più ampio contesto degli studi condotti in ambito italiano sulla valutazione delle abilità linguistiche nella lingua materna (Corda Costa and Visalberghi, 1995; De Mauro, 1983; GISCEL, 2010; Colombo, 2011). Siccome l'attribuzione di errore ad una forma linguistica è un'operazione delicata poiché si presuppone il riferimento ad un sistema normativo, che di per sé non è oggettivo ma arbitrario, poiché basato su convenzioni sociali, per individuare gli errori abbiamo fatto riferimento al concetto di italiano standard neostandard individuato da Berruto (1987). L'analisi empirica della distribuzione delle varie tipologie di errori in CItA è stato un altro dei criteri adottati nel definire lo schema di annotazione. Sulla base di queste considerazioni, abbiamo scelto di annotare le tipologie di errori a cui si fa tradizionalmente riferimento in letteratura laddove la frequenza di occorrenza nel corpus fosse significativa. Come mostra la Tabella 1, che riporta lo schema di annotazione e le distribuzioni delle diverse categorie di errore considerate, abbiamo scelto di annotare errori riconducibili a tre macro–aree: grammatica, ortografia e lessico. Come indicato anche nel recente Rapporto sulla "Rilevazione degli errori più diffusi nella padronanza della lingua italiana nella prima prova di italiano"[1] redatto nel 2012 dall'INVALSI e dall'Accademia della Crusca, sono queste tre gli ambiti di competenza linguistica rispetto ai quali è possibile valutare la padronanza linguistica di uno studente. Seguendo la ripartizione suggerita dal Rapporto in descrittori specifici, per ciascuna competenza è stata prevista una categoria di errore corrispondente alla categoria morfosintattica coinvolta (colonna *Categoria* della Tabella 1). Inoltre, adottando la strategia suggerita da Granger (2003), per ogni categoria è stato individuato il tipo di modifica proposta per l'errore (colonna *Tipo di modifica*). Il formato di annotazione scelto è ispirato a quello messo a punto in occasione dello "Shared Task on Grammatical Error Correction" 2013. La frase seguente mostra un esempio estratto dal corpus dove sono stati annotati due errori:

> [...] io <M t="3.1" c="dovevo">avevo a</M> salire fin lassù ma mi sono <M t="2.1" c="fatta">fata</M> coraggio [...]

Il tag *<M>* (*Mistake*) e la sua relativa chiusura *</M>* marcano l'area dell'errore annotato. *<M>* ha due attributi: *t* (*type*) il cui valore corrisponde al codice dell'errore e *c* (*correction*) il cui valore è la correzione dell'errore. In questo caso sono stati annotati due errori: un errore d'uso lessica-

[1] http://www.invalsi.it/download/rapporti/es2_0312/RAPPORTO_ITALIANO_prove_2010.pdf

		I anno			II anno			Totale %
Categoria	**Tipo di modifica**	Freq.%	Media	Dev.	Freq.%	Media	Dev.	
Grammatica								
Verbi	**Uso dei tempi**	7,78 (150)	0,99	2,29	15,67 (239)	1,47	4,05	11,26 (389)
	Uso dei modi	4,25 (82)	0,54	1,39	4,92 (75)	0,49	0,99	4,55 (157)
	Concordanza con il soggetto	2,85 (55)	0,37	1,38	4 (61)	0,41	1,27	3,36 (116)
Preposizioni	**Uso errato**	6,48 (125)	0,83	2,58	6,75 (103)	0,66	1,21	6,6 (228)
	Omissione o eccesso	1,03 (20)	0,13	0,40	0,72 (11)	0,07	0,25	0,90 (31)
Pronomi	Uso errato	5,09 (98)	0,65	1,13	3,54 (54)	0,36	0,97	4,4 (152)
	Omissione	0,41 (8)	0,05	0,36	0,59 (9)	0,06	0,39	0,49 (17)
	Eccesso	2,70 (52)	0,35	0,61	1,57 (24)	0,16	0,46	2,2 (76)
	Uso errato del pronome relativo	2,13 (41)	0,27	0,70	1,70 (26)	0,17	0,44	1,94 (67)
Articoli	**Uso errato**	5,81 (112)	0,75	3,72	3,54 (54)	0,35	1,09	4,81 (166)
Congiunzioni e/o connettivi	Uso errato	0,57 (11)	0,07	0,33	0,52 (8)	0,05	0,23	0,55 (19)
Altro		7,31 (141)	0,94	3,66	5,18 (79)	0,49	1,79	6,37 (220)
Ortografia								
Doppie	**Uso per difetto**	6,74 (130)	0,83	2,49	5,05 (77)	0,48	1,56	5,99 (207)
	Eccesso	3,27 (63)	0,42	0,89	3,67 (56)	0,37	1,13	3,45 (119)
Uso dell'*h*	**Per difetto**	3,21 (62)	0,39	1,03	1,64 (25)	0,17	0,62	2,52 (87)
	Per eccesso	1,66 (32)	0,21		1,11 (17)	0,10		1,42 (49)
Monosillabi	**Uso errato dei monosillabi accentati**	4,87 (94)	0,63	1,07	4,07 (62)	0,40	0,83	4,52 (156)
	Uso di *po* o *pò* anziché *po'*	1,66 (32)	0,21	0,72	1,64 (25)	0,17	0,52	1,65 (57)
Apostrofo	**Uso errato**	4,82 (93)	0,61	1,01	4,52 (69)	0,46	0,89	4,69 (162)
Altro		21,77 (420)	2,76	4,58	23,02 (351)	2,27	4,60	22,32 (771)
Lessico								
Vocabolario	**Uso errato**	5,60 (108)	0,70	1,64	6,56 (100)	0,66	1,09	6,02 (208)
Numero totale di errori		1929			1525			

Tabella 1: Schema di annotazione degli errori. Per ogni anno scolastico sono riportati: distribuzione percentuale degli errore e numero di occorrenze (*Freq.%*), occorrenza media degli errori per anno (*Media*), deviazione standard delle medie (*Dev.*). La colonna *Totale %* riporta la percentuale e il numero di occorrenze degli errori nei due anni. Gli errori che variano tra i due anni in modo statisticamente significativo all'analisi della varianza ($p<0.05$) sono stati marcati in grassetto.

le (*t="3.1"*) e un errore ortografico nell'uso per difetto delle doppie (*t="2.1"*).

In quanto segue riportiamo alcuni esempi di annotazione estratti da CItA che esemplificano alcune categorie di errori e le relative correzioni.

Verbi: uso dei tempi. [...] dopo aver fatto le squadre <M t="11" c="abbiamo">avevamo</M> subito iniziato a giocare [...]

Verbi: uso dei modi. [...] il pensiero che mi tormentava di più era che tra poco si <M t="12" c="sarebbe fatto">faceva</M> il campo scuola.

Verbi: concordanza con il soggetto. [...] la mia famiglia ed io <M t="13" c="stavamo">stavo</M> al mare a Torvajanica

Preposizioni: uso errato. <M t="14" c="in">a</M> Romania sono andata <M t="14" c="in">a</M> agosto [...]

Pronomi: uso errato. Proteggere i più deboli è molto coraggioso da parte di chi <M t="16" c="li">lo</M> protegge [...]

Pronomi: eccesso. Alla nostra maestra <M t="18" c="canc">gli</M> piaceva tanto la storia

Pronomi: uso errato del pronome relativo. La scienza non so perché mi fa pensare a un fenomeno costruito su un'altura <M t="19" c="per cui">che</M> ci vuole molto ingegno.

Articolo: uso errato. <M t="111" c="gli">i</M> dei, sapendo che qualcuno aveva preso senza merito il sacro vaso della Giustizia, si rattristarono molto, [...]

Grammatica: altro. Quando vedo <M t="10" c="quel">quelle</M> genere di <M t="10" c="persone">persona</M> mi sento strano.

Vocabolario: uso errato. C'era molta ombra nel giardino e io mi ci <M t="31" c="addormentavo">addormivo</M> sempre.

4 CItA per

Il corpus così annotato può avere diversi tipi di utilizzi. Dal punto di vista applicativo, CItA può essere usato come corpus di riferimento per sviluppare sistemi di identificazione e correzione automatica degli errori per la lingua italiana e/o per costruire modelli predittivi della competenza linguistica di un apprendente L1 (Richter et al., 2015). In quest'ultima direzione vanno gli studi che possono essere condotti confrontando le variazioni degli errori nel passaggio dal primo al secondo anno con i risultati del processo di monitoraggio delle caratteristiche linguistiche estratte dai testi linguisticamente annotati in modo automatico (Barbagli et al., 2014). Un esempio è quello della correlazione statisticamente significativa tra la distribuzione dei pronomi e il loro uso errato che diminuisce tra il primo e il secondo anno. Diminuisce ad esempio l'uso di pronomi personali e clitici, che sono usati in eccesso al primo anno, mentre rimane invariato l'uso di pronomi relativi ma cala la percentuale di errori che li coinvolgono. Il rapporto tra uso dei pronomi e relativi errori risulta pertanto predittivo dell'evoluzione nella competenza d'uso di questa categoria morfosintattica.

CItA può essere utilizzato per monitorare l'evoluzione degli errori nel tempo. La Tabella 1 riporta la distribuzione degli errori sia globalmente sia in ognuno dei due anni scolastici. Analizzando gli errori al passaggio tra primo e secondo anno, si può notare come la distribuzione di quelli ortografici e grammaticali (colonna *Totale* %) sia molto simile (rispettivamente 46,55% e 47,33%) mentre quelli lessicali sono nettamente meno (circa il 6%). Andando a valutare i singoli errori, quelli più frequenti sono quelli ortografici non classificati (22,32%) seguiti dall'uso errato dei tempi verbali (circa la metà dei precedenti), gli errori grammaticali non sottocategorizzati e l'uso errato delle preposizioni. Quando valutiamo la significatività delle variazioni degli errori tra i due anni, vediamo che quasi tutti (quelli marcati in grassetto) variano in maniera significativa, mostrando che esistono delle forti tendenze comuni nel passaggio dal primo al secondo anno. Studiando le distribuzioni di frequenza in modo separato per i due anni (colonna *Freq.%*) e la distribuzione media di ogni errore per anno (colonna *Media*), gli errori più diffusi sono quelli ortografici e grammaticali non sottocategorizzati, l'uso errato dei verbi, delle preposizioni, degli articoli, dei pronomi e l'uso per difetto delle doppie. Sebbene in generale il numero totale degli errori diminuisca nel passaggio dal primo al secondo anno, indagando come variano le distribuzioni di ogni categoria, scopriamo che non tutti i tipi di errore diminuiscono. Il caso più evidente è quello dell'aumento nel secondo anno dell'uso errato dei verbi in generale e dei tempi verbali in particolare. Ciò potrebbe essere riconducibile sia all'evoluzione dello studente sia al diverso tipo di tracce distribuite in classe dai docenti. Mentre nel primo anno le tracce assegnate sono per lo più di tipo narrativo, tipologia testuale che comporta l'uso di una sequenza temporale che potrebbe essere considerata più semplice da riconoscere e da costruire per gli studenti, al secondo anno aumentano le tracce di tipo argomentativo la cui struttura risulta più complessa. Questo ci porta a ipotizzare che gli studenti del secondo anno tentino di utilizzare forme verbali più complesse commettendo più errori. Questo è avvalorato dai risultati del monitoraggio linguistico automatico che rivelano come gli studenti al secondo anno usino strutture verbali più complesse (es. uso di ausiliari in tempi composti).

CItA può inoltre essere utilizzato all'interno di studi socio–pedagogici permettendo di mettere in relazione la distribuzione degli errori con le variabili di sfondo. È possibile così verificare in che misura i cambiamenti che avvengono nella scrittura sono attribuibili a condizioni socio–economiche di sfondo. È ad esempio interessante osservare come le esplorazioni statistiche condotte hanno rivelato che la diminuzione dell'uso errato del lessico dal primo al secondo anno è significativamente correlata con l'abitudine alla lettura. Oppure si può studiare come gli errori grammaticali variano in modo statisticamente significativo rispetto alla collocazione della scuola in centro o periferia di Roma: mentre nelle scuole del centro gli errori diminuiscono nel passaggio dal primo al secondo anno, in due delle quattro scuole in periferia aumentano. Diverso è il caso degli errori ortografici che non variano in modo statisticamente significativo rispetto alle variabili di sfondo considerate. Ciò confermerebbe alcuni studi (Colombo, 2011; Ferreri, 1971; Lavino, 1975; De Mauro, 1977) dove si afferma che la correttezza ortografica è un'abilità che si acquisisce con il tempo poiché richiede la sedimentazione di norme, spesso arbitrarie, che stabiliscono legami non causali tra suono e grafia.

References

A. Abel, A. Glaznieks, L. Nicolas, and E. Stemle. 2014. KoKo: an L1 Learner Corpus for German. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 26–31.

I. Aldabe, L. Amoros, B. Arrieta, A. Díaz de Ilarraza, M. Maritxalar, M. Oronoz, L. Uria. 2005. Learner and Error Corpora Based Computational Systems. *Proceedings of the PALC 2005 Conference*, Poland.

C. Andorno and S. Rastelli. 2009. *Corpora di Italiano L2: tecnologie, metodi, spunti teorici*. Guerra Edizioni.

A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2014. Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado. *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it)*, 9–10 December, Pisa, Italy.

G. Berruto 1987. Sociolinguistica dell'italiano contemporaneo. Carocci, Roma.

A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová, and C. Vettori. 2014. The MERLIN corpus: Learner Language and the CEFR. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

A. Colombo. 2011. *"A me mi" Dubbi, errori, correzioni nell'italiano scritto*. Franco Angeli editore.

M. Corda Costa and A. Visalberghi. 1995. (a cura di) *Misurare e valutare le competenze linguistiche*. La Nuova Italia, Firenze.

D. Dahlmeier, H.T. Ng, and S.M. Wu. 2013. Building a large annotated corpus of learner English: The NUS Corpus of Learner English. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31.

P. Deane and T. Quinlan. 2010. What automated analyses of corpora can tell us about student's writing skills. *Journal of Writing Research*, 2(2):151–177.

T. De Mauro. 1983. Per una nuova alfabetizzazione. Gensini S., Vedovelli M.(a cura di) *Teoria e pratica del glotto-kit. Una carta d'identità per l'educazione linguistica*. Franco Angeli Milano.

T. De Mauro. 1977. *Scuola e linguaggio*. Editori Riuniti, Roma.

M. Dickinson and S. Ledbetter. 2012. Annotating Errors in a Hungarian Learner Corpus. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

S. Ferreri. 1971. Italiano standard, italiano regionale e dialetto in una scuola media di Palermo. Medici M.-Simone R. (a cura di) *L'insegnamento dell'italiano in Italia e all'estero*, I, Roma, Bulzoni, 1971, pp. 205-224.

GISCEL Emilia-Romagna. 2010. La correzione dei testi scritti. Lugarini E. (a cura di) *Valutare le competenze linguistiche*. Franco Angeli Milano, pp. 188-203

S. Granger. 2003. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, 20:465–480.

C. Lavino. 1975 *L'insegnamento dell'italiano. Un'inchiesta campione in una scuola media sarda*. Edes, Cagliari.

A. Lüdeling, M. Walter, E. Kroymann, and P. Adolphs. 2005. Multi–level error annotation in learner corpora. *Proceedings of Corpus Linguistics 2005*.

H.T. Ng, S.M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–12.

H.T. Ng, S.M. Wu, T. Briscoe, C. Hadiwinoto, R.H. Susanto, and C. Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14.

S. Richter, A. Cimino, F. Dell'Orletta, and G. Venturi. 2015. Tracking the Evolution of Language Competence: an NLP–based Approach. *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC-it)*, 2–3 December, Trento, Italy.

W. Zaghouani, N. Habash, H. Bouamor, A. Rozovskaya, B. Mohit, A. Heider, K. Oflazer. 2015. Correction Annotation for Non-Native Arabic Texts: Guidelines and Corpus. *Proceedings of The 9th Linguistic Annotation Workshop*, pp. 129–139.

Entity Linking for Italian Tweets

Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro
Dept. of Computer Science - University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{firstname.lastname}@uniba.it

Abstract

English. Linking entity mentions in Italian tweets to concepts in a knowledge base is a challenging task, due to the short and noisy nature of these short messages and the lack of specific resources for Italian. This paper proposes an adaptation of a general purpose Named Entity Linking algorithm, which exploits the similarity measure computed over a Distributional Semantic Model, in the context of Italian tweets. In order to evaluate the proposed algorithm, we introduce a new dataset of tweets for entity linking that we developed specifically for the Italian language.

Italiano. *La creazione di collegamenti tra le menzioni di un'entità in tweet in italiano ed il corrispettivo concetto in una base di conoscenza è un compito problematico a causa del testo nei tweet, generalmente corto e rumoroso, e della mancanza di risorse specifiche per l'italiano. In questo studio proponiamo l'adattamento di un algoritmo generico di Named Entity Linking, che sfrutta la misura di similarità semantica calcolata su uno spazio distribuzionale, nel contesto dei tweet in Italiano. Al fine di valutare l'algoritmo proposto, inoltre, introduciamo un nuovo dataset di tweet per il task di entity linking specifico per la lingua italiana.*

1 Introduction

In this paper we address the problem of entity linking for Italian tweets. Named Entity Linking (NEL) is the task of annotating entity mentions in a portion of text with links to a knowledge base. This task usually requires as first step the recognition of portions of text that refer to named entities (*entity recognition*). The linking phase follows, which usually subsumes the entity disambiguation, i.e. selecting the proper concept from a restricted set of candidates (e.g. New York city or New York state). NEL together with Word Sense Disambiguation, i.e. the task of associating each word occurrence with its proper meaning given a sense inventory, is critical to enable automatic systems to make sense of unstructured text.

Initially developed for reasonably long and clean text, such as news articles, NEL techniques usually show unsatisfying performance on noisy, short and poorly written text constituted by microblogs such as Twitter. These difficulties notwithstanding, with an average of 500 billion posts being generated every day[1], tweets represent a rich source of information. Twitter-based tasks like user interest discovery, tweet recommendation, social/economical analysis, and so forth, could benefit from such a kind of semantic features represented by named entities linked to a knowledge base. Such tasks become even more problematic when the tweet analysis involves languages different from English. Specifically, in the context of Italian language, the lack of language-specific resources and annotated tweet datasets complicates the assessment of NEL algorithms for tweets.

Our main contributions to this problem are:

- An adaptation of a Twitter-based NEL algorithm based on a Distributional Semantic Model (DSM-TEL), which needs no specific Italian resources since it is completely unsupervised (Section 3).
- An Italian dataset of manually annotated tweets for NEL. To the best of our knowledge, this is the first Italian dataset of such a type. Section 2 reports details concerning the annotation phase and statistics about the

[1] http://www.internetlivestats.com/twitter-statistics/

dataset.

- An evaluation of well known NEL algorithms available for Italian language on this dataset, comparing their performance with our DSM-TEL algorithm in terms of both entity recognition and linking. Section 4 shows and analyses the results of that evaluation.

2 Dataset

One of the main limitations to the development of specific algorithms for tweet-based entity linking in Italian lies on the dearth of datasets for training and assessing the quality of such techniques. Hence, we built a new dataset by following the guidelines proposed in the #Microposts2015 Named Entity Linking (NEEL) challenge[2] (Rizzo et al., 2015). We randomly selected 1,000 tweets from the TWITA dataset (Basile and Nissim, 2013), the first corpus of Italian tweets. For each tweet, we first select the named entities (NE). A NE is a string in the tweet representing a proper noun, excluding the preceding article (like "il", "lo", "la", etc.) and any other prefix (e.g. "Dott.", "Prof.") or post-posed modifier. More specifically, an entity is any proper noun that: 1) belongs to one of the categories specified in a taxonomy and/or 2) can be linked to a DBpedia concept. This means that some concepts have a NIL DBpedia reference; these concepts belong to one of the categories but they have no corresponding concept in DBpedia. The taxonomy is defined by the following categories: Thing[3], Event, Character, Location, Organization, Person and Product.

We annotated concepts by using the canonicalized dataset of Italian DBpedia 2014[4]. For specific Italian concepts that are not linked to an English article, we adopt the localized version of DBpedia. Finally, some concepts have an Italian Wikipedia article but they are not in DBpedia; in that case we linked the entity by using the Wikipedia URL. Entities represented neither in DBpedia nor Wikipedia are linked to NIL.

The annotation process poses some challenges specific to the Twitter context. For example, entities can be part of a user mention or tag; all these strings are valid entities: #[Alemanno], and @[CarlottaFerlito]. The '#' and '@' characters are not considered as part of the annotation.

This first version of the dataset was annotated by only one annotator, and comprises 756 entity mentions, with a mean of about 0.75 entities for each tweet. The distribution of entities in categories is as follows: 301 Persons, 197 Organizations, 124 Locations, 96 Products, 18 Things, 11 Events and 9 Characters. 63% of tweets links to a DBpedia concept, about 30% of them is annotated as NIL, 6% links to an URL of a Wikipedia page, while only one entity links to an Italian DBpedia concept.

The dataset[5] is composed of two files: the data and the annotation file. The data file contains pairs of tweet id and text, each listed on a different line. The annotation file consists of a line for each tweet id, which is followed by the start and the end offset[6] of the annotation, the linked concept and the category. All values are separated by the TAB character. For example, for the tweet: *"290460612549545984 @CarlottaFerlito io non ho la forza di alzarmi e prendere il libro! Help me"*, we have the annotation: *"290460612549545984 1 16 http://dbpedia.org/resource/Carlotta˙Ferlito Person"*.

3 DSM-TEL algorithm

We propose an Entity Linking algorithm specific for Italian tweets that adapts the original method proposed during the NEEL challenge (Basile et al., 2015b). Our algorithms consists of two-steps: the initial identification of all possible entity mentions in a tweet followed by the linking of the entities through the disambiguation algorithm. We exploit DBpedia/Wikipedia twice in order to (1) extract all the possible surface forms related to entities, and (2) retrieve glosses used in the disambiguation process. In this case we use as gloss the extended abstract assigned to each DBpedia concept. To speed up the recognition of entities we build an index where each surface form (entity) is paired with the set of all its possible DBpedia concepts. The surface forms are collected by analysing all the internal links in the Italian Wikipedia dump. Each internal link reports the surface form and the linked Wikipedia page that corresponds to a DB-

[2] http://www.scc.lancs.ac.uk/research/workshops/microposts2015/challenge/

[3] Languages, ethnic groups, nationalities, religions, ...

[4] This dataset contains triples extracted from Italian Wikipedia articles whose resources have an equivalent English article.

[5] Available at: https://github.com/swapUniba/neel-it-twitter

[6] Starting from 0.

pedia resource. The index is built by exploiting the Lucene API[7]. Specifically for each possible surface form a document composed of two fields is created. The first field stores the surface form, while the second one contains the list of all possible DBpedia concepts that refer to the surface form in the first field. The entity recognition module exploits this index in order to find entities in a tweet. Given a tweet, the module performs the following steps:

1. Tokenization of the tweet using the Tweet NLP API[8]. We perform some pre-processing operations to manage hashtags and user mentions; for example we split tokens by exploiting upper-case characters: "CarlottaFerlito" $\Longrightarrow$ "Carlotta Ferlito";
2. Construction of a list of candidate entities by exploiting all n-grams up to six words;
3. Query of the index and retrieval of the top 100 matching surface forms for each candidate entity;
4. Scoring each surface form as the linear combination of: a) a string similarity function based on the Levenshtein Distance between the candidate entity and the surface form in the index; b) the Jaccard Index in terms of common words between the candidate entity and the surface form in the index;
5. Filtering the candidate entities recognized in the previous steps: entities are removed if the score computed in the previous step is below a given threshold. In this scenario we empirically set the threshold to 0.66;
6. Finally, we filter candidate entities according to the percentage of words that: (1) are stop words, (2) are common words[9]; and (3) do not contain at least one upper-case character. We remove the entity if one of the aforementioned criteria is above the 33%.

The output of the entity recognition module is a list of candidate entities with their set of candidate DBpedia concepts.

For the disambiguation, we exploit an adaptation of the distributional Lesk algorithm proposed by Basile et al. (Basile et al., 2015a; Basile et al., 2014) for disambiguating named entities. The algorithm replaces the concept of word overlap initially introduced by Lesk (1986) with the broader concept of semantic similarity computed in a distributional semantic space. Let $e_1, e_2, ...e_n$ be the sequence of entities extracted from the tweet, the algorithm disambiguates each target entity e_i by computing the semantic similarity between the glosses of concepts associated with the target entity and its context. The context and the gloss are represented as the vector sum of words they are composed of in a Distributional Semantic Model (DSM). The similarity between the two vectors, computed as the cosine of the angle between them, takes into account the word co-occurrence evidences previously collected through a corpus of documents. We exploit the word2vec tool[10] (Mikolov et al., 2013) in order to build a DSM, by analyzing all the pages in the last Italian Wikipedia dump[11]. The semantic similarity score is combined with a function which takes into account the frequency of the concept usage. More details are reported in (Basile et al., 2015a; Basile et al., 2014; Basile et al., 2015b).

4 Evaluation

The evaluation aims to compare several entity linking tools for Italian language exploiting the proposed dataset. We include in the evaluation our method that is an adaptation of the system that participated in the NEEL challenge for English tweets (Basile et al., 2015b).

We select three tools able to perform entity linking for Italian: TAGME, Babelfy, and TextRazor. TAGME (Ferragina and Scaiella, 2010) has a particular option that enables a special parser for Twitter messages. This parser has been designed to better handle entities in tweets like URL, user mentions and hash-tag. However, some other tools are not developed specifically for Twitter. For example, Babelfy (Moro et al., 2014) is an algorithm for entity linking and disambiguation developed for generic texts that uses BabelNet (Navigli and Ponzetto, 2012) as knowledge source. The third system is TextRazor[12], a commercial system able to recognize, disambiguate and link entities in ten languages, including Italian. Systems are compared using the typical metrics of precision, recall and F-measure. We compute the metrics in two settings: the **exact match** set requires that both

[7] http://lucene.apache.org/

[8] `http://www.ark.cs.cmu.edu/TweetNLP/`

[9] We exploit the list of 1,000 most frequent Italian words: `http://telelinea.free.fr/italien/1000_parole.html`

[10] https://code.google.com/p/word2vec/

[11] We use 400 dimensions for vectors analysing only terms that occur at least 25 times.

[12] `https://www.textrazor.com/`

start and end offsets match the gold standard annotation, while in **non exact match** the offsets provided by the systems can differ of two positions with respect to the gold standard.

Each algorithm provides a different output that needs some post-processing operations in order to make it comparable with our annotation standard. Most of the annotations are made with respect to the canonicalized version of DBpedia, while only 6% of the dataset is annotated using Wikipedia page URLs or the localized version (just one). Babelfy is able to directly provide canonicalized DBpedia URIs. When a BabelNet concept is not mapped to a DBpedia URIs, we return a NIL instance. TAGME returns Italian Wikipedia page titles that we easily translate into DBpedia URIs. We firstly try to map the page title in the canonicalized DBpedia, otherwise into the Italian one. TextRazor supplies Italian Wikipedia URLs or English Wikipedia URLs that we map to DBpedia URIs. Our algorithm provides Italian DBpedia URIs that we translate into canonicalized URIs when it is possible, otherwise we keep the Italian URIs. To recap: all algorithms are able to provide canonicalized and localized DBpedia URIs, only Babelfy is limited to canonicalized URIs.

Table 1: Results of the entity recognition evaluation with exact and non exact match.

	Exact match			Non exact match		
System	P	R	F	P	R	F
Babelfy	.431	.161	.235	.449	.168	.244
TAGME	.363	.458	.405	.391	.492	.436
TextRazor	.605	.310	.410	.605	.310	.410
DSMTEL	.470	.505	**.487**	.495	.532	**.513**

Table 1 reports the results about the entity recognition task with respect to exact and non exact match respectively. DSM-TEL provides the best results followed by TextRazor (exact match) and TAGME (non exact match), while the low performance of Babelfy proves that it is not able to tackle the irregular language used in tweets. In all the cases TextRazor achieves the best precision.

Entity linking performance are reported in Tables 2. It is important to underline that a correct linking requires the proper recognition of the entity involved. TextRazor achieves the best performance in the entity linking task with an F-measure in both exact and non exact match of 0.280.

Moreover, in order to understand if the recognition and linking tasks pose more challenges for Italian language, we evaluated all the systems on an English dataset. Although the two datasets are not directly comparable (due to the different sizes and the number of entities involved per tweet), we run an experiment over the Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge dataset (Rizzo et al., 2015) (Table 2). The evaluation results show a different behaviour of the systems for the English language. F-measure values are slightly lower than for Italian and TextRazor almost always outperforms other systems, with the only exception of TAGME for the linking with non exact match.

Table 2: Results of the entity linking evaluation with exact and non exact match.

	Exact match			Non exact match		
System	P	R	F	P	R	F
Babelfy	.318	.119	.173	.322	.120	.175
TAGME	.226	.284	.252	.235	.296	.262
TextRazor	.413	.212	**.280**	.413	.212	**.280**
DSM-TEL	.245	.263	.254	.254	.272	.263

Table 3: F-Measure results for English #Microposts2015 NEEL dataset.

	Recognition		Linking	
System	Exact	No Exact	Exact	No Exact
Babelfy	.134	.137	.102	.104
TAGME	.352	.381	.290	**.311**
TextRazor	**.460**	**.485**	**.294**	.295
DSMTEL	.442	.467	.284	.299

5 Conclusion

We tackled the problem of entity linking for Italian tweets. Our contribution is threefold: 1) we build a first Italian tweet dataset for entity linking, 2) we adapted a distributional-based NEL algorithm to the Italian language, and 3) we compared state-of-the-art systems on the built dataset. As for English, the entity linking task for Italian tweets turn out to be quite difficult, as pointed out by the very low performance of all systems employed. As future work we plan to extend the dataset in order to provide more examples for training and testing data.

References

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, Georgia, June. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015a. Uniba: Combining distributional semantic models and sense distribution for multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 360–364, Denver, Colorado, June. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Fedelucio Narducci. 2015b. UNIBA: Exploiting a Distributional Semantic Model for Disambiguating and Linking Entities in Tweets. In *Proceedings of the the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference (WWW 2015)*, volume 1395, pages 62–63. CEUR-WS.

Paolo Ferragina and Ugo Scaiella. 2010. Fast and accurate annotation of short texts with wikipedia pages. *IEEE Software*, 29(1):70–75.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proc. of SIGDOC '86*, SIGDOC '86, pages 24–26. ACM.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of ICLR Work.*

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.

Giuseppe Rizzo, Amparo Elizabeth Cano Basave, Bianca Pereira, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. 2015. Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In *Proceedings of the the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference (WWW 2015)*, volume 1395, pages 44–53. CEUR-WS.

Deep Tweets: from Entity Linking to Sentiment Analysis

Pierpaolo Basile[1], Valerio Basile[2], Malvina Nissim[2], Nicole Novielli[1]
[1]Department of Computer Science, University of Bari Aldo Moro
[2]Center for Language and Cognition Groningen, Rijksuniversiteit Groningen
{pierpaolo.basile,nicole.novielli}@uniba.it,
{v.basile,m.nissim}@rug.nl

Abstract

English. The huge amount of information streaming from online social networking is increasingly attracting the interest of researchers on sentiment analysis on micro-blogging platforms. We provide an overview on the open challenges of sentiment analysis on Italian tweets. We discuss methodological issues as well as new directions for investigation with particular focus on sentiment analysis of tweets containing figurative language and entity-based sentiment analysis of micro-posts.

Italiano. *L'enorme quantità di informazione presente nei social media attira sempre più l'attenzione della ricerca in sentiment analysis su piattaforme di micro-blogging. In questo articolo si fornisce una panoramica sui problemi aperti riguardo l'analisi del sentimento di tweet in italiano. Si discute di problemi metodologici e nuove direzioni di ricerca, con particolare attenzione all'analisi della polarità di tweet contenenti linguaggio figurato e riguardo specifiche entità nel micro-testo.*

1 Introduction

Flourished in the last decade, sentiment analysis is the study of the subjectivity and polarity (positive vs. negative) of a text (Pang and Lee, 2008). Traditionally, sentiment analysis techniques have been successfully exploited for opinionated corpora, such as news (Wiebe et al., 2005) or reviews (Hu and Liu, 2004). With the worldwide diffusion of social media, sentiment analysis on micro-blogging (Pak and Paroubek, 2010) is now regarded as a powerful tool for modelling socio-economic phenomena (O'Connor et al., 2010; Jansen et al., 2009).

The success of the tasks of sentiment analysis on Twitter at SemEval since 2013 (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015) attests this growing trend (on average, 40 teams per year participated). In 2014, Evalita also successfully opens a track on sentiment analysis with SENTIPOLC, the task on sentiment and polarity classification of Italian Tweets (Basile et al., 2014). With 12 teams registered, SENTIPOLC was the most popular task at Evalita 2014, confirming the great interest of the NLP community in sentiment analysis on social media, also in Italy.

In a world where e-commerce is part of our everyday life and social media platforms are regarded as new channels for marketing and for fostering trust of potential customers, such great interest in opinion mining from Twitter isn't surprising. In this scenario, what is also rapidly gaining more and more attention is being able to mine opinions about *specific aspects* of objects. Indeed, interest in Aspect Based Sentiment Analysis (ABSA) is increasing, and SemEval dedicates now a full task to this problem, since 2014 (Pontiki et al., 2014). Given a target of interest (e.g., a product or a brand), ABSA traditionally aimed at summarizing the content of users' reviews in several commercial domains (Hu and Liu, 2004; Ganu et al., 2013; Thet et al., 2010). In the context of ABSA, an interesting task is represented by finer-grained assignment of sentiment to entities. To this aim, mining information from micro-blogging platforms also involves reliably identifying entities in tweets. Hence, entity linking on twitter is gaining attention, too (Guo et al., 2013).

Based on the above observations, we discuss open issues in Section 2. In Section 3, we propose an extension of the SENTIPOLC task for Evalita 2016 by also introducing entity-based sentiment analysis as well as polarity detection of messages containing figurative language. Finally, we discuss the feasibility of our proposal in Section 4.

2 Open Challenges

From an applicative perspective, microposts comprise an invaluable wealth of data, ready to be mined for training predictive models. Analysing the sentiment conveyed by microposts can yield a competitive advantage for businesses (Jansen et al., 2009) and mining opinions about specific aspects of entities being discussed is of paramount importance in this sense. Beyond the pure commercial application domain, analysis of microposts can serve to gain crucial insights about political sentiment and election results (Tumasjan et al., 2010), political movements (Starbird and Palen, 2012), and health issues (Michael J. Paul, 2011).

By including explicit reference to entities, ABSA could broaden its impact beyond its traditional application in the commercial domain. While classical ABSA focus on the sentiment/opinion with respect to a particular aspect, entity-based sentiment analysis (Batra and Rao, 2010) tackles the problem of identifying the sentiment about an entity, for example a politician, a celebrity or a location. Entity-based sentiment analysis is a topic which has been unexplored in evaluation campaigns for Italian, and which could gain the interest of the NLP community.

Another main concern is the correct polarity classification of tweets containing figurative language such as irony, metaphor, or sarcasm (Karoui et al., 2015). Irony has been explicitly addressed so far in both the Italian and the English (Ghosh et al., 2015) evaluation campaigns. In particular, in the SENTIPOLC irony detection task, participants were required to develop systems able to decide whether a given message was ironic or not. In a more general vein, the SemEval task invited participants to deal with different forms of figurative language and the goal of the task was to detect polarity of tweets using it. In both cases, participant submitted systems obtaining promising performance. Still, the complex relation between sentiment and figurative use of language needs to be further investigated. While, in fact, irony seems to mainly act as a polarity reverser, other linguistic devices might impact sentiment in different ways.

Traditional approaches to sentiment analysis treat the subjectivity and polarity detection mainly as text classification problems, exploiting machine-learning algorithms to train supervised classifiers on human-annotated corpora. Sentiment analysis on micro-blogging platforms poses new challenges due to the presence of slang, misspelled words, hashtags, and links, thus inducing researchers to define novel approaches that include consideration of micro-blogging features for the sentiment analysis of both Italian (Basile et al., 2014) and English (Rosenthal et al., 2015) tweets. Looking at the reports of the SemEval task since 2013 and of the Evalita challenge in 2014, we observe that almost all submitted systems relied on supervised learning.

Despite being in principle agnostic with respect to language and domain, supervised approaches are in practice highly domain-dependent, as systems are very likely to perform poorly outside the domain they are trained on (Gamon et al., 2005). In fact, when training classification models, it is very likely to include consideration of terms that associate with sentiment because of the context of use. It is the case, for example, of political debates, where names of countries afflicted by wars might be associated to negative sentiments; analogous problems might be observed for the technology domain, where killer features of devices referred in positive reviews by customers usually become obsolete in relatively short periods of time (Thelwall et al., 2012). While representing a promising answer to the cross-domain generalizability issue of sentiment classifiers in social web (Thelwall et al., 2012), unsupervised approaches have not been exhaustively investigated and represent an interesting direction for future research.

3 Task Description

Entity linking and sentiment analysis on Twitter are challenging, attractive, and timely tasks for the Italian NLP community. The previously organised task within Evalita which is closest to what we propose is SENTIPOLC 2014 (Basile et al., 2014). However, our proposal differs in two ways. First, sentiment should be assigned not only at the message level but also to entities found in the tweet. This also implies that entities must be first recognised in each single tweet, and we expect systems also to link them to a knowledge base. The second difference has to do with the additional irony layer that was introduced in SENTIPOLC. Rather than dealing with irony only, we propose a *figurative* layer, that encompasses irony and any other shifted sentiment.

The entity linking task and the entity-based polarity annotation subtask of the sentiment analysis

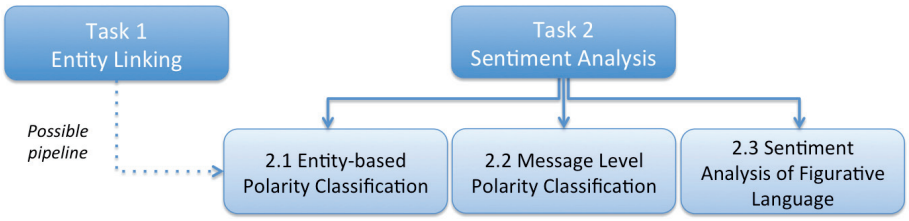


Figure 1: Task organization scheme

task can be seen as separate or as a pipeline, for those who want to try develop an end-to-end system, as depicted in Fig. 1.

3.1 Task 1 - Entity Detection and Linking

The goal of entity linking is to automatically extract entities from text and link them to the corresponding entries in taxonomies and/or knowledge bases as DBpedia or Freebase. Only very recently, entity linking in Twitter is becoming a popular tasks for evaluation campaigns (see Baldwin et al. (2015)).

Entity detection and linking tasks are typically organized in three stages: 1) identification and typing of entity mention in tweets; 2) linking of each mention to an entry in a knowledge-base representing the same real world entity, or NIL in case such entity does not exist; 3) cluster all NIL entities which refer to the same entity.

3.2 Task 2 - Message Level and Entity-Based Sentiment Analysis

The goal of the SENTIPOLC task at Evalita 2014 was the sentiment analysis at a message level on Italian tweets. SENTIPOLC was organized so as to include subjectivity and polarity classification as well as irony detection.

Besides the traditional task on message-level polarity classification, in the next edition of Evalita special focus should be given to entity-based sentiment analysis. Given a tweet containing a marked instance of an entity, the classification goal would be to determine whether positive, negative or neutral sentiment is attached to it.

As for the role of figurative language, the analysis of the performance of the systems participating in SENTIPOLC irony detection subtask shows the complexity of this issue. Thus, we believe that further investigation of the role of figurative language in sentiment analysis of tweets is needed, by also incorporating the lesson learnt from the task on figurative language at SemEval 2015 (Ghosh et al., 2015). Participants would be required to predict the overall polarity of tweets containing figurative language, distinguishing between the literal meaning of the message and its figurative, intended meaning.

4 Feasibility

The annotated data for entity linking tasks (such as our proposed Task 1) typically include the start and end offsets of the entity mention in the tweet, the entity type belonging to one of the categories defined in the taxonomy, and the URI of the linked DBpedia resource or to a NIL reference. For example, given the tweet `@FabioClerici sono altri a dire che un reato. E il "politometro" come lo chiama #Grillo vale per tutti. Anche per chi fa #antipolitica`, two entities are annotated: `FabioClerici` (offsets 1-13) and `Grillo` (offsets 85-91). The first entity is linked as NIL since `Fabio Clerici` has not resource in DBpedia, while `Grillo` is linked with the respective URI: `http://dbpedia.org/resource/Beppe_Grillo`. Analysing similar tasks for English, we note that organizers provide both training and test data. Training data are generally used in the first stage, usually approached by supervised systems, while the linking stage is generally performed using unsupervised or knowledge based systems. As knowledge base, the Italian version of DBpedia could be adopted, while the entity taxonomy could consist of the following classes: Thing, Event, Character, Location, Organization, Person and Product.

As for Task 2, it is basically conceived as a follow-up of the SENTIPOLC task. In order to ensure continuity, it makes sense to carry out the annotation using a format compatibile with the existing dataset. The SENTIPOLC annotation scheme consists in four binary fields, indicating the presence of subjectivity, positive polarity, negative polarity, and irony. The fields are not mutually ex-

clusive, for instance both positive and negative polarity can be present, resulting in a *mixed* polarity message. However, not all possible combinations are allowed. Table 1 shows some examples of annotations from the SENTIPOLC dataset.

Table 1: Proposal for an annotation scheme that distinguishes between literal polarity (`pos`, `neg`) and overall polarity (`opos`, `oneg`).

`subj`	`pos`	`neg`	`iro`	**`opos`**	**`oneg`**	description
0	0	0	0	**0**	**0**	objective tweet
1	1	0	0	**1**	**0**	subjective tweet positive polarity no irony
1	0	0	0	**0**	**0**	subjective tweet neutral polarity no irony
1	0	1	0	**0**	**1**	subjective tweet negative polarity no irony
1	0	1	1	**1**	**0**	subjective tweet negative literal polarity positive overall polarity
1	1	0	1	**0**	**1**	subjective tweet positive literal polarity negative overall polarity

With respect to the annotation adopted in SENTIPOLC, two additional fields are reported to reflect the task organization scheme we propose in this paper, including the sentiment analysis of tweet containing figurative language. These fields, highlighted in bold face, encode respectively the presence of positive and negative polarity *considering the eventual polarity inversion due to the use of figurative language*, thus the existing `pos` and `neg` fields refer to *literal* polarity of the tweet. For the annotation of the gold standard dataset for SENTIPOLC, the polarity of ironic messages has been annotated according to the intended meaning of the tweets, so for the new task the literal polarity will have to be manually annotated in order to complete the gold standard. Annotation of items in the figurative language dataset could be the same as in the message-level polarity detection task of Evalita, but tweets would be opportunistically selected only if they contain figurative language, so as to reflect the goal of the task.

For the entity-based sentiment analysis subtask, the boundaries for the marked instance will be also provided by indicating the offsets of the entity for which the polarity is annotated, as it was done for SemEval (Pontiki et al., 2014; Pontiki et al., 2015). Participants who want to attempt entity-based sentiment analysis only can use the data that contains the gold output of Task 1, while those who want to perform entity detection and linking only, without doing sentiment analysis, are free to stop there. Participants who want to attempt both tasks can treat them in sequence, and evaluation can be performed for the whole system as well as for each of the two tasks (for the second one over gold input), as it will be done for teams that participate in one task only.

For both tasks, the annotation procedure could follow the consolidated methodology from the previous tasks, such as SENTIPOLC. Experts label manually each item, then agreement is checked and disagreements are resolved by discussion.

Finally, so far little investigation was performed about unsupervised methods and the possiblity they offer to overcome domain-dependence of approaches based on machine learning. In a challenge perspective, supervised systems are always promising since they guarantee a better performance. A possible way to encourage teams to explore original approaches could be to allow submission of two separate runs for supervised and unsupervised settings, respectively. Ranking could be calculated separately, as already done for the constrained and unconstrained runs in SENTIPOLC. To promote a fair evaluation and comparison of supervised and unsupervised systems, corpora from different domains could be provided as train and test sets. To this aim, it could be possible to exploit the *topic* field in the annotation of tweets used in the SENTIPOLC dataset, where a flag indicates whether the tweets refer to the political domain or not. Hence, the train set could be built by merging political tweets from both the train and the test set used in SENTIPOLC. A new test set would be created by annotating tweets in one or more different domains.

To conclude, we presented the entity linking and sentiment analysis tasks as related to one another, as shown in the pipeline in Figure 1, specifying that participants will be able to choose which portions of the tasks they want to concentrate on. Additionally, we would like to stress that this proposal could also be conceived as two entirely separate tasks: one on sentiment analysis at the entity level, including entity detection and linking, and one on sentiment analysis at the message level, including the detection of figurative readings, as a more direct follow-up of SENTIPOLC.

Acknowledgements

This work is partially funded by the project "Investigating the Role of Emotions in Online Question & Answer Sites", funded by MIUR under the program SIR 2014.

References

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Association for Computational Linguistics (ACL)*. ACL, Association for Computational Linguistics, August.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proc. of EVALITA 2014*, Pisa, Italy.

Siddharth Batra and Deepak Rao. 2010. Entity based sentiment analysis on twitter. *Science*, 9(4):1–12.

Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. 2005. Pulse: Mining customer opinions from free text. In *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*, IDA'05, pages 121–132, Berlin, Heidelberg. Springer-Verlag.

Gayatree Ganu, Yogesh Kakodkar, and AméLie Marian. 2013. Improving the quality of predictions using textual information in online user reviews. *Inf. Syst.*, 38(1):1–15.

AAniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and Jhon 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 470–475, Denver, Colorado, USA. Association for Computational Linguistics.

Stephen Guo, Ming-Wei Chang, and Emre Kiciman. 2013. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1020–1030, Atlanta, Georgia, June. Association for Computational Linguistics.

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*, pages 168–177.

Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188.

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 *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 644–650.

Mark Dredze Michael J. Paul. 2011. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 265–272.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Intl AAAI Conf. on Weblogs and Social Media (ICWSM)*, volume 11, pages 122–129.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. of the Seventh Intl Conf. on Language Resources and Evaluation (LREC'10)*.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, January.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion . 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June. 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, August.

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, Denver, Colorado, June.

Kate Starbird and Leysia Palen. 2012. (how) will the revolution be retweeted?: Information diffusion and the 2011 egyptian uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 7–16, New York, NY, USA. ACM.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Tun Thura Thet, Jin-Cheon Na, and Christopher S.G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.*, 36(6):823–848.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Web and Social Media*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

Enhancing the Accuracy of Ancient Greek WordNet by Multilingual Distributional Semantics

Yuri Bizzoni[1], Riccardo Del Gratta[1], Federico Boschetti[1], Marianne Reboul[2]
[1]ILC-CNR, Pisa
[2]Université de Paris 4, Paris
{yuri.bizzoni,riccardo.delgratta}@gmail.com,
federico.boschetti@ilc.cnr.it, marianne.reboul@free.fr

Abstract

English. We discuss a method to enhance the accuracy of a subset of the Ancient Greek WordNet based on the Homeric lexicon and the related conceptual network, by using multilingual semantic spaces built from aligned corpora.

Italiano. *Esponiamo un metodo per migliorare l'accuratezza di un sottoinsieme dell' Ancient Greek WordNet, basato sul lessico Omerico e sulla relativa rete concettuale, attraverso l'uso di spazi semantici plurilingui costruiti su corpora paralleli allineati.*

1 Introduction

The Ancient Greek WordNet (AGWN) represents the first attempt to build a WordNet for Ancient Greek (Bizzoni et al., 2014).

The AGWN synsets are aligned to Princeton WordNet (PWN) (Fellbaum, 1998), to Italian WordNet (IWN) (Roventini et al., 2003), developed at the Institute for Computational Linguistic "A. Zampolli" in Pisa, to the Italian section of MultiWordNet,[1] developed at Bruno Kessler Foundation and to a Latin WordNet (LWN) created with the same criteria of AGWN and linked to Minozzi's Latin WordNet (Minozzi, 2009) and (McGillivray, 2010), developed at the University of Verona. In this way the user is allowed to find the equivalents of a set of synonyms into different languages. The AGWN can be freely accessed through a Web interface,[2] which allows enabled users to add or delete words in the synsets, adapt the glosses and validate the lexico-semantic relations.[3] We first created AGWN by bootstrapping Greek-English pairs from bilingual dictionaries and by assigning Greek words to PWN synsets associated to the corresponding English translations. As a drawback of this method, a large number of synsets and lexico-semantic relations are spuriously over-generated by English homonymy and polysemy. As exposed in (Bizzoni et al., 2014), to have PWN as a pivoting resource[4] propagates the same drawback to other connected WordNet in CoPhiWordNet Platform. In order to improve the accuracy of a subset of AGWN synsets related to the Homeric lexicon and the related conceptual network, we have automatically extracted word translations from Greek-Italian parallel texts by applying distributional semantic strategies illustrated in the following sections and verified how many of these translation were in CoPhiWn. According to the methodology explained in (Francis Bond and Uchimoto, 2008), trilingual resources (in our case the original Greek-English pairs extracted from dictionaries and the Greek-Italian pairs extracted from aligned translations) are useful to enhance the accuracy of a bootstrapped WordNets.

2 Translation Mining through Semantic Spaces

We present a way to automatically improve the accuracy of Ancient Greek word translations by applying the principles of distributi-

[1]http://multiwordnet.fbk.eu
[2]GUI beta-version at http://www.languagelibrary.eu/new_ewnui

[3]In the following, when we use the term CoPhiWordNet Platform (CoPhiWn) we mean the three WordNets: AGWN, IWNand PWN.
[4]For example,PWN links through ILI (Vossen, 1998) AGWN to IWN

onal semantics to aligned corpora (Dumais et al., 1997) and (Yuri, 2015). We will first explain the ratio of this method and then show how it is useful to improve AGWN in several ways (see Section 2.7). Although Ancient Greek obviously does not have native speakers, we dispose of a great variety of translations of the same classical texts written in several languages and different historical periods. The study of large diachronical corpora of translations is both relevant in classical studies and a valuable source of information to build or improve the accuracy of multilingual lexico-semantic resources (see Section 3).

2.1 Aligning long and literary-biased translations to the original text

We applied a strategy to automatically align Greek-Italian parallel corpora through two main steps: in the first step we segmented texts in small portions; in the second step we linked those texts together. The result is that each Ancient Greek segment is aligned to its translations. After the segment-to-segment alignment, we applied the distributional semantics method illustrated below, in order to identify word-to-word translations.

2.2 Distributional Semantics

It is argued by several linguists (Miller, 1971) and (Firth, 1975) that one of the best ways to define the meaning of a word is the study of the relations with the other words in the close context. So it is possible to hypothesize that we learn the meaning of many new words thanks to the way they are linked to words we already know, and in general, that we learn the meaning of words by perceiving their verbal as well as non-verbal context. We can study semantic similarities between terms by quantifying their distribution: similar words will have similar contexts. In the same way, we can suppose that, in an aligned parallel corpus, a word and its translation will tend to appear in the same aligned segments. For this reason, the contextual segment of the original Greek word and the contextual aligned segment of the translation have the same identifier.

2.3 Semantic Spaces based on aligned corpora

There are several kinds of linguistic contexts that can be selected to study word similarity (Lenci, 2008):

- window-based collocates: two words co-occur if they appear in a given context window;
- text regions: two words co-occur if they appear in a same textual area such as a document, a paragraph, and so on;
- syntactic collocates: two words co-occur if they appear in a same syntactic pattern, for example if they are the direct objects of a verb, etc.

Although the most typical approach to distributional semantics is the use of window-based collocates, this kind of context becomes useless in multilingual corpora, since words in different languages do not share a common context. We use the method based on text regions collocates, which considers every couple of aligned segments as the default textual area. Word vectors of 0s and 1s in both languages are constructed accordingly to the absence/presence of the word in the aligned couple.
Thus, Ancient Greek and Italian words are mingled together in the vectorial space.[5]

2.4 Words and their translations tend to be neighbors

With a similar procedure, Ancient Greek and Italian equivalent words will happen to have similar vectors, since they will appear in the same aligned chunks. Consequently they will be close in the resulting semantic space. To compute the proximity of vectors we used the cosine similarity measure (Sahlgren, 2006).

2.5 Parts of Speech TRanslations

Performance on nouns is higher than performance on verbs, adjectives and adverbs, due to larger translational fluctuations for the latter parts of speech. Anyway, although verbs

[5] In our experiment the resulting vector has a dimension of $\sim 60k$

are more polysemous than nouns, we apparently are able to find relevant verb translations: *uccidere - kteíno (to kill)*, *morire - thnésko* (to die), *amare - philéo (to love)* and even *essere - eimí (to be)*. The same holds for adjectives, but, however,we found acceptable results also in this category: *bello - kalòs (beautiful)*, *nobile - agauòs (noble)*. Interestingly, from color adjectives we were only able to retrieve black and white translations: *neromélas (black)*, *bianco-leukós (white)*. Color adjectives in Ancient Greek are naturally complex to analyze, since it is hard to retrieve their exact meaning in absence of speakers; this indetermination apparently propagates to our outcomes.

Finally, it is also relevant to observe that extremely polysemous categories like adverbs in some cases find a correct translation: *ek - fuori (out)*, *non-ou - non (not)*.

2.6 Data Presentation and Some Results

We extracted the five most similar items for 121 Ancient Greek words (randomly chosen from different groups of frequency) from a semantic space built on the original texts, i.e five complete Iliad translations and four complete Odyssey translation in Italian aligned to the original texts. The original data resulted in 605 rows (121 time 5pairs); when it comes to verify whether a Greek/Italian pair is mapped in CoPhiWn, we expect that the modern polysemy, the one inducted by English to Italian mapping will increase the number of pairs. Indeed, we found that 605 pairs correspond to 736 Greek-English-Italian possible triples. However, only 176 triples have been successfully found in CoPhiWn. A manual validation of the resulting set excluded 13 triples which are caused by the modern polysemy reducing the found triples to 164. Not surprisingly, the coverage of the triples in CoPhiWn $\sim 23\%$ is quite close to the coverage of AGWN, cf. (Bizzoni et al., 2014) ($\sim 28\%$).

2.7 AGWN: strenghtening bilingual links

If an Ancient Greek word is linked to an Italian word in CoPhiWn and it is distributionally near to the same Italian word in a semantic space, the probability that this link is correct is high. For instance, the word *pólemos*, frequent in Homer, is linked in CoPhiWn to an Italian synset composed by the words *guerra, battaglia, ostilità*. The first two terms appear also to be the nearest Italian terms to the word *pólemos (war, battle)* in our semantic space. This match helps us to increase the probability that *guerra* and *battaglia* are sound translations of *pólemos*, and thus that the Italian and Greek synsets are correctly interlinked.

In CoPhiWn the word *hémar (day)* is linked to the synonyms *giorno, giornata*, and in our semantic space it appears very similar to the word *giorno* only. But the distributional information from our semantic space reinforces the association between *hémar* and the overall Italian synset.

This way to retrieve crosslingual information from textual corpora is highly helpful to discover errors due to the employ of polysemy in different languages. For instance, in CoPhiWn, the word *astér (star)* is linked both to the synset associated to the word *stella*, glossed as star in the sky and to the synset associated to the word *divo*, glossed as star in the show business, due to the intermediation of the English word *star*,[6] while, as expected, *astér* is distributionally similar only to *stella* in our semantic space. The word *dóru (spear and mast)* is linked on one hand to *asta, arma* synset and on the other hand to *prora, prua*, glossed as parts of the boat, which is synecdochically related to the mast, but in our semantic space it appears near only to the words of the first group, allowing us to score higher only the first equivalence. It is important to remember that we can incur in cases of stylistically biased translations and synonyms: *árma (charriot)* can be *cocchio* or *carro* in different translations.

Additional examples are the following: the most similar terms to Italian *mare* in our semantic space are *thálassa, háls, póntos*, three words indicating the concept of sea clustered together by their common translation. *scudo* (shield) is associated both to *aspís* and *sakós*, *soffio* (breath) leads to *pnóe* and *ánemos*, through *popolo* (people) we find *láos, démos* and among the most similar words of *dolore* (pain) we find both *pénthos* and *álgos*. With the same mechanisms that allow to find word to word

[6] This is one effect of the modern polysemy described in Section 2.6.

translations, we can find also some small sets of potential synonyms in the same language looking at their distributional behavior: so *aithér* is near to *oúranos* and *hétor* is near to *thumós*.

2.8 CoPiWn! (CoPiWn!): supporting hypernym/hyponym relations

A system based on distributional semantics tends to cluster together not only bilingual synonyms and translations, but also hypernyms and hyponyms. They tend to have distributionally similar, although not identical, behaviors, and it can easily happen that a word is translated with a hypernym, or more rarely with a hyponym, in another language. Systems to discriminate between hypernyms and synonyms in semantic spaces could become very useful in this context. See for example (Benotto, 2013) and Lenci et al. 2012.

3 Conclusions and Future Work

We have elaborated a system to enhance the accuracy of Ancient Greek WordNet. This system appears to be useful to verify the soundness of automatically generated links between the Ancient Greek WordNet and WordNet in other languages. The method aims at increasing the precision of the Greek-Italian pairs within their translations, since it removes modern polysemy and discards translations in CoPhiWn that are not supported by actual texts' translations.

References

Giulia Benotto. 2013. Modelli distribuzionali delle relazioni semantiche: il caso dell'iperonimia. *Animali, Umani, Macchine. Atti del convegno 2012 del CODISCO*.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta , Monica Monachini, and Gregory Crane. 2014. The Making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, MA, USA.

John Rupert Firth. 1975. *Modes of meaning*. College Division of Bobbs-Merrill Company.

Kyoko Kanzaki Francis Bond, Hitoshi Isahara and Kiyotaka Uchimoto. 2008. Boot-strapping a wordnet using multiple existing wordnets. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics*, 20(1):1–31.

Barbara McGillivray. 2010. Automatic selectional preference acquisition for Latin verbs. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, pages 73–78. ACL.

George A Miller. 1971. Empirical methods in the study of semantics. *Semantics, an interdisciplinary reader in philosophy, linguistics, and psychology*, pages 569–585.

Stefano Minozzi. 2009. The Latin WordNet Project. In Peter Anreiter and Manfred Kienpointner, editors, *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137 of *Innsbrucker Beiträge zur Sprachwissenschaft*, pages 707–716.

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, XVIII-XIX, Pisa-Roma, IEPI*, 2:745–791.

Magnus Sahlgren. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Bizzoni Yuri. 2015. The Italian Homer - The Evolutions of Translation Patterns between the XVIII and the XXI century. Master's thesis, University of Pisa.

Deep Neural Networks for Named Entity Recognition in Italian

Daniele Bonadiman[†], **Aliaksei Severyn**[*], **Alessandro Moschitti**[‡†]
[†]DISI - University of Trento, Italy
[*]Google Inc.
[‡]Qatar Computing Research Institute, HBKU, Qatar
{bonadiman.daniele,aseveryn,amoschitti}@gmail.com

Abstract

English. In this paper, we introduce a Deep Neural Network (DNN) for engineering Named Entity Recognizers (NERs) in Italian. Our network uses a sliding window of word contexts to predict tags. It relies on a simple word-level log-likelihood as a cost function and uses a new recurrent feedback mechanism to ensure that the dependencies between the output tags are properly modeled. The evaluation on the Evalita 2009 benchmark shows that our DNN performs on par with the best NERs, outperforming the state of the art when gazetteer features are used.

Italiano. *In questo lavoro, si introduce una rete neurale* deep *(DNN) per progettare estrattori automatici di entitá nominate (NER) per la lingua italiana. La rete utilizza una finestra scorrevole di contesti delle parole per predire le loro etichette con associata probabilitá, la quale è usata come funzione di costo. Inoltre si utilizza un nuovo meccanismo di retroazione ricorrente per modellare le dipendenze tra le etichette di uscita. La valutazione della DNN sul* dataset *di Evalita 2009 indica che è alla pari con i migliori NER e migliora lo stato dell'arte quando si aggiungono delle* features *derivate dai dizionari.*

1 Introduction

Named Entity (NE) recognition is the task of detectings phrases in text, e.g., proper names, which directly refer to real world entities along with their type, e.g., people, organizations, locations, etc. see, e.g., (Nadeau and Sekine, 2007).

Most NE recognizers (NERs) rely on machine learning models, which require to define a large set of manually engineered features. For example, the state-of-the-art (SOTA) system for English (Ratinov and Roth, 2009) uses a simple averaged perceptron and a large set of local and non-local features. Similarly, the best performing system for Italian (Nguyen et al., 2010) combines two learning systems that heavily rely on both local and global manually engineered features. Some of the latter are generated using basic hand-crafted rules (i.e., suffix, prefix) but most of them require huge dictionaries (gazetteers) and external parsers (POS taggers and chunkers). While designing good features for NERs requires a great deal of expertise and can be labour intensive, it also makes the taggers harder to adapt to new domains and languages since resources and syntactic parsers used to generate the features may not be readily available.

Recently, DNNs have been shown to be very effective for automatic feature engineering, demonstrating SOTA results in many sequence labelling tasks, e.g., (Collobert et al., 2011), also for Italian language (Attardi, 2015).

In this paper, we target NERs for Italian and propose a novel deep learning model that can match the accuracy of the previous best NERs without using manual feature engineering and only requiring a minimal effort for language adaptation. In particular, our model is inspired by the successful neural network architecture presented by Collobert et al. (2011) to which we propose several innovative and valuable enhancements: (i) a simple recurrent feedback mechanism to model the dependencies between the output tags and (ii) a pre-training process based on two-steps: (a) training the network on a weekly labeled dataset and then (b) refining the weights on the supervised training set. Our final model obtains 82.81 in F1 on the Evalita 2009 Italian dataset (Speranza, 2009), which is an improvement of $+0.81$ over the Zanoli and Pianta (2009) system that won the competition. Our model only uses the words in the sentence, four morphological features and a

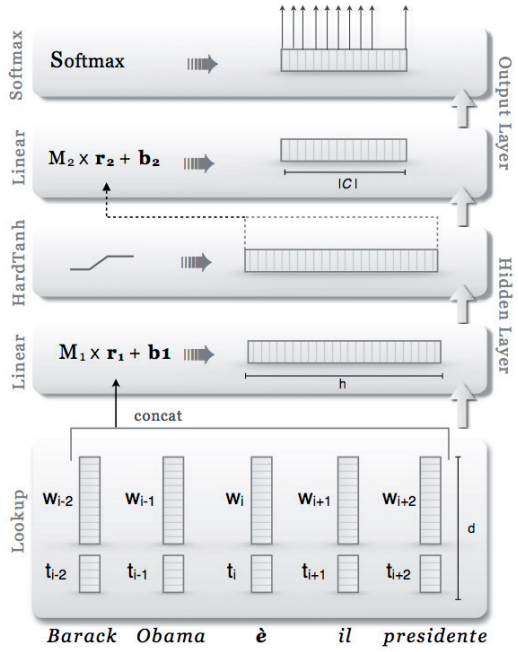


Figure 1: The architecture of Context Window Network (CWN) of Collobert et al. (2011).

gazetteer. Interestingly, if the gazetteer is removed from our network, it achieves an F1 of 81.42, which is still on par with the previous best systems yet it is simple and easy to adapt to new domains and languages.

2 Our DNN model for NER

In this section, we first briefly describe the architecture of the Context Window Network (CWN) from Collobert et al. (2011), pointing out its limitation. We then introduce our Recurrent Context Window Network (RCWN), which extends CWN and aims at solving its drawbacks.

2.1 Context Window Network

We adopt a CWN model that has been successfully applied by Collobert et al. (2011) for a wide range of sequence labelling NLP tasks. Its architecture is depicted in Fig. 1. It works as follows: given an input sentence $s = [w_1, \ldots, w_n]$, e.g., *Barack Obama è il presidente degli Stati Uniti D'America*[1], for each word w_i, the sequences of word contexts $[w_{i-k/2+1}, .., w_i, .., w_{i+k/2}]$ of size k around the target word w_i ($i = 1, .., n$) are used as input to the network.[2] For example, the Fig. 1 shows a network with $k = 5$ and the input sequence for the target word *è* at position $i = 3$.

[1] *Barack Obama is the president of the United States of America.*

[2] In case the target word i is at the beginning/end of a sentence, up to $(k-1)/2$ placeholders are used in place of the empty input words.

The input words w_i from the vocabulary V are mapped to d-dimensional word embedding vectors $\mathbf{w}_i \in \mathbb{R}^d$. Embeddings $\mathbf{w}_i$ for all words in V form an embedding matrix $\mathbf{W} \in \mathbb{R}^{|V| \times d}$, which is learned by the network. An embedding vector $\mathbf{w}_i$ for a word w_i is retrieved by a simple lookup operation in $\mathbf{W}$ (see lookup frame in Fig. 1). After the lookup, the k embedding vectors of the context window are concatenated into a single vector $\mathbf{r_1} \in \mathbb{R}^{kd}$, which is passed to the next hidden layer hl. It applies the following linear transformation: $hl(\mathbf{r_1}) = \mathbf{M_1} \cdot \mathbf{r_1} + b_1$, where the matrix of weights $\mathbf{M_1}$ and the bias b_1 parametrize the linear transformation and are learned by the network. The goal of the hidden layer is to learn feature combinations from the word embeddings of the context window.

To enable the learning of non-linear discriminative functions, the output of hl is passed through a non-linear transformation also called activation function, i.e., a *HardTanh*() non-linearity, thus obtaining, $\mathbf{r_2}$. Finally, the output classification layer encoded by the matrix $\mathbf{M_2} \in \mathbb{R}^{|C| \times h}$ and the bias b_2 are used to evaluate the vector $\mathbf{p} = softmax(\mathbf{M_2} \times \mathbf{r_2} + b_2)$ of class conditional probabilities, i.e., $\mathbf{p}_c = p(c|\mathbf{x}), c \in C$, where C is the set of NE tags, h is the dimension of the hl and $\mathbf{x}$ is the input context window.

2.2 Our model

The CWN model described above has several drawbacks: (i) each tag prediction is made by considering only local information, i.e., no dependencies between the output tags are taken into account; (ii) publicly available annotated datasets for NER are usually too small to train neural networks thus often leading to overfitting. We address both problems by proposing: (i) a novel recurrent context window network (RCWN) architecture; (ii) a network pre-training technique using weakly labeled data; and (iii) we also experiment with a set of recent techniques to improve the generalization of our DNN to avoid overfitting, i.e., we use *early stopping* (Prechelt, 1998), *weight decay* (Krogh and Hertz, 1992), and *Dropout* (Hinton, 2014).

2.2.1 Recurrent Context Window Network

We propose RCWN for modeling dependencies between labels. It extends CWN by using m previously predicted tags as an additional input, i.e., the previously predicted tags at steps $i-m, \ldots, i-1$ are used to predict the tag of the word at position i, where $m < k/2$. Since we proceed from left to right, words in the context window w_j with

Dataset	Articles	Sentences	Tokens
Train	525	11,227	212,478
Test	180	4,136	86,419

Table 1: Splits of the Evalita 2009 dataset

Dataset	PER	ORG	LOC	GPE
Train	4,577	3,658	362	2,813
Test	2,378	1,289	156	1,143

Table 2: Entities distribution in Evalita 2009

$j > i - 1$, i.e., at the right of the target word, do not have their predicted tags, thus we simply use the special unknown tag, `UNK`, for them.

Since NNs provide us with the possibility to define and train arbitrary embeddings, we associate each predicted tag type with an embedding vector, which can be trained in the same way as word embeddings (see vectors for tags t_i in Fig. 1). More specifically, given k words $w_i \in \mathbb{R}^{d_w}$ in the context window and previously predicted tags $t_i \in \mathbb{R}^{d_t}$ at corresponding positions, we concatenate them together along the embedding dimension obtaining new vectors of dimensionality $d_w + d_t$. Thus, the output of the first input layer becomes a sequence of $k(d_w + d_t)$ vectors.

RCWN is simple to implement and is computationally more efficient than, for example, NNs computing *sentence log-likelihood*, which require Viterbi decoding. RCWN may suffer from an error propagation issue as the network can misclassify the word at position $t - i$, propagating an erroneous feature (the wrong label) to the rest of the sequence. However, the learned tag embeddings seem to be robust to noise[3]. Indeed, the proposed network obtains a significant improvement over the baseline model (see Section 3.2).

3 Experiments

In these experiments, we compare three different enhancements of our DNNs on the data from the Evalita challenge, namely: (i) our RCWN method, (ii) pre-training on weakly supervised data, and (iii) the use of gazetteers.

3.1 Experimental setup

Dataset. We evaluate our models on the Evalita 2009 Italian dataset for NERs (Speranza, 2009) summarized in Tab. 1. There are four types of NEs: person (PER), location (LOC), organization (ORG) and geo-political entity (GPE), (see Tab. 2). Data is annotated using the IOB tagging schema, i.e., for inside, outside and beginning of a entity, respectively.

Training and testing the network. We use (i) the Negative Log Likelihood cost function, i.e., $-log(\mathbf{p}_c)$, where c is the correct label for the target word, (ii) stochastic gradient descent (SGD) to learn the parameters of the network and (iii) the backpropogation algorithm to compute the updates. At test time, the tag c, associated with the highest class conditional probability $\mathbf{p}_c$, is selected, i.e., $c = \text{argmax}_{c \in C}\, \mathbf{p}_c$.

Features. In addition to words, all our models also use 4 basic morphological features: *all lowercase*, *all uppercase*, *capitalized* and *it contains uppercase character*. These can reduce the size of the word embedding dictionary as showed by (Collobert et al., 2011). In our implementation, these 4 binary features are encoded as one discrete feature associated with an embedding vector of size 5, i.e., similarly to the preceding tags in RCWN. Additionally, we use a similar vector to also encode gazetteer features. Gazetteers are collections of names, locations and organizations extracted from different sources such as the Italian phone book, Wikipedia and stock marked websites. Since we use four different dictionaries one for each NE class, we add four feature vectors to the network.

Word Embeddings. We use a fixed dictionary of size $100K$ and set the size of the word embeddings to 50, hence, the number parameters to be trained is $5M$. Training a model with such a large capacity requires a large amount of labelled data. Unfortunately, the sizes of the supervised datasets available for training NER models are much smaller, thus we mitigate such problem by pre-training the word embeddings on huge unsupervised training datasets. We use word2vec (Mikolov et al., 2013) skip-gram model to pre-train our embeddings on Italian dump of Wikipedia: this only took a few hours.

Network Hyperparameters. We used $h = 750$ hidden units, a learning rate of 0.05, the word embedding size $d_w = 50$ and a size of 5 for the embeddings of discrete morphological and gazetteer features. Differently, we used a larger embedding, $d_t = 20$ for the NE tags.

Pre-training DNN with gazetters. Good weight initialization is crucial for training better NN models (Collobert et al., 2011; Ben-

[3] We can use the same intuitive explanation of error correcting output codes.

Model	F1	Prec.	Rec.
Baseline	78.32	79.45	77.23
RCWN	81.39	82.63	80.23
RCWN+Gazz	83.59	84.85	82.40
RCWN+WLD	81.74	82.93	80.63
RCWN+WLD+Gazz	**83.80**	**85.03**	**82.64**

Table 3: Results on 10-fold cross-validation

gio, 2009). Over the years different ways of pre-training the network have been designed: layer-wise pre-training (Bengio, 2009), word embeddings (Collobert et al., 2011) or by relying on distant supervised datasets (Severyn and Moschitti, 2015). Here, we propose a pre-training technique using an off-the-shelf NER to generate noisily annotated data, e.g., a sort of distance/weakly supervision or self-training. Our Weakly Labeled Dataset (WLD) is built by automatically annotating articles from the local newspaper "L'Adige", which is the same source of the training and test sets of Evalita challenge. We split the articles in sentences and tokenized them. This unlabeled corpus is composed of 20.000 sentences. We automatically tagged it using EntityPro, which is a NER tagger included in the TextPro suite (Pianta et al., 2008).

3.2 Results

Our models are evaluated on the Evalita 2009 dataset. We applied 10-fold cross-validation to the training set of the challenge[4] for performing parameter tuning and picking the best models.

Table 3 reports performance of our models averaged over 10-folds. We note that (i) modeling the output dependencies with RCWN leads to a considerable improvement in F1 over the CWN model of Collobert et al. (2011) (our baseline); (ii) adding the gazetteer features leads to an improvement both in Precision and Recall, and therefore in F1; and (iii) pre-training the network on the weakly labeled training set produces improvement (although small), which is due to a better initialization of the network weights.

Table 4 shows the comparative results between our models and the current state of the art for Italian NER on the Evalita 2009 official test set. We used the best parameter values derived when computing the experiments of Table 3. Our model using both gazetteer and pre-training outperforms all the systems participating to the Evalita 2009 (Zanoli and Pianta, 2009; Gesmundo, 2009). It should be noted that Nguyen et al. (2010) obtained better results using a CRF classifier followed by a reranker (RR) based on tree kernels. However, our approach only uses one learning algorithm, which is simpler than models applying multiple learning approaches, such as those in (Nguyen et al., 2010) and (Zanoli and Pianta, 2009). Moreover, our model outperforms the Nguyen et al. (2010) CRF baseline (which is given in input to the tree-kernel based reranker) by ~ 2.5 points in F1. Thus it is likely that applying their reranker on top of our model's output might produce a further improvement over SOTA.

[4]The official evaluation metric for NER is the F1, which is the harmonic mean between Precision and Recall.

Models	F1	Prec.	Rec.
Gesmundo (2009)	81.46	**86.06**	77.33
Zanoli and Pianta (2009)	82.00	84.07	80.02
Nguyen et al. (2010) (CRF)	80.34	83.43	77.48
Nguyen et al. (2010) + RR	**84.33**	85.99	**82.73**
RCWN	79.59	81.39	77.87
RCWN+WLD	81.42	82.74	80.14
RCWN+Gazz	81.47	83.48	79.56
RCWN+WLD+Gazz	82.81	85.69	80.10

Table 4: Comparison with the best NER systems for Italian. Models below the double line were computed after the Evalita challenge.

Finally, it is important to note that our model obtains an F1 comparable to the best system in Evalita 2009 without using any extra features (we only use words and 4 morphological features). In fact, when we remove the gazetteer features, our method still obtains the very high F1 of 81.42.

4 Conclusion

In this paper, we propose a new DNN for designing NERs in Italian. Its main characteristics are: (i) the RCWN feedback method, which can model dependencies of the output label sequence and (ii) a pre-training technique involving a weakly supervised dataset. Our system is rather simple and efficient as it involves only one model at test time. Additionally, it does not require time-consuming feature engineering or extensive data processing for their extraction.

In the future, we would like to apply rerankers to our methods and explore combinations of DNNs with structural kernels.

Acknowledgments

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

References

Giuseppe Attardi. 2015. Deepnl: a deep learning nlp pipeline. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 109–115, Denver, Colorado, June. Association for Computational Linguistics.

Yoshua Bengio. 2009. Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1):1–127.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (almost) from Scratch. *The Journal of Machine Learning Research*, 1(12):2493–2537.

Andrea Gesmundo. 2009. Bidirectional Sequence Classification for Named Entities Recognition. *Proceedings of EVALITA*.

Geoffrey Hinton. 2014. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1):1929–1958.

A. Krogh and J. Hertz. 1992. A Simple Weight Decay Can Improve Generalization. *Advances in Neural Information Processing Systems*, 4:950–957.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification.

Truc-vien T Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2010. Kernel-based Reranking for Named-Entity Extraction. In *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Poster*, pages 901–909. Association for Computational Linguistics.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro tool suite. In *Proceedings of LREC*, pages 2603–2607. Citeseer.

Lutz Prechelt. 1998. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

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

Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. *Proceedings of SEMEVAL*.

Manuela Speranza. 2009. The named entity recognition task at evalita 2009. In *Proceedings of EVALITA*.

R Zanoli and E Pianta. 2009. Named Entity Recognition through Redundancy Driven Classi ers. *In: Proceedings of EVALITA 2009. Reggio Emilia, Italy*.

Exploring Cross-Lingual Sense Mapping in a Multilingual Parallel Corpus

Francis Bond[1], Giulia Bonansinga[2]
[1]Linguistics and Multilingual Studies, Nanyang Technological University, Singapore
[2]Filologia, Letteratura e Linguistica, Università di Pisa
bond@ieee.org, giuliauni@gmail.com

Abstract

English. Cross-lingual approaches can make sense annotation of existing parallel corpora inexpensive, thus giving new means to improve any supervised Word Sense Disambiguation system. We compare two such approaches that can be applied to any multilingual parallel corpus, as long as large inter-linked sense inventories exist for all the languages involved.

Italiano. *La disponibilità di corpora annotati a livello semantico è cruciale nei modelli di apprendimento supervisionato per Word Sense Disambiguation. Qualsiasi corpus parallelo multilingue può essere disambiguato -almeno parzialmente- sfruttando le similarità e le differenze tra le lingue incluse, facendo ricorso a reti semantiche quali WordNet.*

1 Introduction

Cross-lingual Word Sense Disambiguation (CL-WSD) aims to automatically disambiguate a text in one language by exploiting its differences with other language(s) in a parallel corpus. Since the introduction of a dedicated task in SemEval-2013 (Lefever and Hoste, 2013), work on CL-WSD has increased, but parallel corpora have been used to this purpose for a long time; see for instance Brown et al. (1991), Gale et al. (1992), Ide et al. (2002), Ng et al. (2003) and, more recently, Chan and Ng (2005) and Khapra et al. (2011). Diab and Resnik (2002) exploit the semantic information inferred by translation correspondences in parallel corpora as a clue for WSD; Gliozzo et al. (2005) represent the milestone behind one of the approaches here evaluated, i.e. sense disambiguation exploiting the polysemic differential between two languages. As Bentivogli and Pianta (2005) pointed out, Word Sense Disambiguation (WSD) is so challenging mainly because most approaches require large amounts of high-quality sense-annotated data. Ten years later, the **knowledge acquisition bottleneck** still needs to be addressed for most languages.

Given an ambiguous word in a parallel corpus, having access to the *semantic space* (here intended as all the senses associated to its lemma) of each of its aligned translations allows one to exploit similarities and differences in the languages involved and, consequently, to make more educated guesses of the intended meaning. This simple, yet powerful, intuition can be decisive, if not in disambiguating all words, at least in reducing ambiguity and thus the human effort in annotating a whole text from scratch.

We explore two approaches of annotating a multilingual parallel corpus in English, Italian and Romanian built upon SemCor (SC) (Landes et al., 1998). We describe it in Section 2 along with a brief outline of the first approach, **sense projection (SP)**, which was pioneered by Bentivogli and Pianta (2005). In Section 3 we list the requirements and the necessary preprocessing steps common to both approaches. In Section 4 we present the second approach, **multilingual sense intersection (SI)**. Section 5 discusses the results achieved on the multilingual corpus with each method. We conclude in Section 6 anticipating future work.

2 SemCor, a corpus made multilingual by sense projection

Developed at Princeton University, SC (Landes et al., 1998) is a sense-annotated subset of the Brown Corpus of Standard American English (Kučera and Francis, 1967). SemCor includes 352 texts, each around 2,000 words long; in 186 texts all content words are annotated, while in the remaining 166 only verbs are.

MultiSemCor: Bentivogli and Pianta (2005) built an English-Italian parallel corpus by manually translating 116 texts from SC all-words component into Italian. Using the word alignment as a bridge, the Italian component was automatically sense-annotated by projection of the annotations available in English. Assuming that translations preserve the meaning of a text, if a sense-annotated source text is aligned to its translation(s), then the annotations can be transferred, as long an inter-linked sense inventory is used by all languages. In this study, a multilingual WordNet with reference to WordNet 1.6 (WN 1.6), MultiWordNet[1] (MWN) (Pianta et al., 2002), was used.

Following Bentivogli and Pianta (2005), we replicated SP on MultiSemCor (MSC) after converting all sense annotations to WordNet 3.0 (WN 3.0).

MultiSemCor+: Lupu et al. (2005) developed the Romanian SemCor (RSC) to build MultiSemCor+, which extended MSC with aligned Romanian translations. The MSC+ originally presented consists of 34 translations aligned to English (Lupu et al., 2005). Since then, the English-Romanian parallel corpus based on SC has grown, currently consisting of 81 texts (82 in the version released) (Ion, 2007) annotated following WN 3.0. Of these, 50 have Italian translations in MSC.

In conclusion, SP can bootstrap the creation of sense-annotated parallel corpora by exploiting existing resources in well-represented languages, with word alignment and connected sense inventories as the only requirements.

3 Preprocessing and requirements

Mapping to WN 3.0: As a preprocessing step, we mapped all annotations in MSC to WN 3.0. This is convenient in itself, as the corpus will be redistributed with reference to a widely used sense inventory, as comparison with related work will be easier. The English component is annotated with *sense keys*, stable across different WN versions, so the conversion was straightforward. On the sense keys alone, 95% of the WN 1.6 synsets can be correctly mapped to WN 3.0.[2] The Italian texts use an offset-based encoding that is not consistent across WN versions; fortunately, there are freely available mappings[3] inferred by exploiting both graph and non-structural information (Daudé et al., 2000; Daudé et al., 2001).

Sense inventories: Table 1 shows the coverage of WNs for our target languages. The Open Multilingual WordNet (OMW)[4] is an open-source multilingual database that connects all open WNs linked to the English WN, including Italian (Pianta et al., 2002) among the 28 languages supported (Bond and Paik, 2012; Bond and Foster, 2013).

Another valid option for the multilingual sense inventory would be BabelNet, created from the automatic integration of WN 3.0, OMW, Wikipedia and many other resources (Navigli and Ponzetto, 2012), with an estimated accuracy of 91% for the WN-Wikipedia mapping (Navigli et al., 2013). However, we chose to use OMW since we wanted to test our hypothesis on resources that were purposely built to be mapped to one another.

The Romanian WordNet (RW) was created within the BalkaNet project (Stamou et al., 2002). The current version has 59,348 synsets in its latest release (Barbu Mititelu et al., 2014). The synsets were mapped to WN 3.0 with precision of 95% (Tufiş et al., 2013).

	Synsets	Senses
English	117,659	206,978
Italian	34,728	69,824
Romanian	59,348	85,238

Table 1: Coverage of the WNs used.

Aligning RSC to MSC: RSC is not word-aligned to any component of the parallel corpus, so it fails in meeting a necessary requirement to perform sense mapping. However, as the sentence alignment is available, we attempted to align all Romanian sense-annotated words to their English and Italian counterparts. For each aligned sentence pair, we first align all candidate pairs sharing the same sense annotation. If any words are left unaligned after this step, the remaining alignments are inferred by taking into account PoS information and synset similarity scores. Suppose the first step alone has aligned all Romanian content words but one, and that the corresponding English sentence has three content words left that are candidates for the alignment. Then, the aligner computes the most likely match by looking for

[1] http://multiwordnet.fbk.eu/

[2] According to the HyperDic project: http://www.hyperdic.net/en/doc/mapping

[3] http://www.talp.upc.edu/index.php/technology/tools/45-textual-processing-tools/98-wordnet-mappings/

[4] http://compling.hss.ntu.edu.sg/omw/summx.html

PoS correspondence and for higher proximity in the WN network, by looking at a combination of the *path similarity score* and the *shortest path distance*. This latter alignment strategy (the only possible source of errors) achieved 97% precision on a small sample (12%) of the alignments found.

4 Multilingual Sense Intersection

Unlike SP, SI does not require any of the texts in a parallel corpus to be sense-annotated, so it can be applied to a wider range of existing resources. Its logical foundation is in that a polysemous word in a language is likely to be translated in different words in other languages, so the comparison with the semantic space of each translation should help select the sense actually intended. Consider, for instance, the problem of disambiguating the English word *administration* in Example 1.

(1) EN *The jury praised the administration and operation of the Atlanta Police Department.*

IT *Il jury ha elogiato l'amministrazione e l'operato del Dipartimento di Polizia di Atlanta.*

RO *Juriul a lăudat administrarea şi conducerea Secţ iei de poliţie din Atlanta.*

Given the alignments, we can retrieve the set of synsets associated with the lemmas in the Italian and Romanian translations. Figure 1 shows how the intersection helps detecting the correct sense, which is the only one shared by all the lemmas.

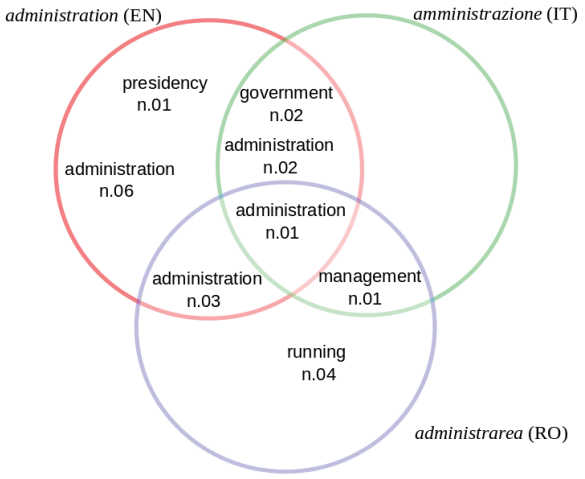


Figure 1: Disambiguation via SI

Most often, however, such a comparison will only partially reduce the ambiguity, especially as such a fine-grained sense inventory as WN is used. Yet, other approaches (employment of human annotators, or recourse to baselines) can be applied in a second phase to solve the disambiguation task, once it has been simplified.

The algorithm disambiguates one side of our multilingual parallel corpus at a time, having as target all texts aligned with at least one other component.[5] Table 2 displays the basic statistics of each corpus and, for the sake of clarity, the number of words to be annotated (target words) before the migration to WN 3.0, as the changes in the WN structure do not set ideal conditions for a meaningful comparison with previous work.

We use sense frequency statistics (SFS) whenever the target word is not fully disambiguated. These were calculated over all texts in the corpus **except** the one being annotated.

	#texts	Tokens	Target words	After mapping
EN	116	258,499	119,802	118,750
IT	116	268,905	92,420	92,022
RO	82	175,603	48,634	48,364

Table 2: Statistics for each text in the multilingual parallel corpus.

%	EN	IT	RO
Disambiguated	27.15	30.92	36.67
MFS-Subset	34.39	26.51	12.89
MFS-Overlap	13.59	26.69	50.45
No alignment	24.14	12.08	-
No match	0.67	0.65	-
No synset found	0.05	3.14	-

Table 3: Distribution of SI outcomes.

Algorithm: Given an ambiguous target word, each of its aligned translations in the parallel sentences contributes to the disambiguation process by bringing in all its 'set of senses' retrieved from the inter-linked sense inventory.

Intersection is then performed over each non-empty set retrieved. If the *overlap* only consists of one sense, then the target word is `Disambiguated` (see Table 3). If the overlap contains more than one sense, then it is further intersected with the set of most frequent senses available for the target lemma. If resorting to MFS statistics leads to an overlap containing one sense, the word is disambiguated (`MFS-Subset`); if the overlap still results in more than one sense, the

[5] With the exception of the English corpus, which we have considered made of the 116 texts included in MSC.

Method	English		Italian		Romanian	
	Precision	Coverage	Precision	Coverage	Precision	Coverage
MFS (baseline)	**0.761**	**0.998**	0.599	0.999	0.531	1
SP	-	-	**0.971**	**0.927**	**0.903**	**1**
SP (Bentivogli & Pianta)	-	-	0.879	0.764	-	-
3-way SI	0.750	0.778	0.653	0.915	0.590	1

Table 4: Comparison of the results scored with SP, SI and MFS baseline.

most frequent one among the ones left is selected (`MFS-Overlap`). In the rare case in which no other language contributes to disambiguate, we assign the current target lemma its MFS. Disambiguation also fails when no match, synset or alignment is found. See also Table 3 for the distribution of all of the possible scenarios that may emerge.

5 Evaluation and discussion

Table 4 shows the precision and coverage scores achieved with the approaches here analyzed, along with the Most Frequent Sense (MFS) baseline. We report the original results for SP (Bentivogli and Pianta, 2005) and ours after the mapping to WN 3.0; we evaluate on different figures (see Table 2) as a part of the original annotations was lost in the mapping process. We performed SP also on the current release of RSC for completeness.

Coverage is overall reasonably high for all languages with SI and very high with the baseline. On the other hand, the precision achieved resorting to SFS is significantly lower for Italian, which makes more valuable the not very high score obtained by SI. Average ambiguity reduction is 54% (EN), 53% (IT) and 55% (RO).

Although SI and MFS perform comparably, we remind that SFS were computed on the same corpus, which is also not extremely large. Thus, we would expect MFS to compare at least slightly worse in more general cases (unfortunately, external statistics are hard to come by). This would make SI a valid and inexpensive cross-lingual disambiguation approach. We also performed 2-way intersection for each corpus pair. We find a slight decrease in precision (of 0.01 to 0.03) compared to the three-way intersection, depending on the corpus. While further restricting the semantic space does help in reducing ambiguity, the improvement is not striking. According to our error analysis, this is corpus-dependent, as the manually assigned correct senses against which we evaluate are very specific. Instead, as the WNs vary largely in coverage, senses found by intersection, though actually shared in all languages, are close, but not quite the same, to the very specific ones selected by the human annotator. In conclusion, coarse-grained evaluation would give a higher score, and in general the senses found by intersection would be just good enough in most cases. Also, as Italian and Romanian are quite similar, we would expect more differences if we added a language from a different language family.

6 Conclusions

To our knowledge, this is the first attempt to disambiguate a parallel corpus by using multilingual SI. The more languages are considered, the more ambiguity should be reduced and the better SI is expected to perform. In future work, we plan to include the Japanese SemCor (Bond et al., 2012) to test our hypothesis that translations from a different language family will discriminate further. We also plan to use a different parallel corpus built on open translations of *The Adventure of the Speckled Band* by Sir Arthur Conan Doyle. We will also try to calculate SFS from untagged text, following McCarthy and Carroll (2003).

Furthermore, we are investigating alternative ways to solve the ambiguity left whenever SI does not lead to a single synset; for instance, we plan to apply some implementation of Lesk (Lesk, 1986) on the subset found by SI. Finally, we aim to port to WN 3.0 the sense clustering carried out by Navigli (2006) to perform a coarse-grained evaluation, which would ignore minor sense distinctions. An initial comparison with Babelfly (Moro et al., 2014) would certainly be enlightening as well.

All data and scripts derived by our work will be made available, except for those derived from RSC, as its license currently forbids it.

Acknowledgments

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2010-5094-7.

References

Verginica Barbu Mititelu, Stefan Daniel Dumitrescu, and Dan Tufiş, 2014. *Proceedings of the Seventh Global Wordnet Conference*, chapter News about the Romanian Wordnet, pages 268–275. ACL.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(03):247, September.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362. Association for Computational Linguistics.

Francis Bond and Kyonghee Paik. 2012. A Survey of WordNets and their Licenses. In *GWC 2012*, pages 64–71.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 56–63.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the ACL*, Morristown, NJ. Morristown, NJ: Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *AAAI*, volume 5, pages 1037–1042.

Jordi Daudé, Lluís Padró, and German Rigau. 2001. A complete WN1.5 to WN1.6 mapping. In *Proceedings of NAACL Workshop" WordNet and Other Lexical Resources: Applications, Extensions and Customizations". Pittsburg, PA*.

Jordi Daudé, Lluís Padró, and German Rigau. 2000. Mapping wordnets using structural information. In *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000).*, Hong Kong.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Stroudsburg, PA, USA. Association for Computational Linguistics.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods.

Alfio Massimiliano Gliozzo, Marcello Ranieri, and Carlo Strapparava. 2005. Crossing parallel corpora and multilingual lexical databases for WSD. In *Computational Linguistics and Intelligent Text Processing*, pages 242–245. Springer.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 61–66. Association for Computational Linguistics.

Radu Ion. 2007. *Metode de dezambiguizare semantica automata. Aplicat ii pentru limbile englezas i romana ("Word Sense Disambiguation methods applied to English and Romanian")*. Ph.D. thesis, Research Institute for Artificial Intelligence (RACAI), Romanian Academy, Bucharest.

Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011. Together we can: Bilingual bootstrapping for wsd. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 561–569, Stroudsburg, PA, USA. Association for Computational Linguistics.

Henry Kučera and W. Nelson Francis. 1967. Computational analysis of present-day American English.

Shari Landes, Claudia Leacock, and Randee I Tengi. 1998. Building semantic concordances. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 199–216. MIT Press, Cambridge, MA.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. *Proc. of SemEval*, pages 158–166.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Monica Lupu, Diana Trandabat, and Maria Husarciuc. 2005. A Romanian SemCor aligned to the English and Italian MultiSemCor. In *1st ROMANCE FrameNet Workshop at EUROLAN*, pages 20–27.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

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.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 455–462. Association for Computational Linguistics.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an Aligned Multilingual Database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, and Maria Grigoriadou. 2002. Balkanet: A multilingual semantic network for the balkan languages. *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25.

Dan Tufiş, Verginica Barbu Mititelu, Dan Ştefănescu, and Radu Ion. 2013. The Romanian wordnet in a nutshell. *Language Resources and Evaluation*, 47(4):1305–1314, December.

ISACCO: a corpus for investigating spoken and written language development in Italian school–age children

Dominique Brunato, Felice Dell'Orletta
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it

Abstract

English. We present ISACCO (Italian school–age children corpus)[1], a new corpus of oral and written retellings of Italian-speaking children attending the primary school. All texts were digitalized and automatically enriched with linguistic information allowing preliminary explorations based on NLP features. Written retellings were also manually annotated with a typology of linguistic errors. The resource is conceived to support research and computational modeling of "later language acquisition", with an emphasis for comparative assessment of oral and written language skills across early school grades.

Italiano. *Presentiamo ISACCO (Italian school–age children corpus), un nuovo corpus di riassunti orali e scritti prodotti da bambini italiani della scuola primaria. Tutti i testi sono stati digitalizzati e arricchiti automaticamente con informazione linguistica per consentire esplorazioni preliminari basate su caratteristiche estratte con strumenti di TAL. I riassunti scritti sono stati anche annotati a mano con una tipologia di errori linguistici. La risorsa è pensata per lo studio e la definizione di modelli computazionali degli stadi più avanzati del processo di acquisizione linguistica, con un'enfasi per la valutazione comparativa delle abilità linguistiche orali e scritte nei primi anni scolastici.*

1 Introduction

The use of naturalistic data to investigate child language features and development has a well-established tradition in L1 acquisition research. The most notable example is the CHILDES database (MacWhinney, 2000), which contains transcripts of spoken interactions involving children of different ages for over 25 languages, Italian included. Yet, CHILDES data refer especially to preschool children, with only a minor section dedicated to their older mates, thus making this resource less adequate for studying how language skills evolve during early schooling. The rapid and remarkable changes children's language undergoes before age five justify the amount of research for the earliest stages of acquisition. However, over the last two decades also "later language acquisition" has gained increasing interest (Tolchinsky, 2004), prompted by the awareness that "becoming a native speaker is a rapid and highly efficient process, but becoming a proficient speaker takes a long time" (Berman, 2004). Indeed, under explicit teaching language keeps growing through school–age years in a way that affects all domains and modalities (Koutsoftas, 2013). Regarding the methodological approach to inspect children's data, more attention has been recently paid to text analysis techniques drawn from computational linguistics and Natural Language Processing (NLP). The use of a statistical parser is reported e.g. by Sagae et al. (2005) and Lu (2009) to automate sophisticated measures of syntactic development, reaching performances comparable to those obtained by manual annotation. Computational methods are also employed in diagnostic settings, e.g. to identify markers of Autism Spectrum Disorders in children's speech by integrating features from automatic morpho–syntactic and syntactic annotation (Prud'hommeaux and Roark, 2011), as well as metrics of semantic similarity (Rouhizadeh et al., 2015). Despite the focus of this paper is on the resource, we will also present preliminary analyses aiming at showing how a NLP perspective applied to a corpus like ISACCO can

[1]The resource will be made publicly available at: http://www.italianlp.it/software–data.

serve as the starting point to conduct computational explorations at multiple levels, which may become particularly useful in view of their applicability to large–scale corpora. It should be possible to test the effect of the diamesic variation on the linguistic complexity of children's texts and to assess changes across schooling levels (cf. section 3.1). The same can be done with respect to the "content", to evaluate whether these variables affect text comprehension and recall. To this aim, the output of an ontology learning system can provide a mean to compare the quantity of 'matched' ideas between the child's retelling and the content of the heard story (cf. section 3.2), so that to identify patterns of typical development to be used for comparison e.g. in clinical settings, with children showing atypical language development.

2 The corpus

2.1 Participants

Fifty-six TD (typically developing) children from the 2nd to the 4thgrade of primary school participated in the task. They were all recruited from a public primary school located in the suburbs of Pisa and examined in the last month of the school year. All children were Italian monolingual speakers, except from two, who were also included in the survey since they had no significant exposure to other languages. Details of the sample group are given in Table 1.

Grade	Male	Female	Age Mean (SD)
Second	11	8	8.1 m (3.6 m)
Third	10	11	9.0 m (5.6 m)
Fourth	9	7	10.0 m (4.2 m)

Table 1: Children sample group (SD=Standard deviation; m=months).

2.2 Methodology

To collect ISACCO, we inspired to the work of (Silva et al., 2010) for Spanish, who assessed children's oral and written performance in a retelling task by exposing them to the same story to avoid a possible text bias. Differently from them, we excluded the 1stgrade pupils, following the teachers' suggestions pointing out that free written retelling is usually introduced in the curriculum by the end of the second year. We then selected a narrative text from a 3rdgrade book, which was intended to be not too challenging for the youngest nor too easy for the oldest group[2]. Children were tested in two sessions, with a gap of two weeks, so that to prevent memory bias. The first session was devoted to collect oral productions; this was done by reading the story aloud once to the whole class and repeating it again to a restricted group of students, which was randomnly chosen by teachers, while their mates carried out another activity related to the story (e.g. drawing a picture). Each selected child was tested individually, in a quiet room, and after hearing the story again was asked to retell it to the experimenter. All retellings were recorded and then transcribed, as detailed in Section 2.3.

	Oral retellings	
Grade	**Number of texts**	**Number of tokens**
Second	19	2.029
Third	21	2.994
Fourth	16	2.406
Tot	56	7.429
	Written retellings	
Second	43	4.508
Third	44	4.984
Fourth	38	4.417
Tot	125	13.909

Table 2: Corpus of oral and written retellings.

In the second session, the same story was read again to the whole class and this time all students produced a written retelling. No limit of time was given and they were left free to write in capital letters or italics. Although for the purpose of comparative analysis only the writings of the 56 children tested in the first session were needed, we digitalized all written retellings; such a corpus offers indeed valuable material for research on writing development with a view to its computational modeling.

2.3 Oral data transcription

Children's oral retellings were manually trascribed adding some "natural punctuations" (Powers, 2005) (i.e. periods and commas) according to speech pauses and intonations, to identify major sentence boundaries. These "row" transcripts were then enriched with additional "xml-style" labels to annotate typical phenomena of spoken language (e.g. false starts, disfluencies), as defined in the following tagset:

- tag *fs*: to mark a false start (covering both a single or a sequence of words).

[2]The story is titled "La statua nel parco", by Roberto Piumini.

- tag *rip*: to mark a repeated word. It has the attribute *number* for the number of repetitions made by the child;
- tag *int*: to mark a long interruption (e.g. when the child did not recall the story)

2.4 Linguistic annotation of errors

After being digitalized, written texts were manually annotated with typologies of linguistic errors, following the tagset defined by Barbagli et al. (2015). Errors are distinguished into three macro–areas, according to the domain of linguistic knowledge affected, i.e.: ortography, grammar and lexicon. Each macro–class is further sub-divided into more classes codifying the linguistic category and the target modification for the misused units. Table 3 reports the error tagset and the quantitative distributions for each category according to the school grade.

3 Preliminary explorations of the corpus

This section presents preliminary explorations comparing oral and written retellings with respect to both linguistic structure and content. All analyses were conducted by comparing the statistical distribution of linguistic and lexico–semantic features automatically extracted from the corpora by means of NLP tools. Specifically, all texts were automatically tagged with the part–of–speech tagger described in Dell'Orletta (2009) and dependency–parsed by the DeSR parser (Attardi, 2006) using Support Vector Machines as learning algorithm. It goes without saying that the typology of texts under examination is particularly challenging for *general-purpose* text analysis tools; this is not only due to the features of spoken language but also to missing punctuation (especially in the 2ndgrade writings), which already impacts on the coarsest levels of text analysis, i.e. sentence splitting. Although we plan to evaluate more in detail the impact of these non–standard patterns on linguistic annotation, we believe that some features extracted from linguistically annotated texts are robust enough to offer a first insight into the linguistic structure of children's texts according to age and modality, as well as with respect to the content.

3.1 First results on linguistic structure

Table 4 shows a subset of linguistic features for which the average difference value between oral and written samples was significant[3]. Starting from superficial features, it emerges that oral retellings are on average longer than the written ones ([1]); in line with previous findings in the literature, such a difference may be due to the heavy cognitive demands initially posed by writing affecting memory and causing a loss of information. Oral retellings also tend to exhibit slightly shorter words. This finding can be elaborated by looking at the POS distribution, where we find a greater distribution of words belonging to functional categories (particularly, Pronouns [7] and Conjunctions [4,8]) in oral than in written texts. Such a difference affects lexical density [10], which is higher in writing, as typically reported for adults (Halliday, 1989). Coming to the grammatical structure, when children retell the story orally they tend to produce more complex sentences, as suggested by the predominance of conjunctions, especially subordinating ones. Such a distribution, together with that of adverbs [3], can also give some indications on the way modality affects children's language at discourse structure, which appears less cohesive when they write rather than when they retell the story verbally. Last, it is interesting to note that a well-known factor of syntactic complexity, i.e. the length of dependency links [11], is not significantly influenced by the way children retell the story.

Linguistic Feature	Oral	Written	Diff.
[1] Text length (in token)	125.11	109.46	**-15.64**
[2] Word length	4.54	4.55	**+0.01**$*$
[3] Adverbs	8.62	4.86	**-3.77**$*$
[4] Coordinating Conj.	6.14	4.83	**-1.31**$*$
[5] Determiners	10.88	14.52	**+3.64**$*$
[6] Nouns	21.80	28.50	**+6.70**$*$
[7] Pronouns	6.70	4.79	**-1.91**$*$
[8] Subordinating Conj.	1.56	0.96	**-0.96**
[9] Verbs	15.51	14.26	**-1.25**$*$
[10] Lexical density	0.539	0.552	**0.012**
[11] Length of depend. links	2.40	2.42	0.02

Table 4: Linguistic features. Significant differences at $p < 0.05$ are bolded, those at $p < 0.005$ are also marked with $*$.

3.2 Analysis of the content

For the analysis of the corpus with respect to the content, we relied on $T2K^2$ (Text–to–Knowledge), a suite of tools based on NLP modules for automatically extracting domain–specific

[3]Wilcoxon's signed rank test was applied for statistical analysis because of the small number of subjects.

		II grade		III grade		IV grade	
Category	**Target modification**	Freq.%	Abs. Value	Freq.%	Abs. Value	Freq.%	Abs. Value
Orthography							
Consonant doubling	Omission	10.59	(45)	1.40	(3)	5.52	(8)
	Excess	2.35	(10)	1.40	(3)	2.07	(3)
Use of *H*	Omission	0.71	(3)	0.93	(2)	0.00	(0)
	Excess	0.24	(1)	0.00	(0)	0.00	(0)
Monosyllabic words	Mispelling of stressed monosyllabic words	2.35	(10)	6.51	(14)	1.38	(2)
	Mispelling of *po'* (e.g. pó or po)	3.76	(16)	4.65	(10)	4.14	(6)
Apostrophe	Misuse	3.76	(16)	0.93	(2)	0.69	(1)
Other		32.94	(140)	33.02	(71)	40.69	(59)
Grammar							
Verbs	Use of tenses	24.00	(102)	15.35	(33)	12.00	(12)
	Use of modes	0.00	(0)	0.00	(0)	0.69	(1)
	Subject-verb agreement	1.88	(8)	6.51	(14)	5.52	(8)
Prepositions	Misuse	1.88	(8)	3.26	(7)	1.38	(2)
	Omission or Excess	1.41	(6)	1.47	(1)	1.38	(2)
Pronouns	Misuse	0.24	(1)	0.47	(1)	1.38	(2)
	Omission	0.24	(1)	0.47	(1)	1.38	(2)
	Excess	0.240	(1)	0.47	(1)	1.38	(2)
	Misuse of relative pronoun	0.24	(1)	0.47	(1)	0.69	(1)
Conjunctions	Misuse	0.24	(1)	0.47	(1)	2.38	(2)
Other		8.00	(34)	11.63	(25)	10.34	(16)
Lexicon							
Vocabulary	Misuse of terms	4.94	(21)	11.63	(25)	11.03	(16)

Table 3: Linguistic errors tagset and quantitative distributions in written retellings.

knowledge from a corpus (Dell'Orletta et al., 2014). Following the assumption that the most relevant concepts of a text have a linguistic counterpart, which is typically conveyed by single and multi–word nominal terms, the process of terminology extraction can be seen as the first step to access the knowledge contained in text. We thus applied the term extraction functionalities of $T2K^2$ both to the original story and to the corpus of children's retellings; the latter was first distinguished into the oral and written sub–corpora (each one taken as a whole) and then by considering each school–grade separately for both modalities. As shown by the excerpt of the output in Table 5, there is a strict correspondence between the ten most salient concepts characterizing the original story and those reported by children, independently from modality. Such findings were also replicated when we analyzed separately the oral and written retellings of the 2nd, 3rd and 4thgrade students, thus suggesting that from age seven children have already mastered the ability to grasp, retain and organize the main concepts of a narrative text like the one here proposed. This analysis, complemented with first data of linguistic profiling, seems to imply that the effect of modality is more relevant at the level of linguistic structure.

Original story	**Oral corpus**	**Written Corpus**
mappamondo	mano	statua
pietra	statua	mano
terra	mappamondo	mappamondo
mano	rondine	geografo
rondine	geografo	rondine
Geografo	terra	parco
statua	primavera	primavera
busto	nido	terra
parco	ragazzo	nido
primavera	giorno	ragazzo

Table 5: Excerpt of automatically extracted domain–terminology.

4 Conclusion

We presented ISACCO, a new resource for the Italian language containing oral and written retellings of children attending the primary school. We showed the potentiality of NLP-based analyses to inspect child language features, both with respect to linguistic and content structure, as well as in relation to diachronic and diamesic variations. Ongoing work is devoted to enlarge the corpus, also in a longitudinal perspective, to elaborate a qualitative analysis of linguistic errors by also looking comparatively at other learner corpora, and to evaluate the impact of child language features on standard linguistic annotation tools.

Acknowledgments

We would like to thank the headmaster of the primary school "Vasco Morroni" of Ghezzano (Pisa), the teachers and all the children taking part in the survey for their contribution in this research.

References

B. MacWhinney. 2000. The CHILDES Project: Tools for Analyzing Talk. *3rd edition. Lawrence Erlbaum Associates, 2000.*

L. Tolchinksy. 2004. The nature and scope of later language development. In *R.A. Berman (Ed.), Language Development across Childhood and Adolescence. Amsterdam: John Benjamins Publishing Company.*

R. Berman. 2004. Between emergence and mastery: the long developmental route of language acquisition. In *R.A. Berman (Ed.), Language Development across Childhood and Adolescence. Amsterdam: John Benjamins Publishing Company.*

A. D. Koutsoftas. 2013. School–age language development: Application of the five domains of language across four modalities. In *N. Capone-Singleton and B.B. Shulman (Eds.), Language development: Foundations, processes, and clinical applications, Second Edition*, pp. 215–229, Burli, April 2013.

K. Sagae, A. Lavie and B. MacWhinney. 2005. Automatic measurement of syntactic development in child language. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pp. 197–204.

X. Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1), 3-28.

E. T. Prud'hommeaux and B. Roark. 2011. Classification of atypical language in autism. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics.*

M. Rouhizadeh, R. Sproat, J. van Santen. 2015. Similarity Measures for Quantifying Restrictive and Repetitive Behavior in Conversations of Autistic Children. *Computational Linguistics and Clinical Psychology Workshop (CLPsych), NAACL, 2015, Denver, CO.*

M. Silva, V. Sánchez Abchi, A. Borzone. 2010. Subordinate clauses usage and assessment of syntactic maturity: A comparison of oral and written retellings in beginning writers. *Journal of Writing Research*, 2(1):47–64.

W. R. Powers. 2005. Transcription techniques for the spoken word. *Lanham, MD: Altamira Press.*

A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi. 2015 (Submitted). CItA: un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1 Annotato con Errori.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.

G. Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06)*, New York City, New York:166–170.

M. A.K. Halliday. 1989. Spoken and Written Language. *Oxford: Oxford University Press.*

F. Dell'Orletta, G. Venturi, A. Cimino, S. Montemagni. 2014. T2K: a System for Automatically Extracting and Organizing Knowledge from Texts. In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2062–2070, 26-31 May, Reykjavik, Iceland.

Inconsistencies Detection in Bipolar Entailment Graphs

Elena Cabrio[1], **Serena Villata**[2]
[2] CNRS, [1,2]University of Nice Sophia Antipolis, France
elena.cabrio@unice.fr; serena.villata@cnrs.fr

Abstract

English. In the latest years, a number of real world applications have underlined the need to move from Textual Entailment (TE) pairs to TE graphs where pairs are no more independent. Moving from single pairs to a graph has the advantage of providing an overall view of the issue discussed in the text, but this may lead to possible inconsistencies due to the combination of the TE pairs into a unique graph. In this paper, we adopt *argumentation theory* to support human annotators in detecting the possible sources of inconsistencies.

Italiano. *Negli ultimi anni, in svariate applicazioni sta sorgendo la necessità di passare da coppie di Textual Entailment (TE) a grafi di TE, in cui le coppie sono interconnesse. Il vantaggio dei grafi di TE è di fornire una visione globale del soggetto di cui si sta discutendo nel testo. Allo stesso tempo, questo può generare inconsistenze dovute all'integrazione di più coppie di TE in un unico grafo. In questo articolo, ci basiamo sulla teoria dell'argomentazione per supportare gli annotatori nell'individuare le possibili fonti di inconsistenze.*

1 Introduction

A Textual Entailment (TE) system (Dagan et al., 2009) automatically assigns to independent pairs of two textual fragments either an *entailment* or a *contradiction* relation. However, in some real world scenarios like analyzing costumer reviews about a service or product, these pairs cannot be considered as independent. For instance, all the reviews about a certain service need to be collected into a single graph, to understand the overall problems/merits of the service. The combination of TE pairs into a unique graph may generate *inconsistencies* due to the wrong relation assignment by the TE system, which could not have been identified if TE pairs were considered independently. The detection of such inconsistencies is usually left to human annotators, which later correct them. The need of processing such graphs to support annotators is therefore of crucial importance, particularly when dealing with big amounts of data. Our research question is *How to support annotators in detecting inconsistencies in TE graphs?*

The term *entailment graph* has been introduced by (Berant et al., 2010) as a structure to model entailment relations between propositional templates. Differently, in this paper we consider *bipolar entailment graphs* (BEGs), where two kinds of edges are considered, i.e., entailment and contradiction, to reason over the graph consistency.

We answer the research question by adopting *abstract argumentation theory* (Dung, 1995), a reasoning framework used to detect and solve inconsistencies in the so-called *argumentation graphs*, where nodes are called *arguments*, and edges represent a *conflict* relation. Argumentation semantics allows to compute *consistent* sets of arguments, given the conflicts among them.

We define the BEGincs (BEG-Inconsistencies) framework, which translates a BEG into an argumentation graph. It then provides to the annotators sets of arguments, following argumentation semantics, that are supposed to be consistent. If it is not the case, the TE system wrongly assigned some relations. Moving from single pairs to an overall graph allows for the detection of inconsistencies otherwise undiscovered. BEGincs does not identify the precise relation causing the inconsistency, but providing annotators with the consistent arguments sets, they are supported in narrowing the causes of inconsistency.

2 BEGincs framework

TE is a directional relation between two textual fragments. In various real world scenarios, these pairs cannot be considered as independent, and they need to be collected into a single graph. We define therefore a new framework involving *entailment graphs*, where pairs of textual fragments connected by semantic relations are also part of a graph that provides an overall view of the statements' interactions (*bipolar entailment graphs*).

Definition 1. *A bipolar entailment graph is a tuple* $BEG = \langle T, E, C \rangle$ *where* T *is a set of text fragments,* $E \subseteq T \times T$ *is an entailment relation between text fragments, and* $C \subseteq T \times T$ *is a contradiction relation between text fragments.*

This opens new challenges for TE, that originally considers the pairs as "self-contained" (i.e., the meaning of one text has to be derived from the meaning of the other). One challenge consists in checking BEGs to identify possible inconsistencies due to wrong relation assignments by the TE system. Figure 1 shows the architecture of the BEGincs framework to support human annotators in detecting inconsistencies in TE graphs.

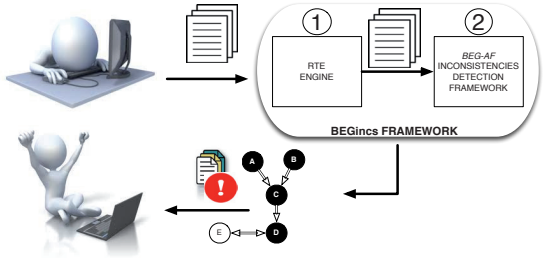


Figure 1: The BEGincs framework architecture.

Annotators provide the dataset to be checked as input of the BEGincs framework, which consists of two main modules: *(1)* a TE module, takes as input the dataset of text fragments, and returns the pairs annotated with the entailment or contradiction relations; and *(2)* a BEG-AF Inconsistencies Detection module, which translates the received BEGs into an argumentation framework such that argumentation semantics can be applied to retrieve consistent sets of arguments. The BEGincs framework returns through a user interface the starting BEGs highlighted with the consistent sets of text fragments. Checking them, annotators are able to detect errors in the annotation produced by the TE module (they will find inconsistent arguments in the returned sets), and correct the erroneous pairs.

2.1 Argumentation theory

An abstract argumentation framework (AF) (Dung, 1995) represents conflicts among elements called *arguments*. It is based on a binary *attack* relation among them, whose role is determined only by their relation with the other arguments. An AF encodes, through the *attack* relation, the existing conflicts within a set of arguments. It identifies then the conflict outcomes, i.e. which arguments should be accepted ("they survive the conflict") and which arguments should be rejected, according to some reasonable criterion. (Dung, 1995) presents several acceptability semantics that produce zero, one, or several *consistent* sets of accepted arguments. Such set of accepted arguments does not contain an argument conflicting with another argument in the set (*conflict free*). Following from this notion, an *admissible* set of arguments is required to be both internally coherent (*conflict-free*) and able to defend its elements. In BEGincs, we adopt admissibility based semantics. Roughly, an argument is accepted if all the arguments attacking it are rejected, and it is rejected if there is at least an argument attacking it which is accepted. The sets of accepted arguments computed using an acceptability semantics are called *extensions*, and the addition of another argument from outside the set will make it *inconsistent*.

2.2 Inconsistencies detection

To reuse abstract argumentation results and semantics for inconsistencies detection, we need to represent both the entailment and the contradiction relations of the bipolar entailment graph under the form of *attacks* between abstract arguments in an argumentation graph (Definition 2).

Definition 2. *A BEG-based argumentation framework is a tuple* $\langle A, \Rightarrow, \Leftrightarrow \rangle$ *where* A *is a set of text fragments called arguments,* $\Rightarrow$ *is a binary entailment relation on* A *(*$\Rightarrow \subseteq A \times A$*), and* $\Leftrightarrow$ *is a binary contradiction relation on* A *(*$\Leftrightarrow \subseteq A \times A$*). The set of arguments is* $\{a, b, \ldots \in A\}$.

BEG-AFs' consistent sets of arguments contain the text fragments that do not conflict with other fragments in the set (they are coherent). BEGincs uses the consistent sets of arguments computed following admissibility based argumentation semantics to support annotators in detecting inconsistencies. We need then to define the semantics of the entailment and contradiction relations in the

BEG-based argumentation framework (i.e. the behavior these relations have to satisfy in terms of conflict, since the only relation between arguments in abstract argumentation is the conflict relation).

Example 1.
T1: Natural gas vehicles run on natural gas, so emit significant amounts of greenhouse gases into the atmosphere, albeit smaller amounts than gasoline-fueled cars. To combat global warming, we should be focusing our energies and investments solely on 0-emission electric vehicles.
H: On the surface, natural gas cars seem alright, but the topic becomes a bit different when they are competing against zero emission alternatives (e.g. electric cars).

In Example 1, the text ($T1$) entails the hypothesis (H), i.e., $T1 \Rightarrow H$. Entailment is a directional relation (Dagan et al., 2009), that holds if the meaning of H can be inferred from the meaning of T, as interpreted by a typical language user. In the pair, T is more specific than H (i.e., the more specific argument entails the more general one). In the argumentation setting, we have to reason over this feature to identify which constraints it poses in terms of conflicts among the text fragments. In particular, the following constraints emerge from the entailment relation: assuming *T entails H* holds, then *(i)* if there is a text fragment T_1 which contradicts H (negative TE) *then* T_1 contradicts also T ($T \equiv T_1$ does not entail $H \equiv T$), and *(ii)* if there is a text fragment T_2 which contradicts T *then* T_2 does not necessary contradict H too. These two constraints hold when a TE pair is inserted into an entailment graph. As a consequence, from the arguments' acceptance viewpoint: given that $T \Rightarrow H$, every time argument H is rejected, argument T is rejected too. We model the *entailment* relation such that, given that T entails H, T is accepted only if H is accepted too (Definit. 3)[1].

Definition 3. *Given a BEG-based argumentation framework $\langle A, \Rightarrow, \Leftrightarrow \rangle$, a translated BEG-based argumentation framework (BEG-AF) is a tuple $\langle \mathcal{A}, \longmapsto \rangle$ such that the set of arguments $\mathcal{A}$ is $\{a, b, \ldots \in A\} \cup \{X_{a,b}, Y_{a,b}, E_{a,b} \mid a, b \in A\}$, where $X_{a,b}, Y_{a,b}$ are the dummy arguments corresponding to the contradiction relation and $E_{a,b}$ is the dummy argument corresponding to the entailment relation, and $\longmapsto$ is a binary conflict relation over $\mathcal{A}$ such that: $b \longmapsto E_{a,b} \longmapsto a$ iff $a \Rightarrow b$.*

We have now to define the semantics of the *contradiction* relation (i.e., negative TE) in BEGs, see Example 2. (Marneffe et al., 2008) claims that contradiction occurs when two sentences *i)* are extremely unlikely to be true simultaneously, and *ii)* involve the same event. Starting from these considerations, the following constraint holds for the contradiction pairs: T and H conflict with each other (i.e. it is not possible to have both in a coherent and consistent set of arguments).

Example 2.
T2: Natural gas is the cleanest transportation fuel available today. If we want to immediately begin the process of significantly reducing greenhouse gas emissions, natural gas can help now. Other alternatives cannot be pursued as quickly.
H: On the surface, natural gas cars seem alright, but the topic becomes a bit different when they are competing against zero emission alternatives (e.g. electric cars).

Definition 4 models contradiction in BEG-AFs. The attack in (Dung, 1995) is directed from an argument to another argument while our contradiction leads to a cycle of attacks.

Definition 4. *Given a BEG-based argumentation framework $\langle A, \Rightarrow, \Leftrightarrow \rangle$, a BEG-AF is a tuple $\langle \mathcal{A}, \longmapsto \rangle$ such that $\mathcal{A}$ is the set of arguments, and $\longmapsto$ is a binary conflict relation over $\mathcal{A}$ such that:*
$a \longmapsto X_{a,b} \longmapsto Y_{a,b} \longmapsto b$, and
$b \longmapsto X_{b,a} \longmapsto Y_{b,a} \longmapsto a$, iff $a \Leftrightarrow b$.

Figure 2 summarizes the translation procedure, which is the core of our framework. We start with a BEG consisting of three text fragments (i.e., arguments A, B, C) from Ex. 1 and 2, where T1 is A, T2 is B, and H is C. The BEG is then translated into a BEG-AF where dummy arguments are introduced to express the semantics of the relations of entailment and contradiction, e.g., dummy argument $E_{A,C}$ represents the relation *A entails C* in the BEG-AF. The only relation allowed in a BEG-AF is the conflict relation $\longmapsto$. Therefore we have that a BEG-AF is a standard abstract AF, and we can apply admissibility based argumentation semantics to retrieve consistent sets of arguments. Acceptability semantics return the extension of the BEG-AF (i.e., the black nodes in Fig. 2), where arguments C,A are accepted, and dummy arguments are filtered out from the set of accepted ones.

We prove now that our BEG-AF actually satisfies the semantics of the entailment relation.

[1] See (Cabrio and Villata, 2013) for a comparison of the *entailment* wrt the *support* relation (Boella et al., 2010).

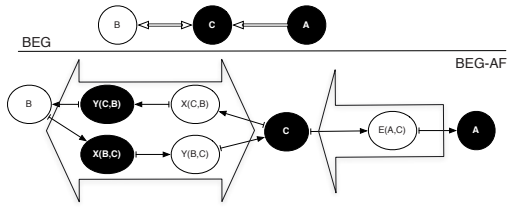


Figure 2: Translation from a BEG to a BEG-AF.

Proposition 1 (Semantics of *entailment*)**.** *Given a BEG-AF, if it holds that $T \Rightarrow H$ and text fragment T is accepted, then fragment H is accepted too.*

Proof. We prove the contrapositive. If it holds that $T \Rightarrow H$ and text fragment H is not accepted, then text fragment T is not accepted. Assume that $T \Rightarrow H$ and assume that argument H is not accepted, then dummy argument $E_{T,H}$ is accepted. Consequently, T is not accepted, i.e., rejected. □

We need to add two nodes, i.e., dummy arguments $X_{a,b}$ and $Y_{a,b}$, to represent a contradiction while we only need one node, i.e., dummy argument $E_{a,b}$, to represent entailment, since preserving the semantics of a contradiction holding between two text fragments means that the two text fragments cannot be together in a consistent set of arguments. To avoid the two being both accepted, we need to introduce two dummy arguments so that: a (accepted) $\longmapsto$ $X_{a,b}$ (rejected), $X_{a,b} \longmapsto Y_{a,b}$ (accepted), and $Y_{a,b} \longmapsto b$ (rejected). In this way, if a is accepted then b is rejected, and viceversa. A unique dummy argument between a and b would not ensure such behavior.

Existing works combine NLP and argumentation theory, e.g. (Chesñevar and Maguitman, 2004; Carenini and Moore, 2006; Wyner and van Engers, 2010; Feng and Hirst, 2011) with different purposes. However, only our previous work (Cabrio and Villata, 2012) combines TE with AF, but here our goal is to introduce a framework for inconsistencies detection in TE annotations.

3 Experimental setting

Data set. We added 60 pairs to the Debatepedia dataset [2] (extracted from a sample of Debatepedia[3] debates (Cabrio and Villata, 2012)), resulting in 160 pairs as training set, and 100 pairs as test set (balanced wrt to entailment/contradiction).

Evaluation. *First step*: we assess the performances of the TE system to correctly assign the TE relations to the pairs of arguments in the dataset. *Second step*: we evaluate how much such performances impact on the flattening of the BEG-AF, i.e., how much a wrong assignment of a relation to a pair of arguments is propagated in the AF. It is actually to detect such wrong assignments that the BEGincs framework has been conceived.

To recognize TE, we tested several algorithms from the EOP[4], i.e. BIUTEE (Stern and Dagan, 2011), TIE[5] and EDITS (Kouylekov and Negri, 2010). BIUTEE obtained the best results on Debatepedia (configuration exploiting all available knowledge resources): Acc:0.71, Rec:0.94, Pr:0.66, F-meas:0.78. As baseline we use a token-based version of the Levenshtein distance algorithm, i.e. EditDistanceEDA in the EOP (Acc:0.58, Rec:0.61, Pr:0.59, F-meas:0.59).

Then, we consider the impact of the best TE configuration on the arguments acceptability. We use admissibility-based semantics to identify the accepted arguments both on *i)* the goldstandard entailment graphs of Debatepedia topics, and *ii)* on the graphs generated using the relations assigned by BIUTEE. On the 10 Debatepedia graphs, BEGincs avg pr:0.68, avg rec:0.91, F-meas:0.77. BIUTEE mistakes in relation assignment propagate in the AF, but results are promising. The incons. detection module takes $\sim$1 sec. to analyze a BEG of 100 nodes and 150 relations.

4 Concluding remarks

We have presented BEGincs, a new formal framework that, translating a BEG into an argumentation graph, returns inconsistent set of arguments, if a wrong relation assignment by the TE system occurred. These inconsistent arguments sets are then used by annotators to detect the presence of a wrong assignment, and if so, to narrow the set of possibly erroneous relations. If no mistakes are produced in relation assignment, by definition BEGincs semantics return consistent arguments sets.

Assuming that in several real world scenarios TE pairs are interconnected, we ask to the NLP community to contribute in the effort of building suitable resources. In BEGincs, we plan to verify and ensure transitivity of BEGs.

[2] The only available dataset of T-H pairs combined into bipolar entailment graphs.

[3] http://idebate.org/

[4] http://bit.ly/ExcitementOpenPlatform

[5] http://bit.ly/MaxEntClassificationEDA

References

J. Berant, I. Dagan, and J. Goldberger. 2010. Global learning of focused entailment graphs. In *ACL*, pages 1220–1229.

G. Boella, D. M. Gabbay, L. W. N. van der Torre, and S. Villata. 2010. Support in abstract argumentation. In P. Baroni, F. Cerutti, M. Giacomin, and G. R. Simari, editors, *COMMA*, volume 216 of *Frontiers in Artificial Intelligence and Applications*, pages 111–122. IOS Press.

E. Cabrio and S. Villata. 2012. Natural language arguments: A combined approach. In *Procs of ECAI, Frontiers in Artificial Intelligence and Applications 242*, pages 205–210.

E. Cabrio and S. Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions;. *Argument & Computation*, 4(3):209–230.

G. Carenini and J. D. Moore. 2006. Generating and evaluating evaluative arguments. *Artif. Intell.*, 170(11):925–952.

C. I. Chesñevar and A.G. Maguitman. 2004. An argumentative approach to assessing natural language usage based on the web corpus. In *Procs of ECAI*, pages 581–585.

I. Dagan, B. Dolan, B. Magnini, and D. Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE)*, 15(04):i–xvii.

P.M. Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358.

V. Wei Feng and G. Hirst. 2011. Classifying arguments by scheme. In *Procs of ACL-2012*, pages 987–996.

M. Kouylekov and M. Negri. 2010. An open-source package for recognizing textual entailment. In *Procs of ACL 2010 System Demonstrations*, pages 42–47.

M.C. De Marneffe, A.N. Rafferty, and C.D. Manning. 2008. Finding contradictions in text. In *Procs of ACL*.

A. Stern and I. Dagan. 2011. A confidence model for syntactically-motivated entailment proofs. In *Proceedings of RANLP 2011*.

A. Wyner and T. van Engers. 2010. A framework for enriched, controlled on-line discussion forums for e-government policy-making. In *Procs of eGov 2010*.

A Graph-based Model of Contextual Information in Sentiment Analysis over Twitter

Giuseppe Castellucci(†), **Danilo Croce**(‡), **Roberto Basili**(‡)
(†)Department of Electronic Engineering
(‡)Department of Enterprise Engineering
University of Roma, Tor Vergata
castellucci@ing.uniroma2.it, {croce,basili}@info.uniroma2.it

Abstract

English. Analyzing the sentiment expressed by short messages available in Social Media is challenging as the information when considering an instance is scarce. A fundamental role is played by *Contextual* information that is available when interpreting a message. In this paper, a Graph-based method is applied: a graph is built containing the contextual information needed to model complex interactions between messages. A Label Propagation algorithm is adopted to spread polarity information from known polarized nodes to the others.

Italiano. *Uno dei principali problemi nella analisi delle opinioni nei Social Media riguarda la quantitá di informazione utile che un singolo messaggio puó fornire. Il contesto di un messaggio costituisce un insieme di informazioni utile ad ovviare questo problema per la classificazione della polaritá. In questo articolo proponiamo di rappresentare le interazioni tra i messaggi in grafi che sono poi utilizzati in algoritmi di Label Propagation per diffondere la polaritá tra i nodi.*

1 Introduction

Sentiment Analysis (SA) (Pang and Lee, 2008) faces the problem of deciding whether a text expresses a sentiment, e.g. positivity or negativity. Social Media are observed to measure the sentiment expressed in the Web about products, companies or politicians. The interest in the analysis of tweets led to the definition of highly participated challenges, e.g. (Rosenthal et al., 2014) or (Basile et al., 2014). Machine Learning (ML) approaches are often adopted to classify the sentiment (Pang et al., 2002; Castellucci et al., 2014; Kiritchenko et al., 2014), where specific representations and hand-coded resources (Stone et al., 1966; Wilson et al., 2005; Esuli and Sebastiani, 2006) are used to train some classifier. As tweets are very short, the amount of available information for ML approaches is in general not sufficient for a robust decision. A valid strategy (Vanzo et al., 2014b; Vanzo et al., 2014a) exploits *Contextual* information, e.g. the *reply-to* chain, to support a robust sentiment recognition in online discussions.

In this paper, we foster the idea that Twitter messages belong to a network where complex interactions between messages are available. As suggested in (Speriosu et al., 2011), tweets can be represented in graph structures, along with words, hashtags or users. A Label Propagation algorithm (Zhu and Ghahramani, 2002; Talukdar and Crammer, 2009) can be adopted to propagate (possibly noisy) sentiment labels within the graph. In (Speriosu et al., 2011), it has been shown that such approach can support SA by determining how messages, words, hashtags and users influence each other. The definition of the graph is fundamental for the resulting inference, e.g. when mixing messages about different topics, sentiment detection can be difficult. We take inspiration from the contexts defined in (Vanzo et al., 2014b). In (Speriosu et al., 2011) no explicit relation between messages is considered. We, instead, build a graph where messages in the same context are linked each other and to the words appearing in them. Moreover, we inject prior polarity of words as available in a polarity lexicon (Castellucci et al., 2015). Experiments are carried out over a subset of the Evalita 2014 Sentipolc (Basile et al., 2014) dataset, showing improvements in the polarity classification with respect to not using networked information.

In the remaining, Section 2 presents our graph-based approach. In Section 3 we evaluate the pro-

posed method with respect to a dataset in Italian and we derive the conclusions in Section 4.

2 Sentiment Analysis through Label Propagation over Contextual Graphs

Twitter messages are not created in isolation, but they live in conversations (Vanzo et al., 2014b; Vanzo et al., 2014a). Graph based methods (Zhu and Ghahramani, 2002; Talukdar and Crammer, 2009) provide a natural way to represent tweets in a graph structure in order to exploit relationships between messages to support the SA task.

2.1 Label Propagation Algorithms

In a classification task, given a graph representing a set of objects whose classes are known (labeled seeds) and a set of unlabeled objects, Label Propagation (LP) algorithms spread the label distribution by exploiting the underlying graph. Labels are spread over a graph $\mathbb{G} = \langle V, E, W \rangle$, where V is a set of n nodes, E is a set of edges and W is an $n \times n$ matrix of weights, i.e. w_{ij} is the weight of the edge between nodes i and j.

The Modified Adsorption (MAD) algorithm (Talukdar and Crammer, 2009) is a particular LP algorithm where the spreading of label distributions provides the labeling of all the nodes in the graph, possibly re-labeling also the seeded ones in order to improve robustness against outliers. MAD is defined starting from the Adsorption algorithm (Baluja et al., 2008), where the labeling of all the nodes in a graph is modeled as a controlled random walk. Three actions drive this random walk: `inject` a seeded node with its seed label; `continue` the walk from the current node to a neighbor; `abandon` the walk. These actions are modeled in the MAD algorithm through a minimization problem whose objective function is:

$$\sum_{l \in V} [\, \mu_1 (\vec{Y}_l - \hat{\vec{Y}}_l)^{\mathsf{T}} S (\vec{Y}_l - \hat{\vec{Y}}_l) + \mu_2 \, \hat{\vec{Y}}_l^{\mathsf{T}} L \hat{\vec{Y}}_l \; + \; \mu_3 | \hat{\vec{Y}}_l - R_l | \,] \qquad (1)$$

where S, L and R are matrices whose role is to model respectively the relationships between a node and its prior labels, the relationships between two similar nodes and the regularization imposed to the labeling of nodes[1]. The objective function aims at imposing the following constraints to the labeling process with these three terms: the algorithm should assign to a labeled vertex l a distribution $\hat{\vec{Y}}_l$ w.r.t. the target classes that is close to the a-priori distribution ($\vec{Y}_l$); moreover, if two nodes are close according to the graph, then their labeling should be similar. Finally, the third term is a regularization factor. More details about MAD are reported in (Talukdar and Crammer, 2009).

In our approach, the MAD algorithm is applied to a graph where each node is labeled with a distribution over some polarity classes[2]. We assume that a subset of the messages have been annotated, and they are used to train a classifier f that ignore the graph structure. The function f is then used to label the remaining messages so that the MAD algorithm is used to determine the final polarities based on the graph structure.

2.2 Contextual Graph: a definition

In order to generalize the contextual models proposed in (Vanzo et al., 2014b), we build a *Contextual Graph* $\mathbb{G}$ of messages as following. Given a message t_j we consider its context $C(t_j)$ as the list of l preceding messages $t_{j-1}, t_{j-2}, \ldots, t_{j-l}$. The context can be defined as the *reply-to* chain of messages (*conversation* context) or the temporally preceding messages sharing at least one hashtag (*hashtag* context). The contextual graph $\mathbb{G}$ is then built by considering pairs of messages (t_o, t_n) in a context, i.e. $t_o, t_n \in C(t_j)$. These are linked with an edge whose weight w_{t_o,t_n} is computed through a function that depends from the distance between t_o and t_n. In particular, we choose $w_{t_o,t_n} = e^{-\lambda |o-n|}$, where λ controls the influence of messages at different distances. These weighted edges are meant to capture the interaction between close messages in the context. We augment the set of vertices V with nodes representing the words appearing in messages. In particular, given $r_1, r_2, \ldots, r_k$ as the words composing t_o, we add k nodes to V, each representing a word r_i. Each word node is connected to its message and the weight w_{t_o,r_i} is computed through the $\sigma(t_o, r_i)$ function[3]. Word nodes are intended to make the graph connected: without them the graph would be composed by many disconnected sub-graphs, i.e. one per context. Moreover, the

[1] The three hyper-parameters μ_1, μ_2 and μ_3 are used to control the importance of each of these terms.

[2] If a node cannot be initialized with any method, the distribution is initialized with a value of $1/c$, where c is the number of classes.

[3] In the experiments reported below, a boolean function is adopted, i.e. $\sigma(t_o, r_i) = 1$ if r_i belongs to t_o.

more words two messages share, the more they are conveying a similar message. Finally, we define the set of seed nodes as a subset of V that are associated to prior labels. As discussed in the next section, these can be either messages or words: the former are seeded through noisy labels computed from a classification function f; the latter are seeded through label distributions derived from a polarity lexicon.

3 Experimental Evaluation

In order to evaluate the *Contextual Graph* and the MAD algorithm, we adopted a subset of the Evalita 2014 Sentipolc dataset (Basile et al., 2014). It consists of short messages annotated with the `subjectivity`, `polarity` and `irony` classes. We selected those messages annotated with polarity and that were not expressing any ironic content to focus our investigations on less ambiguous messages. Thus, the datasets used for our evaluations consist of a training set Tr of 2,449 messages and a testing set Ts of 1,129 messages.

Dataset	w/ conv	w/ hashtag	w/ both
Tr	349(14,27%)	987(40.36%)	80(3.27%)
Ts	169(14.98%)	468(41.48%)	47(4.16%)

Table 1: Dataset statistics w.r.t. contexts.

As in (Vanzo et al., 2014b), we downloaded the conversation and hashtag contexts that were available at the time of downloading[4]. In Table 1 the number of messages involved in the different contexts are shown. In the experiments, messages are classified with respect to the *positive*, *negative* and *neutral* polarity classes. The message distribution with respect to these classes is shown in Table 2.

Dataset	positive	negative	neutral
Tr	761	973	715
Ts	365	464	300

Table 2: Dataset statistics w.r.t. polarities.

3.1 Graph Construction

In the *Contextual Graph*, vertices represent messages and auxiliary information, such as words. In the LP algorithm each vertex can become a seed, i.e. a distribution w.r.t. the polarity classes can be assigned to it. We first investigate a configuration in which only messages are seeded. Experiments are carried out on three types of *Contextual Graph*s. In the first experiment a graph is built by considering contexts where messages are in a *reply-to* relationship, namely *conversation graph*. A second experiment considers instead the *hashtag* contexts, where messages share at least one hashtag. A third experiment considers both *conversation and hashtag* contexts in the same graph representation. In these configurations, vertices representing words are added to the graph but they are not "seeded" (i.e. they are considered as unlabeled nodes). In the fourth experiment, the last graph is enriched by electing as seeds also words whose sentiment polarity is known a-priori, e.g. derived by a polarity lexicon. In the following, we describe how to associate polarity distributions both to messages and words.

Message seeding. A classification function f that feeds the label distributions for messages is derived by a supervised learning process. In particular, we consider the training set Tr described above, and we acquire a Support Vector Machine multi-classifier in a One-Vs-All schema for the *positive*, *negative* and *neutral* polarity classes as in (Castellucci et al., 2014). Two types of features are adopted: the first is a boolean Bag-of-Words (BOW) feature set. The second is a Wordspace (WS) feature set derived from vector representations of the words in a message, obtained through a neural word embedding (Mikolov et al., 2013). We acquired the embedding from a corpus of 10 million tweets downloaded during the first months of 2015. A *skip-gram* model is acquired through the `word2vec`[5] tool and deriving[6] 250-dimensional vectors for about $99,410$ words. The WS feature set for a message t_j is obtained by considering the linear combination of word vectors that appear in t_j. The SVM classifier realizes the function f that assign the initial label distribution, reflecting the classifier confidence in assigning a polarity to each message, i.e. belonging to both train and test datasets, as well as belonging to contexts.

Words seeding. Seeds words are also considered when building the *Contextual Graph*. In particular, we adopt the Distributional Polarity Lexicon (DPL) (Castellucci et al., 2015) that associates each word to the prior information about the positivity, negativity and neutrality. The lexicon is built as follows: a classifier d is acquired from

[4] Results are not directly comparable to other systems participating to the Evalita 2014 challenge as some message was not more available in Twitter.

[5] `https://code.google.com/p/word2vec/`

[6] `word2vec` settings are: min-count=50, window=5, iter=10 and negative=10.

a dataset of generic messages gathered by Twitter considering the occurrence of noisy labels, i.e. emoticons expressing positivity, negativity or neutrality. In a nutshell, given the properties of the vector space WS, we project sentences and words in the same space, in order to transfer the polarity from sentences to words via the classifier d and obtaining the polarity scores of the DPL. The *positivity* and *negativity* scores of a word in DPL are used as seed distribution in the MAD algorithm.

3.2 Experimental Results

A first measure is given by the SVM classifier that is used to assign a polarity distribution to seeds belonging to the test dataset. We measure the mean between the F1 scores of the *positive* and *negative* classes (F1-Pn), and the mean between the F1 scores of all the three classes (F1-Pnn). Different feature representations are exploited in the SVM evaluation, as pointed out in Table 3.

Features	F1-Pn	F1-Pnn
BOW	0.630	0.583
BOW+WS	0.688	0.636

Table 3: SVM results (w/o contexts).

When adopting also the WS features, the performance increases in both the performance measures, with respect to the setting where only BOW features are considered.

Ctx size	F1-Pn	F1-Pnn
3	0.693	0.633
6	0.695	0.634
ALL	0.695	0.637

Table 4: MAD on conversation context.

Ctx size	F1-Pn	F1-Pnn
3	0.696	0.635
6	0.697	0.635
16	0.698	0.634
31	0.701	0.634

Table 5: MAD on hashtag context.

In Tables 4 and 5 the MAD algorithm results[7] are reported w.r.t. the *Conversation* and *Hashtag* contexts, as well to different context sizes, e.g. at size 3 we consider a maximum of 2 messages preceding a target one. The MAD algorithm is able to consistently increase the F1-Pn performance measure, while it is equally performing in the F1-Pnn. When adopting the *Hashtag* context, performances are higher w.r.t. the *Conversation* context setting. This is probably due to the fact the only 15% of the messages belong to a *reply-to* chain, while about 40% of the message belong to a Hashtag context. Moreover, *Hashtag* contexts refer to more coherent sets of messages. It makes the LP algorithm better exploit the graph by assigning similar labeling to nodes in the *Hashtag* context.

In Table 6 we applied the MAD algorithm over a graph built considering both contexts: in this scenario, we tuned and adopted two distinct λ values, i.e. λ_c and λ_h, respectively when weighting messages in conversation and hashtag contexts. Again, the contribution of the contextual information is measured through an increment of both the performance measures. Moreover, the contribution of the two contexts allows to further push the performances up, confirming the need of additional information when dealing with such short messages.

	Message seeding		+DPL	
Ctx Size	F1-Pn	F1-Pnn	F1-Pn	F1-Pnn
3	0.697	0.635	0.703	0.636
6	0.700	0.637	0.705	0.638
16	0.702	0.638	**0.719**	**0.648**
31	0.708	0.640	0.708	0.638

Table 6: MAD on both contexts.

Finally, we injected seed distributions over words through the Distributional Polarity Lexicon (DPL). The lexicon allows injecting a-priori seed on the words in the *Contextual Graph*, resulting in higher performances w.r.t. the case without DPL.

4 Conclusion

In this paper, the *Contextual Graph* is defined as a structure where messages can influence each other by considering both *intra-context* and *extra-context* links: the former are links between messages, while the latter serves to link messages in different contexts through shared words. The application of a Label Propagation algorithm confirms the positive impact of contextual information in the Sentiment Analysis task over Social Media. We successfully injected prior polarity information of words in the graph, obtaining further improvements. This is our first investigation in graph approaches for SA: we only adopted the MAD algorithm, while other algorithms have been defined, since (Zhu and Ghahramani, 2002) and they will be investigated in future works. Moreover, other contextual information could be adopted. Finally, other datasets should be considered, proving the effectiveness of the proposed method that does not strictly depend on the language of messages.

[7] the λ value and the MAD hyper-parameters μ_1, μ_2, μ_3 have been tuned on a validation set in each experiment.

References

Shumeet Baluja, Rohan Seth, D. Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. 2008. Video suggestion and discovery for youtube: Taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 895–904, New York, NY, USA. ACM.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the evalita 2014 sentiment polarity classification task. In *Proc. of the 4th EVALITA*.

Giuseppe Castellucci, Danilo Croce, Diego De Cao, and Roberto Basili. 2014. A multiple kernel approach for twitter sentiment analysis in italian. In *Fourth International Workshop EVALITA 2014*.

Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2015. Acquiring a large scale polarity lexicon through unsupervised distributional methods. In Chris Biemann, Siegfried Handschuh, Andr Freitas, Farid Meziane, and Elisabeth Mtais, editors, *Natural Language Processing and Information Systems*, volume 9103 of *Lecture Notes in Computer Science*, pages 73–86. Springer International Publishing.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of 5th LREC*, pages 417–422.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *JAIR*, 50:723–762, Aug.

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

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*, volume 10, pages 79–86, Stroudsburg, PA, USA. ACL.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proc. SemEval*. ACL and Dublin City University.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD '09, pages 442–457, Berlin, Heidelberg. Springer-Verlag.

Andrea Vanzo, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2014a. A context based model for sentiment analysis in twitter for the italian language. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa, Italy.

Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014b. A context-based model for sentiment analysis in twitter. In *Proceedings of 25th COLING*, pages 2345–2354. Dublin City University and Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of EMNLP*. ACL.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, CMU CALD.

Word Sense Discrimination: A gangplank algorithm

Flavio Massimiliano Cecchini, Elisabetta Fersini
Università degli Studi di Milano-Bicocca
Viale Sarca 336, Ed. U14, 20126 Milano
{flavio.cecchini,fersiniel}@disco.unimib.it

Abstract

English. In this paper we present an unsupervised, graph-based approach for Word Sense Discrimination. Given a set of text sentences, a word co-occurrence graph is derived and a distance based on Jaccard index is defined on it; subsequently, the new distance is used to cluster the neighbour nodes of ambiguous terms using the concept of "gangplanks" as edges that separate denser regions ("islands") in the graph. The proposed approach has been evaluated on a real data set, showing promising performance in Word Sense Discrimination.

Italiano. *L'obiettivo di questo articolo è descrivere un approccio di clustering non supervisionato e basato su grafi per individuare e discriminare i differenti sensi che un termine può assumere all'interno di un testo. Partendo da un grafo di cooccorrenze, vi definiamo una distanza fra nodi e applichiamo un algoritmo basato sulle "passerelle", cioè archi che separano regioni dense ("isole") all'interno del grafo. Discutiamo i risultati ottenuti su un insieme di dati composto da tweet.*

1 Introduction

Word Sense Disambiguation is a challenging research task in Computational Linguistics and Natural Language Processing. The main reasons behind the difficulties of this task are ambiguity and arbitrariness of human language: just depending on its context, the same term can assume different interpretations, or senses, in an unpredictable manner. In the last decade, three main research directions have been investigated (Navigli, 2009; Navigli, 2012): 1) supervised (Zhong and Ng, 2010; Mihalcea and Faruque, 2004), 2) knowledge-based (Navigli and Ponzetto, 2012; Schmitz et al., 2012) and 3) unsupervised Word Sense Disambiguation (Dorow and Widdows, 2003; Véronis, 2004), where the last approach is better defined as "induction" or "discrimination". In this paper we focus on the automatic discovery of senses from raw text, by pursuing an unsupervised Word Sense Discrimination paradigm. We are interested in the development of a method that can be generally independent from the register or the linguistic well-formedness of a text document, and, given an adequate pre-processing step, from language. Among the many unsupervised research directions, i.e. context clustering (Schütze, 1998), word clustering (Lin, 1998), probabilistic clustering (Brody and Lapata, 2009) and co-occurrence graph clustering (Widdows and Dorow, 2002) , we committed to the last one, based on the assumption that word co-occurrence graphs can reveal local structural properties tied to the different senses a word might assume in different contexts.

Given a global word co-occurrence graph, the main goal is to exploit the subgraph induced by the neighbourhood of the word to be disambiguated (a "word cloud"). There, we define separator edges ("gangplanks") and use them as the means to cluster the word cloud: the fundamental assumption is that, in the end, every cluster will correspond to a different sense of the word.

The paper is organized as follows. In section 2 we explain how we build our co-occurrence graph and word clouds by means of a weighted Jaccard distance. In section 3 we describe the gangplank algorithm. In section 4 we present the algorithm's results on our data set and their evaluation. In section 5 we give a brief overview on related work and section 6 presents some short conclusions.

2 Building the word graphs

Given a set of sentences, we derive a global co-occurrence word graph. It is a weighted, undirected graph where each node corresponds to a word (token) and there is an edge between two nodes if and only if the corresponding words co-occur in the same sentence. Edge weights are the respective frequencies of such co-occurrences. It has been shown (i Cancho and Solé, 2001) that a word graph like this tends to behave like a small-world, scale-free graph (Watts and Strogatz, 1998). In short, the graph is very cohesive and its cohesion hinges on few nodes from which nearly every other node can be reached. A similar structure can be quite difficult to handle for our purposes, since on the one hand the largest part of the graph tends to behave as a single, inextricable unit, and on the other hand the graph collapses in a myriad connected components if the hub nodes are removed: we are interested in a clustering between the two extremes. To mitigate this problem, a word filtering step is performed. Stopwords and irrelevant word classes (based on part-of-speech tagging), as e.g. adverbs and adjectives, are removed. Only nouns and verbs are retained.

2.1 A weighted Jaccard distance

Given the previously outlined word graph, we introduce a graph-based distance between nodes derived from Jaccard index that will be enclosed in our clustering algorithm. Given a graph G, each neighbourhood of a node w in G is treated as a multiset[1], where its elements correspond to the neighbour nodes of w and their multiplicity is the weight of the edge that connects them to w, i.e. the number of times they co-occur with w. We have defined the "automultiplicity" of w in this multiset as the greatest weight between all such edges. Given two multisets A and B, now we can define the weighted Jaccard distance as

$$1 - \frac{|A \cap B|}{|A \cup B|},$$

where the intersection is the multiset with the least multiplicity for each element of both (possibly 0, so not counting it) and the union is the multiset with the greatest multiplicity for each element of both. The cardinality of a multiset is the sum of all the multiplicities of its elements. The weighted Jaccard distance provides a measure of how much the contexts of two words overlap, taking into account the importance of each context by means of the weights. A distance of 1 means that contexts do not overlap, and on the contrary a distance of 0 implies a complete overlap. The weighted Jaccard distance can of course be expanded to neighbourhoods of depth greater than 1: for increasing depths, we would take into account contexts of neighbour words, contexts of contexts, and so on. This means that the greater the depth, the less significant the Jaccard distance becomes. It can be shown that, for depth d, any two elements have Jaccard distance (strictly) less than 1 if and only if the shortest path connecting them consists of at most $2d$ edges. This lemma will be used in the next section.

2.2 Word clouds

Given a word w of interest, we extract from G the subgraph G_w induced by the open neighbourhood[2] of w, originating a "word cloud". We again perform word filtering and remove redundant words, this time using principal component analysis, retaining just words that contribute the most to the first, most important component, and considering the corresponding subgraph of G_w (we will denote it by the same notation).

On it, we can define a local weighted Jaccard distance, as explained before. This allows us the transition from the word (sub)graph to a word metric space. From the metric space we build again a weighted, undirected "distance graph" D_w, where two nodes are connected by an edge if and only if their weighted Jaccard distance is strictly less than 1, and weights are the distances between words. As noted at the end of section 2.1, this operation practically coincides with the expansion of G_w where we add edges between nodes that are 2 steps away from each other and reassign a weight corresponding to distance to each edge.

3 The gangplank clustering algorithm

3.1 The algorithm

Our objective is a clustering of D_w that maximizes intra-cluster connections and minimizes inter-cluster connections. As explained in Section 2, our assumption is that clusters of a word cloud

[1] A multiset is a set where an element can recur more than once, and can be defined as a set of couples of the type (element, multiplicity) (Aigner, 2012).

[2] In our case, we consider neighbourhoods of degree 1.

will define different senses, implicitly identified with the clusters themselves.

Our approach focuses on edges. We define an edge e in D_w connecting two nodes u and v to be a gangplank if its weight is strictly greater than the mean of the weights of the edges departing from both its ends u and v: if this happens, then edge e is keeping distant the two local graph's "halves" it connects (the neighbourhood of u excluding v and viceversa). In other words, the two aforementioned halves on both sides of e, seen as different subgraphs of D_w, are on their own more tightly connected regions than the subgraph induced by the union of u's and v's neighbourhoods (and thus including e) would be. To each node v we can assign the number $g(v)$ of gangplanks going out from it; $g(v)$ will be comprised between 0 and $\deg(v)$. We also define the ratio $\gamma(v) = \frac{g(v)}{\deg(v)}$. The smaller $\gamma(v)$, the more we deem v to be in the middle of a tightly connected area, or island.

Our clustering algorithm doesn't set a predetermined number of clusters. It starts by ranking each node v by the ratio $\gamma(v)$ and takes the node with the smallest γ as the seed of the first cluster K. We start then a cycle of expansion and reduction steps. In the expansion step, we add all the neighbours of K to K. Then, in the reduction step, we begin discarding from K all the nodes whose connections are undermining the homogeneous nature of cluster K. More precisely, we rank the nodes in K by the ratio $\gamma_K(u) = \frac{g_K(u)}{\deg_K(u)}$, where $g_K(u)$ and $\deg_K(u)$ are defined as $g(u)$ and $\deg(u)$, but with respect to the subgraph of D_w induced by K. Then, we remove from K the node with the greatest non-zero γ_K, if there is any. Thereafter we update γ_K for each remaining node in K and again remove the node with the greatest non-zero ratio. We continue the reduction step until we can no longer remove any node, and hence no gangplanks are left in cluster K. The cluster's seed will never be removed. Once the reduction step has finished, the expansion and reduction step is repeated, this time ignoring any previously discarded node. The cycle continues until no further expansion is possible. At this point we have obtained the first cluster. Now, we select the yet unclustered node with greatest γ in D_w and start a new cycle of expansion and reduction steps for the corresponding new cluster, and so on, until every node has been clustered.

In the end, we'll obtain a clustering of D_w. However, many clusters will often consist of just one node: these are nodes between more tightly connected areas, which we would like to assign to bigger clusters to avoid dispersion. To this end, we set m_w as the minimum allowed cluster size: m_w is the length of the shortest (filtered) sentence where w appears or 2, whichever is greater. This choice of m_w is motivated by the fact that, in the graph, every sentence forms a clique that we have to consider as a possible cluster. All the clusters whose size is less than m_w are post-processed and their elements reassigned to one of the bigger clusters. Again, we rank these remaining nodes by γ and, starting from the node v with the smallest ratio (the less "noisy" node) and going up, we assign v to the cluster $K_m = \arg\min_K \gamma_K(v)$ (the cluster with less relative gangplank connections to v), until finally all nodes have been clustered. If two or more K_m are eligible, the biggest one is preferred.

A pseudo-code of the proposed gangplank clustering algorithm is reported below.

Algorithm 1 Gangplank clustering algorithm

$\mathcal{K} = \{\}$ ▷ The set of future clusters
$\mathcal{V} = E_{D_w}$ ▷ The set of nodes in D_w
$\mathfrak{s} = \{\}$ ▷ Nodes to be assigned in second step
while $\mathcal{V} \neq \emptyset$ **do**
  $v = \arg\min_{u \in \mathcal{V}} \gamma(u)$
  $K = \{v\}$ ▷ The new, seeded cluster
  $\mathfrak{n} = \{\}$ ▷ Discarded, noisy nodes
  $\mathcal{N} = \bigcup_{u \in K} N_{D_w}(u) \backslash \mathfrak{n}$ ▷ Neighbours of K
  while $\mathcal{N} \neq \mathfrak{n}$ **do**
    $K = K \cup \mathcal{N}$
    while $\exists u \in K \backslash \{v\} : \gamma_K(u) \neq 0$ **do**
      $u = \arg\max_{t \in K} \gamma_K(t)$
      $K = K \backslash \{u\}$
      $\mathfrak{n} = \mathfrak{n} \cup \{u\}$
    end while
    $\mathcal{N} = \bigcup_{u \in K} N_{D_w}(u) \backslash \mathfrak{n}$
  end while
  if $|K| \geq m_w$ **then**
    $\mathcal{K} = \mathcal{K} \cup \{K\}$ ▷ Add the new cluster
  else
    $\mathfrak{s} = \mathfrak{s} \cup K$
  end if
  $\mathcal{V} = \mathcal{V} \backslash K$ ▷ Remove clustered nodes
end while
while $\mathfrak{s} \neq \emptyset$ **do**
  $s = \arg\max_{r \in \mathfrak{s}} \gamma(r)$
  $K_s = \arg\min_{K \in \mathcal{K}} \gamma_{K \cup s}(s)$
  $K_s = K_s \cup \{s\}$ ▷ Expand the cluster
  $\mathfrak{s} = \mathfrak{s} \backslash \{s\}$
end while
return $\mathcal{K}$

3.2 The labelling step

Once we have obtained a given number of cluster-senses relative to the chosen term, we adopt a ma-

Keywords	Tagged tweets	No. of senses	Most common senses ($\geq 10\%$)
blizzard	463	23	snowstorm 43%, video game company 37%
caterpillar	467	23	CAT machines 30%, animal 24%, The Very Hungry Caterpillar 17%, CAT company 16%
england	474	11	country (UK) 65%, national football team 10%, New England (USA) 10%
ford	558	12	Harrison Ford 40%, Ford vehicles 30%, Tom Ford (fashion designer) 25%
india	474	5	country 50%, national cricket team 48%
jfk	474	13	New York airport 61%, John Fitzgerald Kennedy 33%
mcdonald	425	47	McDonald's (restaurants) 38% , McDonald's (company) 31%
mars	440	24	planet 66%, Bruno Mars 17%
milan	594	41	Milano (Italy) 58%, A.C. Milan football team 24%
pitbull	440	7	rapper 49%, dog breed 48%
venice	482	9	Venezia (Italy) 55%, Venice beach (California) 42%

Table 1: Keywords and entities.

jority voting mechanism to label each of the term's occurrences in the original sentences. For each sentence where the disambiguated term appears, we compute the Jaccard distance between the set of the sentence's filtered words and each cluster. Then, we assign the term a label referring to the nearest cluster. It is possible that not every cluster will be assigned to a term's occurrence; these are "weak" clusters that are maybe either too insignificant or too fine-grained.

4 Evaluation on tweets

4.1 Data and tagging

In order to evaluate the performance of the proposed approach from a quantitative point of view, a benchmark data set should be employed. However, data sets like SemEval are not ideal, since they don't present enough samples for each word, therefore yielding a sparse and most often unweighted (i.e. all weights are equal to 1) graph. For these reasons, we created an *ad hoc* data set consisting of 5291 tweets in English, downloaded from Twitter on a single day using eleven different keywords; the data set is summarized in Table 1. Keywords were chosen to be common nouns that may possess many different senses, and were the target of our word sense discrimination. To have a basis for evaluation, we manually tagged keywords occurring in the tweets, in order to create a ground truth.

4.2 Evaluation and results

We evaluated the coherence of our clustering and subsequent word labelling with respect to the data set's "true" labels. For each keyword's cluster-derived labelling, we compare that label's context (i.e. all the words in the corresponding filtered sentences) to all the true labels' contexts by means of the Jaccard distance. We then identify the cluster-derived label with its closest true label. We end up with a mapping σ going from some of our clusters to the true labels. To measure the quality of the proposed solution, accuracy's performance has been computed for each disambiguated term, as reported in Table 2. The global accuracy score we obtained is $62{,}56\%$. It could be argued that accuracies are worse whenever the keyword is not polarized on two senses, as is the case for *caterpillar* or *mcdonald*, with many possible senses and no two of them covering together more than 90% of all senses. This could happen because in this case the word cloud will be fractured in many subunits, the gangplanks algorithm will tendentially split them even more and so surfacing labels will be sparse and somewhat inorganic.

For confrontation, we also ran the *Chinese Whispers* algorithm (Biemann, 2006), which uses a simplified form of Markov clustering, on our graphs, obtaining a global accuracy score of $60{,}1\%$ with a mean number of just 2,27 clusters per keyword (a behaviour mentioned in Section 2). Local scores are shown in Table 2. Accuracies are only better when senses are strongly polarized, e.g. for *pitbull* and *england*. In the latter case, just one cluster is found, so the algorithm's accuracy is the same as the percentage of occurrences of the main sense.

5 Related work

An approach similar to ours, at least in the initial graph construction, can be found in (Véronis, 2004). The weights we put on edges substantially differ from his, but, most markedly, Véronis wants to span some trees from some hub nodes in each word cloud. In other words, Véronis's algorithm is more hierarchical in nature, where ours is more aggregative. Similar considerations can also be made for (Hope and Keller, 2013).

6 Conclusions

The main challenge we encountered for our word sense discrimination algorithm was the difficulty of handling a small-world graph. Apart from that, we have to notice that word clustering just rep-

	blizzard	caterpillar	england	ford	india	jfk	mars	mcdonald	milan	pitbull	venice
Labelling accuracy	**49,9**	42,0	50	**87,8**	**82,7**	**67,5**	**75,5**	40	53,2	64,5	**71,0**
No. of clusters	38	20	46	17	28	14	9	15	32	55	34
No. of labels	13	11	7	6	5	4	5	9	16	5	7
Chinese Whispers	43,0	**43,7**	**65,4**	80,1	63,1	61,2	66,4	**40,7**	**58,2**	**81,1**	55,0

Table 2: Local labelling accuracies for the gangplank and Chinese Whispers clustering algorithm. Accuracies in %. *No. of labels* represent how many labels effectively appear in labelled tweets.

resents the last step of a process that starts with pre-processing and tokenization of a text, which are both mostly of supervised nature. Our future goals will be to investigate the relations between text pre-processing and clustering results, and how to render the whole process completely unsupervised.

References

Martin Aigner. 2012. *Combinatorial theory*, volume 234. Springer Science & Business Media.

Chris Biemann. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, pages 73–80. Association for Computational Linguistics.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111. Association for Computational Linguistics.

Beate Dorow and Dominic Widdows. 2003. Discovering corpus-specific word senses. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 79–82. Association for Computational Linguistics.

David Hope and Bill Keller. 2013. Maxmax: a graph-based soft clustering algorithm applied to word sense induction. In *Computational Linguistics and Intelligent Text Processing*, pages 368–381. Springer.

Ramon Ferrer i Cancho and Richard V Solé. 2001. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

Rada Mihalcea and Ehsanul Faruque. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of ACL/SIGLEX Senseval*, volume 3, pages 155–158.

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.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Roberto Navigli. 2012. A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and practice of computer science*, pages 115–129. Springer.

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Jean Véronis. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics.

Facebook and the Real World: Correlations between Online and Offline Conversations

Fabio Celli, Luca Polonio
University of Trento
{fabio.celli,luca.polonio}@unitn.it

Abstract

English. Are there correlations between language usage in conversations on Facebook and face to face meetings? To answer this question, we collected transcriptions from face to face multi-party conversations between 11 participants, and retrieved their Facebook threads. We automatically annotated the psycholinguistic dimensions in the two domains by means of the LIWC dictionary, and we performed correlation analysis. Results show that some Facebook dimensions, such as "likes" and shares, have a counterpart in face to face communication, in particular the number of questions and the length of statements. The corpus we collected has been anonymized and is available for research purposes.

Italiano. *Ci sono correlazioni tra l'uso del linguaggio nelle conversazioni su Facebook e faccia a faccia? Per rispondere a questa domanda, abbiamo raccolto delle trascrizioni di conversazioni di gruppo tra 11 partecipanti e campionato i loro dati Facebook. Abbiamo annotato automaticamente le dimensioni psicolinguistiche per mezzo del dizionario LIWC e abbiamo estratto le correlazioni tra le due diverse tipologie testuali. I risultati mostrano che alcune dimensioni linguistiche di Facebook, come i "mi piace" e il numero di condivisioni, correlano con dimensioni linguistiche dell'interazione faccia a faccia, come il numero di domande e la lunghezza delle frasi. Il corpus e' stato anonimizzato ed e' disponibile per scopi di ricerca.*

1 Introduction and Background

In recent years we had great advancements in the analysis of communication, in face to face meetings as well as in Online Social Networks (OSN) (Boyd and Ellison, 2007). For example, resources for computational psycholinguistics like the Linguistic Enquiry Word Count (LIWC) (Tausczik and Pennebaker, 2010), have been applied to OSN like Facebook and Twitter for personality recognition tasks (Golbeck et al., 2011) (Schwartz et al., 2013) (Celli and Polonio, 2013) (Quercia et al., 2011). Interesting psychological research analyzed the motivations behind OSN usage (Gosling et al., 2011) (Seidman, 2013) and whether user profiles in OSN reflect acual personaliy or a self-idealization (Back et al., 2010).

Also Conversation Analysis (CA) of face to face meetings, that has a long history dating back to the '70s (Sacks et al., 1974), has taken advantage of computational techniques, addressing detection of consensus in business meetings (Pianesi et al., 2007), multimodal personality recognition (Pianesi et al., 2008) and dectection of conflicts from speech (Kim et al., 2012).

In this paper we make a comparison of the linguistic behaviour of OSN users both online and in face to face meetings. To do so, we collected Facebook data from 11 volunteer users, who participated to an experimental setting where we recorded face to face multiparty conversations of their meetings. Our goal is to discover relationships between a rich set of psycholinguistic dimensions (Tausczik and Pennebaker, 2010) extracted from Facebook metadata and meeting transcriptions. Our contributions to the research in the fields on Conversation Analysis and Social Network Analysis are: the release of a corpus of speech transcriptions aligned to Facebook data in Italian and the analysis of correlations between psycholinguistic dimensions in the two settings.

The paper is structured as follows: in section 2 we describe the corpora and the data collection, in section 3 we explain the method adopted and report the results, in section 4 we draw some conclusions.

2 Data and Method

We collected 11 volunteer Italian native speakers, who provided the consent to use their Facebook metadata, and organized meeting sessions with them to collect spoken linguistic data. The meetings consist in sessions of one hour, where participants, 6 in the first session and 5 in the second one, performed free multi-party conversations. Groups were balanced by gender and aged between 18 and 50 years. There were no restrictions, predefined task or topic to elicitate speech. In order to prevent biases in the interactions we put in the groups persons who do not know each other.

We recorded and manually transcribed a corpus of spoken conversations from the meeting sessions, splitting utterances by turns where a speaker ends its speech or is interrupted by another speaker. Then we annotated each utterance with dialogue act (DA) labels. To select DA labels we referred to Novielli & Strapparava (Novielli and Strapparava, 2010), who performed a dialogue act annotation on meetings transcriptions in Italian. We just added the label "laugh" to their label set. The final dialogue act label set we used is reported in Table 1. The agreement on the annotation of

label	description	example
Req	Questions	what's your name?
St	Statements	Today is sunny
Op	Opinions	I think that..
Agr	Acceptance	ok for me!
Rej	Rejection	no, thanks
In	Opening	hello!
End	Closing	goodbye!
Ans	Answers	My name is ..
Lau	Laughs	haha

Table 1: Dialogue act label set.

dialogue act labels between 2 non-expert labelers is $k = 0.595$ (Fleiss et al., 1981). This moderate agreement score, and the feedback from the annotators, indicate that the task is hard due to the presence of long and complex utterances.

We aligned the data from spoken conversations with public data from the participants' Facebook profiles. Using Facebook APIs, we collected data from 6 months before the meeting session to 1 year later. We collected public status updates, including text messages, links, pictures, and multimedia files posted and received on the participants' walls. We distinguished between statuses posted

metadata	description
fb-friends	number of friends
fb-pics	number of photos
fb-comm	avg number of comments received
fb-likes	avg number of likes received
fb-p-tot	count of all P's posts
fb-p-usr	posts by P on his/her wall
fb-p-oth	posts by others on P's wall
fb-shared	posts of the P shared by others
fb-text	count of textual posts
fb-media	count of non-textual posts
fb-chars	average characters in posts
fb-words	average words of posts

Table 2: Description of Facebook metadata collected.

by the users and statuses posted on the users' wall by others. Eventually we computed the numerical metadata reported in table 2 and we analyzed the textual pots.

We anonymized both the transcription and the Facebook data. The final corpus contains 2 audio files (one hour each) with transcriptions (about 21000 tokens and about 1600 utterances in total; 1750 words and 133 utterances on average per participant), and Facebook data of the participants (about 80000 tokens, about 5800 posts including multimedia status updates). We automatically annotated the textual data in the corpus with the Italian version of LIWC (Alparone et al., 2004). Doing so, we annotated words with 85 psychological dimensions, such as linguistic categories (verbs, prepositions, future tense, past tense, swears, etc.), psychological processes (anxiety, anger, feeling, cognitive mechanisms, etc.), and personal concerns (money, religion, leisure, TV, achievement, home, sleep. etc.). In the next section we report the results of the analysis of the data collected.

3 Experiments and Results

Scope From a communication analysis perspective, face to face meetings and Facebook are two very different settings: in Facebook the communication is written, asynchronous, mediated and with an audience that is a mix of friends and unknown people. On the contrary in face to face meetings the communication is oral, synchronous, not mediated, and the audience is unknown people. In a theory of communication (Shannon and Weaver, 1949), illustrated in Figure 1, all those levels are variables related to the sender, receiver and medium. Here we restrict the scope of this

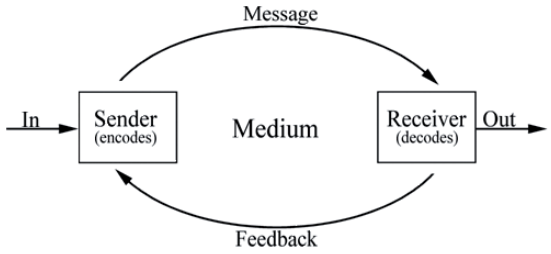


Figure 1: Schema of communication as transmission. We limit the scope of this work to the message level.

work to the analysis of message level, leaving to future work the possibility to extend this analysis to the characteristics of the media or the participants.

Experiments First of all we analyzed the topics in Facebook and meeting transcriptions. We removed the stopwords and we generated two word clouds with the 70 most frequent words in each dataset with 5 as minimum term frequency. We report the word clouds in Figure 2. The comparison of the two clouds reveal that participants

Figure 2: Word clouds of the 70 most frequent words in meeting transcriptions and Facebook data

to the experiments in Facebook discussed and planned actions ("*dormire*", "*andare*") places ("*rimini*", "*copenhagen*") and times ("*sera*", *stasera*", "*domani*") while in meetings they told and discussed mainly about places ("*bologna*", "*rimini*") and people ("*tipo*", "*gente*").

In order to discover relationships between psycholinguistic dimensions in Facebook and face to face meetings, we labelled the texts with LIWC, and we computed how much the psycholinguistic dimensions correlate in the two settings. We observed few, but strong, significative correlations (for significative we mean correlations with p-value smaller than 0.05 and correlation greater than 0.5), reported in table 3.

Word type (LIWC-it)	corr. to both settings
Anxiety	0,510***
Anger	0,580***
Feel	0,571***
Future	-0,532**
Home	-0,715*
TV	0,711*
sleep	0,537***
swears	0,696**

Table 3: Pearson's Correlations between LIWC dimensions in texts from Facebook profiles of the participants and face to face meeting. Only dimensions significantly correlating are reported. Significance is ***=p-value smaller than 0.001; **=p-value smaller than 0.01; *=p-value smaller than 0.05.

The dimensions with strong correlation are related to powerful emotions, difficult to control, like anxiety and anger, but also to the tendency to express feelings and emotions with words. Swears, that is the dimension with the highest combination of *correlation coefficient* and significance, is related as well to a dimension difficult to control. Maybe less interesting for our purposes are other dimensions with high correlations related to the content of discourse, like "home", "TV", "future" and "sleep". We ran automatic topic modeling with a Hierarchical Latent Dirichlet Allocation (Teh et al., 2006) (Blei et al., 2003) to reveal that participants spoke about "TV" and "sleep" in both settings, but about "home" and "future" only in Facebook and not in face to face meetings. This is why these values are negative.

We also compared behavioral data from Facebook and meetings. In particular we computed the correlations between Facebook metadata and dialogue acts annotated in meeting transcriptions, plus metadata from face to face meetings, namely the average length of utterances in words and characters. Results, reported in Table 4, show that

	f2f-req	f2f-st	f2f-op	f2f-agr	f2f-rej	f2f-in	f2f-end	f2f-ans	f2f-lau	f2f-words
fb-friends	0,243	0,130	-0,047	-0,298	-0,080	0,166	-0,475	-0,206	-0,063	-0,156
fb-pics	0,167	-0,157	0,281	-0,198	-0,410	-0,078	-0,253	0,163	-0,185	-0,084
fb-comm	0,439	-0,295	-0,003	0,464	-0,036	-0,287	0,297	-0,525	0,173	-0,064
fb-likes	**0,698***	-0,379	0,308	-0,276	-0,033	0,064	0,383	-0,230	-0,143	0,079
fb-p-tot	0,533	-0,078	-0,020	0,286	-0,117	-0,147	-0,240	-0,553	0,107	-0,135
fb-p-usr	0,140	-0,176	-0,297	0,230	0,174	0,311	-0,475	0,094	0,066	-0,157
fb-p-oth	-0,140	0,176	0,297	-0,230	-0,174	-0,311	0,475	-0,094	-0,066	0,157
fb-shared	-0,204	**0,698***	0,384	-0,352	-0,060	-0,206	-0,292	-0,155	-0,272	**0,619***
fb-text	-0,043	-0,096	-0,142	0,417	0,123	-0,336	0,427	-0,427	0,420	-0,100
fb-media	0,043	0,096	0,142	-0,417	-0,123	0,336	-0,427	0,427	-0,420	0,100
fb-chars	0,305	0,193	0,276	-0,042	-0,209	-0,475	0,269	-0,442	-0,161	0,309
fb-words	0,247	0,215	0,217	-0,005	-0,166	-0,453	0,275	-0,426	-0,124	0,283

Table 4: Pearson's correlations between metadata from Facebook and dialogue act labels from face to face meetings. *=p-value smaller than 0.05.

there are few, but very interesting, significative correlations. The number of likes received by the participants on Facebook correlate positively with a tendency to ask questions in meetings. This is quite surprising and perhaps reveals a will to engage the audience asking questions. Crucially, other significative correlations are related to shares generated in Facebook by the participants. In particular this is correlated with long statements in face to face meetings. In practice, people posting contents that are reshared online, in face to face meetings tend to produce long statements and talk more than the others.

4 Discussion and Conclusions

In this paper, we attempted to analyse the correlations between psycholinguistic dimensions observed in Facebook and face to face meetings. We found that the type of words significantly correlated to both settings are related to strong emotions (anger and anxiety), We suggest that these are linguistic dimensions difficult to control and tend to be constant in different settings. Crucially, we also found that likes received on Facebook are correlated to the tendency to ask questions in meetings. Literature on impression formation/management report that people with high self-esteem in meetings will elicit self-esteem enhancing reactions from others (Hass, 1981). This could explain the link between the tendency to ask questions in meetings with unknown people and the tendency to post contents that elicit likes in Facebook. Moreover, the tendency to ask questions in spoken conversations is correlated to observed emotional stability (Mairesse et al., 2007) and that emotionally stable users in Twitter tend to have more replies in conversations than neurotic users (Celli and Rossi, 2012). We suggest that the correlation we found can be partially explained by these two privious findings.

Another very interesting finding is that the tendency to be reshared on Facebook correlates to the tendency to speak a lot in face to face meetings. Again, literature about impression formation/management can explain this, because people with high self-esteem tend to engage people and to speak a lot, while people adopting defensive strategies tend to be assertive less argumentative. In linguistics it is an open debate whether virality depends from the influence of the source (Zaman et al., 2010) or the content of message being shared (Guerini et al., 2011) (Suh et al., 2010). In particular, the content that evokes high-arousal positive (amusement) or negative (anger or anxiety) emotions is more viral, while content that evokes low arousal emotions (sadness) is less viral (Berger and Milkman, 2012). Given that the tendency to express both positive and negative feelings and emotions in spoken conversations is a feature of extraversion (Mairesse et al., 2007), and that literature in psychology links the tendency to speak a lot to extraversion (Gill and Oberlander, 2002), observed neuroticism (Mairesse et al., 2007) and dominance (Bee et al., 2010). we suggest that the correlation between long turns in meetings and highly shared contents in Facebook may be due to extraversion, dominance and high self-esteem.

We are going to release the dataset we collected on demand.

Aknowledgements

We wish to thank the artist Valentina Perazzini for the contribution in the collection of data and Luca Rossi (University of Copenhagen) for the discussions.

References

Francesca R Alparone, S. Caso, A. Agosti, and A Rellini. 2004. The italian liwc2001 dictionary. Austin, TX: LIWC.net.

Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*.

Nikolaus Bee, Colin Pollock, Elisabeth André, and Marilyn Walker. 2010. Bossy or wimpy: expressing social dominance by combining gaze and linguistic behaviors. In *Intelligent Virtual Agents*, pages 265–271. Springer.

Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research*, 49(2):192–205.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Danah Boyd and Nicole Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Fabio Celli and Luca Polonio. 2013. Relationships between personality and interactions in facebook. In *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, pages 41–54. Nova Science Publishers, Inc.

Fabio Celli and Luca Rossi. 2012. The role of emotional stability in twitter conversations. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 10–17. Association for Computational Linguistics.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2:212–236.

Alastair Gill and Jon Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 253–262. ACM.

Samuel D Gosling, Adam A Augustine, Simine Vazire, Nicholas Holtzman, and Sam Gaddis. 2011. Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking*, 14(9):483–488.

Marco Guerini, Carlo Strapparava, and Gözde Özbal. 2011. Exploring text virality in social networks. In *Proceedings of ICWSM*, pages 1–5.

Glen R Hass. 1981. Presentational strategies and the social expression of attitudes: Impression management within limits. *Impression management theory and social psychological research*, pages 127–146.

Samuel Kim, Fabio Valente, and Alessandro Vinciarelli. 2012. Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5089–5092. IEEE.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500.

Nicole Novielli and Carlo Strapparava. 2010. Exploring the lexical semantics of dialogue acts. *J Comput Linguist Appl*, 1(1-2):9–26.

Fabio Pianesi, Massimo Zancanaro, Bruno Lepri, and Alessandro Cappelletti. 2007. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41(3-4):409–429.

Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. 2008. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 180–185. IEEE.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

Andrew H Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):773–791.

Gwendolyn Seidman. 2013. Self-presentation and belonging on facebook: How personality influences social media use and motivations. *Personality and Individual Differences*, 54(3):402–407.

Claude E Shannon and Warren Weaver. 1949. *The mathematical theory of communication*. University of Illinois press.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*, pages 177–184. IEEE.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Tauhid R Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern. 2010. Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, pages 599–601.

La scrittura in emoji tra dizionario e traduzione

Francesca Chiusaroli
Università di Macerata
francesca.chiusaroli@unimc.it

Abstract

English The paper presents an analysis of semantics and uses of emoji in digital writing, mainly through the observation of some recent applications in translation. The purpose is to discuss the hypothesis of setting up an emoji multilingual dictionary and translator through a process of selection and assessment of conventional semantic values. Translation cases may show how images can convey common and universal meanings, beyond specific peculiarities, so as they can stand as models in the perspective of an interlanguage. The analysis will move from the definition of "scritture brevi" (short writings) as developed in Chiusaroli and Zanzotto 2012a 2012b, and now at www.scritturebrevi.it.

Italiano *Il presente contributo propone un'analisi sulla semantica e sugli usi degli emoji nella scrittura digitale, in particolare attraverso l'osservazione di alcune recenti applicazioni nell'ambito della traduzione. Scopo dell'analisi è di discutere l'ipotesi della costituzione di un dizionario e traduttore emoji multilingue, attraverso un procedimento a posteriori di selezione e fissazione dei valori semantici convenzionali. La dimensione traduttiva consente di valutare la capacità designativa dell'immagine, oltre le specificità delle lingue, per esprimere significati comuni e universali, dunque tali da potersi costituire come modelli nella prospettiva della lingua veicolare e dell'interlingua. L'analisi muoverà dalla nozione di "scritture brevi" quale si trova definita in Chiusaroli e Zanzotto 2012a, 2012b, e ora in www.scritturebrevi.it.*

1. Introduzione

L'odierna popolarità degli *emoji* negli ambienti digitali non trova adeguato riscontro in termini di impieghi razionali, a motivo dell'alto grado di vaghezza implicito nella figura. Nonostante le diffuse dichiarazioni e i continui annunci sull'avvento di un nuovo idioma universale per immagini, resta l'impraticabilità di fatto di un simile linguaggio espressivo, evidentemente carente sul piano "strutturale", della *langue*. L'assenza di un sistema condiviso, infatti, instaura una costante condizione di ambiguità semantica che preclude l'affermazione e gli usi dell'auspicato codice generale,[1] richiamando e riproducendo così il destino delle tradizioni grafiche storiche, che, come è noto, hanno sperimentato i limiti dei sistemi pittografici o avviato la loro specializzazione linguistica.

Mentre l'*emoticon* - combinazione sequenziale di caratteri per l'espressione facciale come :-) - si configura sempre più come un solido elemento disambiguante per la comunicazione delle componenti emozionali nell'ambito della scrittura "digitata", utile per contrastare l'indeterminatezza affidata alla parola in forma scritta con l'aggiunta del fondamentale tratto/richiamo prosodico, appare al contrario scarsamente definita la semantica degli *emoji*, la serie sempre più ricca di simboli di tastiera che riproducono referenti e "oggetti" del discorso attraverso distinte forme pittografiche. Proprio il carattere iconico, infatti, inteso ad assicurare la comprensione oltre, o contro, le barriere linguistiche specifiche, dà luogo piuttosto a variabili soluzioni di lettura del medesimo segno, con effetti sulla corretta o univoca trasmissione/comprensione del messaggio.

2. La dimensione nomenclaturista

Rispetto alle comuni pratiche d'uso, estemporanee e soggettive, il riferimento a un sistema linguistico specifico appare come un utile e idoneo strumento di uniformazione, capace di limitare la proliferazione incontrollata delle forme e dei contenuti. Scopo del presente contributo è di valutare l'ipotesi di una collocazione degli *emoji* nella prospettiva di un codice veicolare norma-

[1] Si veda il dichiarato insuccesso del pur avvincente *Emojili* (http://emoj.li/), esperimento di un *emoji-only network*, un social network vincolato alla comunicazione esclusiva tramite *emoji*.

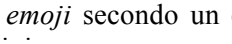
lizzato, ovvero per la capacità di porsi quali segni di un sistema intermediario, ed eventualmente automatico, per la traduzione multilingue, attraverso un procedimento di trasferimento e applicazione di valori semantici comuni, generali e condivisi, secondo un metodo di pianificazione (meta)linguistica *a posteriori*. La funzione codificatrice dell'intermediazione linguistica può provvedere alla prioritaria assegnazione di valori logografici alle figure, con speciale efficacia nei contesti traduttivi. Il disegno, che visivamente, per la pregnanza pittografica, rinvia a un'ampia sommatoria di valori semantici, può acquisire, attraverso lo strumento traduttivo, significati convenuti, linguistici prima, e poi logografici, consentendo la fissazione di corrispondenze utili all'impiego degli *emoji* secondo un codice convenzionale e condiviso.

Rispondono all'istanza della regolarizzazione iniziative come l'acquisizione degli *emoji* nello standard *unicode*,[2] oppure gli elenchi a base semantica e nomenclatoria, con relativa versione in lingua, principalmente inglese, sulla cui base risultano strutturati i lessici delle tastiere *emoji* internazionali.[3] La tendenza universalizzante caratterizza anche le collezioni enciclopediche,[4] da cui l'individuazione delle macrocategorie generali: *People*, *Nature*, *Food & Drink*, *Celebration*, *Activity*, *Travel & Places*, *Objects & Symbols*. Nella prospettiva lessicale e nomenclaturista si interpretano applicazioni "traduttive" come *Emoji Fortunes* (http://emojifortun.es/), un sistema automatico di produzione di brevi messaggi, composti di sequenze fortuite di tre *emoji* con le rispettive equivalenze in lingua inglese:

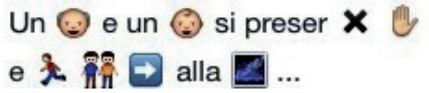
Proprio a partire dal criterio traduttivo, la corrispondenza può essere evidentemente trasferita ad altre lingue, determinando infine la codificazione di un repertorio *emoji* funzionante come dizionario veicolare.

3. La grammatica

L'organizzazione in aree semantiche riprodotta nelle tastiere *emoji* dei dispositivi digitali non definisce di per sé la forma morfologica, rendendo così evidenti i limiti della scrittura in *emoji* nella rappresentazione sintattica per la resa dei contenuti relativi a enunciati e proposizioni. Quando la traduzione si sposta dalla parola al testo, la selezione della forma difficilmente è operata sulla base della nomenclatura predisposta, bensì tende a essere dettata dall'estemporaneo rinvenimento e dall'abbinamento intuitivo e improvvisato. Una distinzione categoriale come *Persone*, *Oggetti* vs. *Attività* non comporta, ad esempio, l'assoluta e aprioristica assegnazione dei valori linguistici grammaticali sulla base delle funzioni "sostantivo"/"verbo". Così l'*emoji* "lampadina" vale anche per indicare il verbo "illuminare":



Fa capo al blog *Scritture Brevi* un esperimento di scrittura tramite *emoji* (da cui il caso precedente), che consiste nella traduzione in figure di brevi stringhe testuali:



L'impostazione esplicitamente ludica dell'iniziativa, e l'opzione della scrittura mista (in lettere e *emoji*), si pongono come incentivi all'approccio creativo nelle interpretazioni dei segni, generando plurime applicazioni in senso grammaticale, e orientate di volta in volta sul significato o sul significante, con interessanti soluzioni in favore della dimensione plurilingue, linguistica specifica e internazionale:

Oltre a produrre omografie (è il caso, appena osservato, dell'*emoji* "stella/star"), la qualità pittografica dell'immagine induce naturalmente problemi di "sinonimie", per le affinità semantiche tra i segni:

Nella dimensione dell'atto linguistico o momento della *parole*, l'immagine assume così il proprio significato soprattutto in rapporto alle condizioni cotestuali, ovvero secondo i principi delle relazioni sintagmatiche e paradigmatiche innestate dal testo. L'approccio libero e creativo non agisce, come prevedibile, nella direzione

[2] http://blog.unicode.org/search/label/emoji

[3] Ad esempio www.emojisites.com e https://themeefy.com/TitashNeogi6/whatisemoji

[4] Sull'esempio di Wikipedia si struttura *Emojipedia*: http://emojipedia.org/

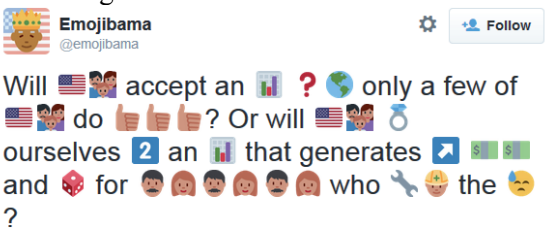
della limitazione del senso, bensì, al contrario, attesta la vastissima gamma funzionale dei segni. È invece, in questo caso, il contorno testuale a poter assumere la funzione di mediazione e a ridurre il grado di ambiguità, fino a favorire l'interpretazione attesa.

4. Universalismo vs. relativismo

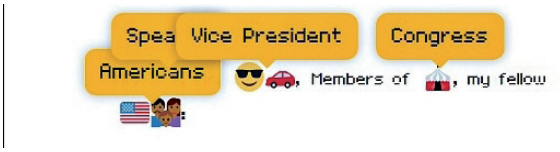
Contro la tendenza generalizzante della scrittura per immagini, l'adesione al principio traduttivo e glossatorio può far emergere le specificità linguistico-semantiche, la corrispondenza istituita andando nella direzione della riproduzione di sensi peculiari del codice fonte. Il sistema delle conoscenze rappresentato dalla lingua nazionale, con gli annessi portati storici e culturali, diventa allo stesso tempo valore aggiunto nel trasferimento del contenuto in figure, insieme evidenziando, come sempre, il ruolo della componente relativista nell'interpretazione.

Per l'aspetto connotativo e in relazione soprattutto alla *sentiment analysis* si veda, ad esempio, la differenziazione degli usi degli *emoji* su base etnica o regionale rilevata dall'*Emoji Report* di SwiftKey dell'aprile 2015,[5] che illustra la selezione di categorie diverse per l'espressione dello stesso "umore".

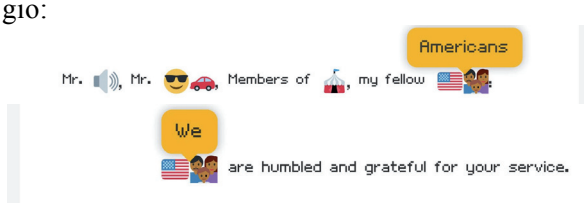
Risponde alla strategica attenzione per la rete l'esperimento promosso dalla testata statunitense *The Guardian* di rendere disponibile una traduzione in *emoji* dei discorsi di Barak Obama: *Emojibama*.[6] L'interesse pragmatico comunicativo appare come lo scopo più evidente dell'iniziativa, senz'altro prevalente rispetto alla ricerca linguistica:



La scelta di una scrittura mista mette ancora in risalto il ruolo fondamentale del cotesto, ma è ugualmente interessante la soluzione di rendere disponibile una lettura/traduzione (in lingua), ottenibile attraverso il semplice movimento del cursore sopra l'immagine, che provvede in via definitiva alla disambiguazione:



Parallelamente all'impatto sulla comunicazione universale, proprio il particolarismo linguistico caratterizza l'operazione, come mostrano certe soluzioni traduttive volte alla rappresentazione del soggetto-comunità destinatario del messaggio:



Nell'esempio, la rappresentazione dell'elemento pronominale ("we") attraverso un digramma (bandiera americana + gruppo familiare) contestualizza opportunamente il discorso rispetto all'uditorio (USA), e non riproduce astrattamente la categoria morfologica ("we"= noi statunitensi). Il procedimento di generalizzazione dell'immagine trova pertanto corrispondenza nella specifica riscrittura, ma si rivela poco adeguato nella prospettiva dell'interlingua.

Analoga problematica emerge nell'applicazione incoerente dei valori semantici, quale è il caso dell'adozione del numerale per il valore fonetico, secondo le comuni pratiche del *texting* (2 = *to*), evidentemente inidoneo all'eventuale lettura in una lingua diversa dall'inglese.

5. Testo letterario e frasario

Tra i progetti di traduzione in *emoji* spicca, per la considerevole dimensione "fisica" e per l'alto grado di sperimentalismo, il caso di *Emoji Dick*, "a crowd sourced and crowd funded translation of Herman Melville's Moby Dick into Japanese emoticons called emoji", per la cura di Fred Benenson.[7]

Il lavoro in *crowdsourcing* di circa 800 traduttori (ciascuna frase tradotta tre volte, con successiva selezione delle soluzioni ritenute migliori tramite votazione di gruppo) ha prodotto un imponente bagaglio di forme e frasi costituite. Il legame con un testo canonico, di cui si hanno traduzioni accreditate e "d'autore", rilascia un

[5] http://www.scribd.com/doc/262594751/SwiftKey-Emoji-Report#scribd

[6] http://www.theguardian.com/us-news/ng-interactive/2015/jan/20/-sp-state-of-the-union-2015-address-obama-emoji, con relativo account di Twitter @emojibama

[7] Per il testo e il progetto: https://www.kickstarter.com/projects/fred/emoji-dick

repertorio potenzialmente utile all'ipotesi di una applicazione multilingue, ovvero per l'eventuale definizione di un codice *emoji* stabilizzato sulla base dell'adattamento a lingue diverse della stessa versione in immagini. La scelta della redazione collettiva rende ragione della volontà di uscire dai margini della pratica idiosincratica, inevitabile nelle produzioni individuali, operando nel senso dell'aggregazione e della riduzione delle versioni all'unità minima del significato. Tale prospettiva di unificazione non si sottrae tuttavia ai limiti della composizione personale, evocativa e non letterale, per l'adozione del metodo a base di frase che praticamente impedisce l'articolazione e l'annotazione degli elementi del codice, come mostra l'incommensurabilità sostanziale col testo originale nella versione "interlineare", mostrato in Figura 1 (nella pagina seguente).

Diversamente dalla scrittura letteraria, dove la cifra stilistica dominante agevola la soluzione personale e suggestiva, il collegato progetto del traduttore automatico [8] sembra più opportunamente rivolto alla resa di espressioni della lingua comune, relative alla vita quotidiana, efficacemente realizzabili attraverso la pratica della glossatura *ad verbum*, pertanto più utile alla prospettiva interlinguistica:

6. Conclusioni

Proprio il richiamo alla corrispondenza biunivoca appare come l'elemento più significativo per un metodo che intenda considerare gli *emoji* non soltanto quali elementi dell'atto di *parole* (unico, sempre diverso), bensì come segni di un codice formalizzato e condiviso, il più possibile coerente, univoco e razionale.

Al di là della dimensione idiosincratica o creativa, oltre la vaghezza e l'equivocità dell'uso individuale, l'ipotesi della scrittura in *emoji* come sistema veicolare deve consegnare alla pratica un codice idoneo alla comunicazione internazionale e multilingue, un sistema dunque costruito da una preliminare selezione secondo un corretto equilibrio di coerenza ed efficacia, e capace di riprodurre le idee e di ridurre la superficiale varietà per cogliere la struttura, o il senso, profondi. La priorità assegnata alla definizione dell'interlingua in *emoji* terrà in debito conto specificità e occasionalismi in quanto imprescindibili nell'atto comunicativo storicamente e culturalmente collocato e, come tali, inclusi nell'inventario secondo la prospettiva gerarchica delle relazioni semantiche (iponimie, iperonimie) e formali, sintagmatiche e associative. Diverso ruolo sarà assegnato a significati non universalmente trasferibili o traducibili. Secondo un criterio tassonomico saranno dunque collocati pittogrammi specifici, allorché espressivi di valori storico-culturali peculiari, e nondimeno riconducibili alle forme di base, rispetto alle quali essi si porranno quali estensioni per aggiunta di elementi modificatori. Si tratta di un metodo per altro già adottato dai sistemi di tastiera nel recente rilascio degli *emoji* relativi alla rappresentazione delle notazioni etniche come il colore della pelle, i capelli, e altre caratteristiche fisiche o dell'orientamento etico-sociale, che stanno ampliando significativamente il repertorio predisposto, in tal modo abbandonando la cifra simbolica e adeguando la dimensione pittografica alla riproduzione sempre più fedele dei *realia*.

L'obiettivo della lingua-scrittura comune, storicamente ricercato dai programmi universalisti dall'epoca della linguistica cartesiana, può così trovare oggi un'adeguata occasione di affermazione nella scrittura in *emoji*: nuova scrittura potente per la popolarità, e fondata sul presupposto della comunicazione condivisa e globalizzata. L'ampliamento della rete sociale diventa fattore limitante della inevitabile deriva arbitrarista, ma è soprattutto l'ancoraggio al piano linguistico, attraverso lo strumento glossatorio, a garantire la costituzione del codice, traducibile in segni linguistici, come tale vincolato all'orizzonte di pensiero che la singola lingua predispone, come ogni lingua parziale e imperfetto, e tuttavia proprio per questo rigoroso ed efficace, l'unico in grado di consentire la comunicazione.

[8] https://www.kickstarter.com/projects/fred/the-emoji-translation-project

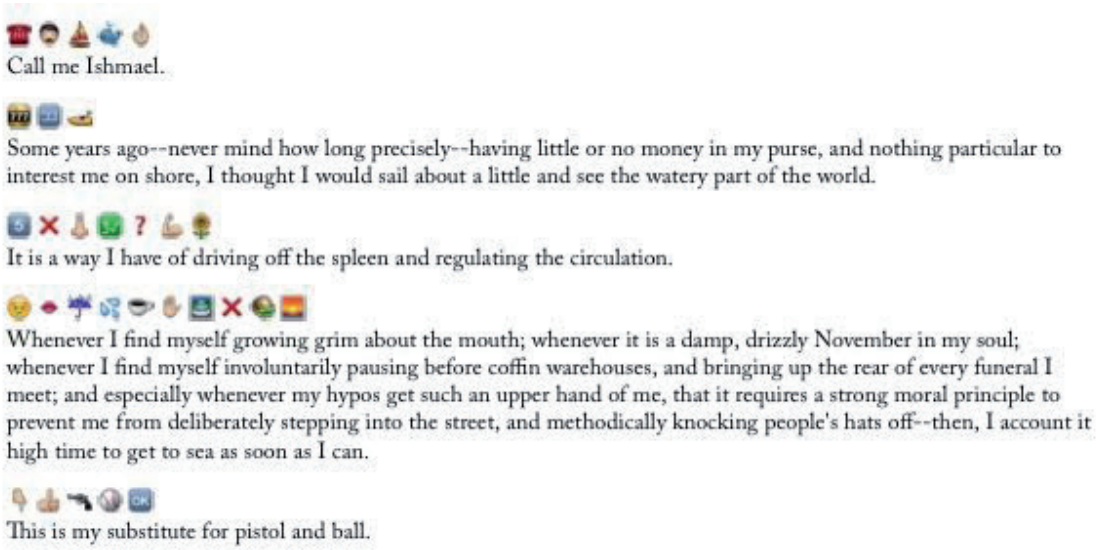


Figura 1

Bibliografia di riferimento

Giorgio Raimondo Cardona. 1981. *Antropologia della scrittura*. Loescher. Torino. Nuova ed. con prefazione di Armando Petrucci. 2009. Utet, Torino.

Giorgio Raimondo Cardona. 1986. *Storia universale della scrittura*. Mondadori. Milano.

Francesca Chiusaroli. 1998. *Categorie di pensiero e categorie di lingua. L'idioma filosofico di John Wilkins*. Il Calamo, Roma.

Francesca Chiusaroli. 2001. *Una trafila secentesca di* reductio. In Vincenzo Orioles (a cura di). *Dal 'paradigma' alla parola. Riflessioni sul metalinguaggio della linguistica*. Atti del Convegno, Università degli studi di Udine - Gorizia, 10-11 febbraio 1999. Il Calamo, Roma: 33-51.

Francesca Chiusaroli. 2012. *Scritture Brevi oggi. Tra convenzione e sistema*. In Francesca Chiusaroli, Fabio Massimo Zanzotto (a cura di). *Scritture brevi di oggi*. Quaderni di Linguistica Zero. 1. Università degli studi di Napoli L'Orientale, Napoli: 4-44.

Francesca Chiusaroli. 2015. *Scritture brevi e identità del segno grafico: paradigmi ed estensioni semiotiche*. In Laura Mariottini (a cura di). *Identità e discorsi. Studi offerti a Franca Orletti*. RomaTrE-Press, Roma: 251-264.

Francesca Chiusaroli. 2015. *Scritture Brevi per la realizzazione del falso*. In Gabriella Catalano, Marina Ciccarini, Nicoletta Marcialis (a cura di). *La verità del falso. Studi in onore di Cesare G. De Michelis*. Viella, Roma: 75-85.

Francesca Chiusaroli. (in stampa). *Scritture brevi e tendenze della scrittura nella comunicazione di Twitter*. In *Linguaggio e apprendimento linguistico: metodi e strumenti tecnologici*. Atti del XV Congresso Internazionale di Studi dell'Associazione Italiana di Linguistica Applicata (AItLA). Università del Salento, Lecce, 19-21 febbraio 2015.

Francesca Chiusaroli. (in stampa). *Scritture brevi in emoji, dalla scrittura alla lettura*. In Francesca Chiusaroli, Marina Ciccarini (a cura di). *Brevitas. Letture e scritture a confronto.* Workshop, Università di Roma "Tor Vergata", 25-26 febbraio 2015.

Francesca Chiusaroli. (in stampa). *Emoji, hashtag, TVB… Scritture brevi, categorie per un dizionario*. In *Scritture brevi: forme, modelli e applicazioni, per l'analisi e per il dizionario*: Secondo convegno interannuale Prin SCRIBE e Scritture Brevi, 28-30 maggio 2015, Università di Macerata.

Francesca Chiusaroli and Fabio Massimo Zanzotto (a cura di). 2012a. *Scritture brevi di oggi*. Quaderni di Linguistica Zero. 1. Università degli studi di Napoli L'Orientale, Napoli.

Francesca Chiusaroli and Fabio Massimo Zanzotto (a cura di). 2012b. *Scritture brevi nelle lingue moderne*. Quaderni di Linguistica Zero. 2. Università degli studi di Napoli L'Orientale, Napoli

Francesca Chiusaroli and Fabio Massimo Zanzotto. 2012. *Informatività e scritture brevi del web*. In Francesca Chiusaroli, Fabio Massimo Zanzotto (a cura di). *Scritture brevi nelle lingue moderne*. Quaderni di Linguistica Zero. 2. Università degli studi di Napoli L'Orientale, Napoli: 3-20.

Noam Chomsky. 1966. *Cartesian linguistics: a chapter in the history of rationalist thought*. Harper & Row, New York.

David Crystal. 2001. *Language and the Internet*. Cambridge UP, Cambridge.

David Crystal. 2003. *English as a global language*. Cambridge UP, Cambridge. II ed.

David Crystal. 2004. *A glossary of netspeak and textspeak*. Edinburgh UP, Edinburgh.

Eli Dresner and Susan C. Herring. 2010. *Functions of the non-verbal in CMC: emoticons and illocutionary force.* Communication Theory 20: 249-268.

Umberto Eco. 1993. *La ricerca della lingua perfetta nella cultura europea*. Laterza, Roma-Bari.

Umberto Eco. 2007. *Dall'albero al labirinto. Studi storici sul segno e l'interpretazione*. Bompiani, Milano.

Vyvyan Evans. 2014. *The language myth. Why language is not an istinct*. Cambridge, Cambridge UP.

Vyvyan Evans. (in stampa). *The emoji code: language and the future of communication*.

Adrian Frutiger. 1996. *Segni & simboli. Disegno, progetto e significato*. Trad. it. Stampa alternativa e graffiti, Roma.

Jack Goody. 1989. *Il suono e i segni*. Trad. it. Il Saggiatore, Milano.

André Leroi-Gourhan. 1977. *Il gesto e la parola. I. Tecnica e linguaggio. II. La memoria e i ritmi*. Trad. it. Einaudi, Torino.

Aleksandăr Lûdskanov. 2008. *Un approccio semiotico alla traduzione. Dalla prospettiva informatica alla scienza traduttiva*. Hoepli, Milano.

Zoe Mendelson. 2014. *Under the hood of the all-emoji programming language*. Co.Labs. Januar, 09, 2014.

Walter Ong. 1986. *Oralità e scrittura. Le tecnologie della parola*. Trad. it. Il Milano, Bologna.

Elena Pistolesi. 2014. *Scritture digitali*. In Giuseppe Antonelli, Matteo Motolese, Lorenzo Tomasin (eds.). *Storia dell'italiano scritto. Vol. III: Italiano dell'uso*. Roma, Carocci: 349-375.

Silvestri Domenico. (in stampa). *Primissime forme di scritture brevi: dai pittogrammi "metonimici" protosumerici alle complementazioni fonetiche ittite*. In Francesca Chiusaroli, Fabio Massimo Zanzotto. *Scritture brevi nella storia delle scritture*, Quaderno monografico di Linguistica Zero. Università degli studi di Napoli L'Orientale, Napoli.

On Mining Citations to Primary and Secondary Sources in Historiography

Giovanni Colavizza, Frédéric Kaplan
EPFL, CDH, DH Laboratory, Lausanne, Switzerland
{giovanni.colavizza,frederic.kaplan}@epfl.ch

Abstract

English. We present preliminary results from the Linked Books project, which aims at analysing citations from the historiography on Venice. A preliminary goal is to extract and parse citations from any location in the text, especially footnotes, both to primary and secondary sources. We detail a pipeline for these tasks based on a set of classifiers, and test it on the *Archivio Veneto*, a journal in the domain.

Italiano. *Presentiamo i primi risultati del progetto Linked Books, per l'analisi delle citazioni della storiografia su Venezia. Ci prefiggiamo l'estrazione e l'analisi delle citazioni da ogni posizione nei testi, specialmente note a pi pagina, sia a fonti primarie che secondarie. Discutiamo una serie di classificatori con questo obiettivo, valutandone i risultati su Archivio Veneto, una rivista del settore.*

1 Introduction

The Linked Books project is part of the Venice Time Machine[1], a joint effort to digitise and study the history of Venice by digital means. The project goal is to analyse the history of Venice through the lens of citations, by network analytic methods. Such research is interesting because it could unlock the potential of the rich semantics of the use of citations in humanities. A preliminary step is the extraction and normalization of citations, which is a challenge in itself. In this paper we present the first results on this last topic, over a corpus of journals and monographs on the history of Venice, digitised in partnership with the Ca' Foscari Humanities Library and the Marciana Library.

[1] http://vtm.epfl.ch/.

Our contribution is three-fold. First, we address the problem of extracting citations in historiography, something rarely attempted before. Secondly, we extract citation from footnotes, with plain text as input. Lastly, we deal at the same time with two different kind of citations: to primary and to secondary sources. A primary source is a documentary evidence used to support a claim, a secondary source is a scholarly publication (Wiberley Jr, 2010). In order to solve this problem, we propose a pipeline of classifiers dealing with citation detection, extraction and parsing.

The paper is organised as follows: a state of the art in Section 2 is followed by a methodological section explaining the pipeline and applied computational tools. A section on experiments follows, conclusions and future steps close the paper.

2 Related work

Sciences have largely used quantitative citation data to study their practices, whilst humanities remained largely outside of the process (Ardanuy, 2013). Difficulties of a concrete nature along with peculiar features of humanistic discourse make the task not trivial.

The lack of citation data for the humanities is well recognised, both for monographs and other kind of secondary literature (Heinzkill, 1980; Larivière et al., 2006; Linmans, 2009; Hammarfelt, 2011; Sula and Miller, 2014). Furthermore, citations are deployed within humanities in multifaceted ways, posing further challenges to their extraction and understanding (Grafton, 1999; Hellqvist, 2009; Sula and Miller, 2014).

One core element of citations in humanities, and especially so History, is the distinction between primary and secondary sources, and the quantitative and qualitative importance of both (Frost, 1979; Hellqvist, 2009). Little previous work on the use of primary sources via citations exist, with few exceptions in the domains of biblical stud-

ies and Classics (Murai and Tokosumi, 2008; Romanello, 2014).

The literature on citation extraction mirrors this scenario. As far as the citations to secondary sources are concerned, the development of automatic citation indexing systems has been a well explored area of research over the last two decades, starting from the seminal work of Giles et al. (1998). Increasingly, researchers are also tackling the problem of locating citations within the structure of documents (Lopez, 2009; Kim et al., 2012b; Heckmann et al., 2014). The extraction of citations to primary sources is instead a largely unexplored area, where recent effort has been produced within the fields of Classics (Romanello et al., 2009; Romanello, 2013; Romanello, 2014) and law (Francesconi et al., 2010; Galibert et al., 2010).

3 Approach

We propose a three-staged incremental pipeline including the following steps:

1. **Text block detection** of contiguous lines of text likely to contain citations, usually footnotes. The motivation for this preliminary step, inspired by Kim et al. (2012b), is to individuate the footnote space of a publication, as footnotes can span multiple pages.

2. **Citation extraction** within their boundaries over one or more contiguous text lines. This stage entails a token by token classification. A further sub-step is the classification of a citation as being *Primary* or *Secondary*, meaning to primary or secondary sources respectively.

3. **Citation parsing**, token by token, to detect all relevant components over a set of 50 mutually exclusive classes (e.g. *Author*, *Title* and *PublicationDate* for citations to secondary sources, or *Archive*, *Fond* and *Series* for primary sources).

The first step is dealt with using a SVM classifier,[2] initially trained with a small set of morphological features.

The second and last steps are approached with a group of CRF classifiers trained over a rich set of features, considering a bi-gram and tri-gram context, both backwards and forward. We train the models with Stochastic Gradient Descent and L2 regularisation, using the CRFSuite and default parameters (Okazaki, 2007).

Conditional Random Fields and Supporting Vector Machines are state-of-the-art models in the field of citation extraction since the work of Peng and McCallum (2006), and were introduced first by Cortes and Vapnik (1995) and Lafferty et al. (2001) respectively.

4 Experiments

The corpus is first digitised,[3] then OCRed using a commercial product with no extra training.[4] Our tests are based on an annotated sample of pages from the *Archivio Veneto*—a scholarly journal in Italian specialised in the History of Venice—randomly selected from a corpus of 92 issues from the year 1969 to 2013. The sample consists of 1138 annotated pages, for a total of 6257 annotated citations. Proper evaluation of the OCR quality and inter-annotator agreement are still pending at this stage. The annotation phase has been carried out with Brat.[5] No text format features—i.e. italics or type module—are used for the moment, and will be considered in a subsequent phase of the project.

4.1 Text block detection

The first classification step is a boolean one, where we are interested in knowing if a line of text, or a group of contiguous lines, is likely to contain citations, therefore likely to be a footnote. Text blocks are defined as groups of k contiguous lines of text. This step is required by the nature of footnotes, which can span over multiple pages demanding their proper identification in order to define the input space for subsequent stages in the pipeline. For each block we extract the following features: 1- **General**: line number (to detect footnotes); 2- **Morphological**[6]: punctuation frequency, frequency of digits, frequency of uppercase and lower-case characters, number of white spaces, number of characters, frequency of abbreviations according to multiple patterns, average word length, average frequency of specific punctuation symbols (“.”, “,”, “(”, “)”, “[”, “]”); 3- **Boolean**: if the chunk begins with a possible

[2] Using Python sklearn package.

[3] With 4DigitalBooks DLmini scanners.

[4] Abbyy FineReader Corporate 12.

[5] `http://brat.nlplab.org/`.

[6] Frequencies are always assessed character by character.

acronym or with a digit. After experimental tuning, we settle for a poly-linear model of degree 2 over a set of alternatives (degrees 1 to 10), which has the added value of maximizing recall, the most important metric at this early stage. The best division into text-blocks is found to be with $k = 2$. The evaluation of this step, based on a randomly-selected third of the annotated data (3633 blocks, 2204 negative and 1429 positive), is reported in Table 1.

Task	Precision	Recall	F1-score
no-citation	0.96	0.95	0.96
citation	0.92	0.95	0.93
avg / total	0.95	0.95	0.95

Table 1: Evaluation results for Text block detection.

Our results compare with others applying similar filtering methods (Kim et al., 2012a). In the future we will test a confidence classification with threshold lower than 0.5, as to further improve recall over precision.

4.2 Citation extraction

Given a text block likely to contain citations, we address the problem of citation extraction, meaning tokenizing the block and tagging each token as being part of a citation or not. For this phase and the next, text blocks are merged as to avoid any input being considered twice or more in the training and test sets. We merge together contiguous text lines likely to contain a citation, and consider k extra context (lines of text without citations) before and after. The set of features used for this step is organised in the following classes:[7]

1. **Shape of the token**: according to each character being upper-case, lower-case or punctuation. E.g. "UUU." for a token of length 4 with 3 upper-case characters and a final dot.

2. **Type of the token**: according to a set of classes such as if the token is a digit, or made of all upper-case letters, etc.

3. **Boolean features**: if the token is a 2 or 4 digit number, if it contains digits, if it contains upper or lower case characters, etc.

4. **Other features**: the token itself and it's position in the current line.

A more limited set of features is also considered in a bi and tri-gram conditioning over a sliding window within the preceding and following 3 tokens, namely: the tokens themselves, their shape and type, their position in the line.

The evaluation was conducted on a set of 19852 tokens (5240 primary and 14612 secondary) and 1056 text blocks, corresponding to a random third of the annotated corpus. The most balanced context turned out to be $k = 2$, results in Table 2. The performance is acceptably high in terms of overall item accuracy (0.95). In general, a higher context k means trading off precision for recall. Instance accuracy is apparently much lower (0.504), we must however remember that an instance at this level is a text block, possibly containing several non contiguous citations. Instance accuracy at the citation level improves to 0.78, and 0.84 if we tolerate for 1 token of difference between the golden standard and automatic tagging of a citation. We therefore attain results comparable to those Lopez (2010) got for the task of individuating non-patent references in patent text bodies.

Task	Precision	Recall	F1-score
no-citation	0.978	0.917	0.947
citation	0.926	0.98	0.953
avg / total	0.952	0.949	0.95

Table 2: Evaluation results for Citation extraction.

We further explored if a classifier trained with the same features could properly distinguish citations to primary and secondary sources. For this task each citation is parsed independently, assuming proper segmentation from the previous step. We attain an overall item accuracy of 0.967 and instance accuracy of 0.928 over the same training and testing sets. The fact that this classifier performs well allows us to consider the macro-category (primary or secondary) as a feature in the parsing step. Results in Table 3.

4.3 Citation parsing

This step involves the parsing of an extracted citation in order to individuate its components. The same set of features as before is used for each token, with the addition of:

- **Enhanced boolean features**: if the token is

[7] The full list of features is available upon request and partially inspired by Okazaki (2007).

Task	Precision	Recall	F1-score
primary	0.968	0.904	0.935
secondary	0.966	0.989	0.978
avg / total	0.967	0.947	0.956

Table 3: Evaluation results for Primary and Secondary Citation classification.

a time span (e.g. "1600-1700"), if it might be a Roman number, or an abbreviation.

- **The macro-category** (primary or secondary), as an indicator of the typology of the citation.

Task	Precision	Recall	F1-score
Author	0.939	0.958	0.948
Title	0.873	0.989	0.928
Pub.Place	0.927	0.899	0.913
Pub.Year	0.927	0.861	0.893
Pagination	0.961	0.978	0.969
Archive	0.968	0.912	0.939
ArchivalRef.	0.909	0.884	0.896
Folder	0.955	0.938	0.947
Registry	0.957	0.901	0.928
Cartulation	0.938	0.908	0.921
Foliation	0.862	0.890	0.875

Table 4: Evaluation results for Citation parsing: without macro-category feature.

Task	Precision	Recall	F1-score
Author	0.94	0.957	0.948
Title	0.9	0.984	0.94
Pub.Place	0.931	0.908	0.919
Pub.Year	0.945	0.893	0.918
Pagination	0.953	0.984	0.968
Archive	0.969	0.919	0.943
ArchivalRef.	0.901	0.895	0.898
Folder	0.956	0.942	0.949
Registry	0.971	0.901	0.935
Cartulation	0.964	0.934	0.949
Foliation	0.892	0.884	0.888

Table 5: Evaluation results for Citation parsing: with macro-category feature.

We test over a random 30% of the corpus and report only results of parsing with no extra context, which predictably gave the best results. Overall item and instance accuracy are 0.884 and 0.575 without the macro-category feature, and 0.893 and 0.592 with it. The testing set is comparable and proportional in size, yet different in sampling to the one used in step 2. Results in Table 4 and Table 5 only report the most significant classes in order to understand a citation, for citations secondary (above) and primary sources (below) respectively.[8]

The macro-category has only a marginal, albeit positive impact. Furthermore, some categories are either under-represented in terms of training instances, or easily mistaken for another one, contributing to the overall degradation of results. Such is the case for *Editor* or *Curator*, frequently classified as *Author*. In general several categories could be grouped, and lookup features—over list of names or library catalogues—should greatly improve our results.

The model performs well for the most significant categories, in comparison to models trained on more data and/or fewer categories and/or on references and not footnote citations. Specifically, we improve on Lopez (2010), Kim et al. (2012b), Romanello (2013), and compare to Heckmann et al. (2014).

5 Conclusions and future work

We presented a pipeline for recognizing and parsing citations to primary and secondary sources from historiography on Venice, with a case study on the *Archivio Veneto* journal. A first filtering step allows us to detect text blocks likely to contain citations, usually footnotes, by a SVM classifier trained on a simple set of morphological features. We then detect citation boundaries and macro-categories (to primary and secondary sources) using more rich features and CRFs. The last step in our pipeline is the fine-grained parsing of each extracted citation, in order to prepare them for further processing and analysis.

In the future we plan to design more advanced feature sets, first of all considering text format features. Secondly, we will implement the next package of our chain: an error-tolerant normalizer which will uniform all citations to the same primary or secondary source within a publication, as a means to minimise the impact of classification errors during previous steps.

[8] The full list of results is available upon request.

Acknowledgments

We thank Maud Ehrmann and Jean-Cédric Chappelier, EPFL, for useful comments.

The project is funded by the Swiss National Fund under Division II, project number 205121_159961.

References

Jordi Ardanuy. 2013. Sixty years of citation analysis studies in the humanities (1951-2010). *Journal of the American Society for Information Science and Technology*, 64(8):1751–1755.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. 2010. Semantic processing of legal texts - where the language of law meets the law of language.

Carolyn O. Frost. 1979. The use of citations in literary research: A preliminary classification of citation functions. *The Library Quarterly*, pages 399–414.

Olivier Galibert, Sophie Rosset, Xavier Tannier, and Fanny Grandry. 2010. Hybrid citation extraction from patents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 530–534.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. CiteSeer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98.

Anthony Grafton. 1999. *The Footnote: a Curious History*. Harvard University Press.

Björn Hammarfelt. 2011. Interdisciplinarity and the intellectual base of literature studies: citation analysis of highly cited monographs. *Scientometrics*, 86(3):705–725.

D. Heckmann, A. Frank, M. Arnold, P. Gietz, and C. Roth. 2014. Citation segmentation from sparse and noisy data: a joint inference approach with Markov logic networks. *Digital Scholarship in the Humanities*.

Richard Heinzkill. 1980. Characteristics of references in selected scholarly english literary journals. *The Library Quarterly*, pages 352–365.

Björn Hellqvist. 2009. Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, 61(2):310–318.

Young-Min Kim, Patrice Bellot, Elodie Faath, and Marin Dacos. 2012a. Annotated bibliographical reference corpora in Digital Humanities. In *Language Resources and Evaluation Conference*, pages 494–501.

Young-Min Kim, Patrice Bellot, Elodie Faath, and Marin Dacos. 2012b. Automatic annotation of incomplete and scattered bibliographical references in Digital Humanities papers. In *Conférence en Recherche de Information et Applications*, pages 329–340.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Vincent Larivière, Yves Gingras, and Éric Archambault. 2006. Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, 68(3):519–533.

A. J. M. Linmans. 2009. Why with bibliometrics the humanities does not need to be the weakest link: Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics*, 83(2):337–354.

Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries*, pages 473–474.

Patrice Lopez. 2010. Automatic extraction and resolution of bibliographical references in patent documents. In *Advances in Multidisciplinary Retrieval*, pages 120–135.

Hajime Murai and Akifumi Tokosumi. 2008. Extracting concepts from religious knowledge resources and constructing classic analysis systems. In *Large-Scale Knowledge Resources. Construction and Application*, pages 51–58.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Fuchun Peng and Andrew McCallum. 2006. Information extraction from research papers using Conditional Random Fields. *Information Processing & Management*, 42(4):963–979.

Matteo Romanello, Federico Boschetti, and Gregory Crane. 2009. Citations in the digital library of Classics: extracting canonical references by using Conditional Random Fields. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 80–87.

Matteo Romanello. 2013. Creating an annotated corpus for extracting canonical citations from classics-related texts by using active annotation. In *Computational Linguistics and Intelligent Text Processing*, volume 7816, pages 60–76.

Matteo Romanello. 2014. Mining citations, linking texts. *Institute for the Study of the Ancient World Papers 7.24*.

Chris A. Sula and Matt Miller. 2014. Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing*, 29(3):452–464.

Stephen E. Wiberley Jr. 2010. Humanities literatures and their users. In *Encyclopedia of Library and Information Sciences*, pages 2197–2204.

Visualising Italian Language Resources: a Snapshot

Riccardo Del Gratta, Francesca Frontini, Monica Monachini, Gabriella Pardelli, Irene Russo, Roberto Bartolini, Sara Goggi, Fahad Khan, Valeria Quochi, Claudia Soria, Nicoletta Calzolari
Istituto di Linguistica Computazionale "A . Zampolli"
CNR Pisa, Italy
name.surname@ilc.cnr.it

Abstract

English. This paper aims to provide a first snapshot of Italian Language Resources (LRs) and their uses by the community, as documented by the papers presented at two different conferences, LREC2014 and CLiC-it 2014. The data of the former were drawn from the LOD version of the LRE Map, while those of the latter come from manually analyzing the proceedings. The results are presented in the form of visual graphs and confirm the initial hypothesis that Italian LRs require concrete actions to enhance their visibility.

Italiano. *Questo articolo ha l'obiettivo di fornire una fotografia del contesto delle Risorse Linguistiche italiane e dei loro usi da parte della comunità scientifica; i dati usati sono tratti dagli articoli presentati a due diverse conferenze del settore, LREC2014 e CLiC-it 2014. I primi sono derivati dalla LRE Map in versione LOD, mentre i secondi sono stati ottenuti da un'analisi manuale degli atti della conferenza. I risultati sono presentati e analizzati sotto forma di grafi e confermano l'ipotesi che le risorse linguistiche italiane richiedano azioni mirate ad aumentare la loro visibilità.*

1 Introduction

The availability of Language Resources (LRs) - such as corpora, computational lexicons, parsers, etc. - is crucial to most NLP technologies (Machine Translation, Crosslingual Information Retrieval, Multilingual Information Extraction, Automatic Document Indexing, Question Answering, Natural Language Interfaces, etc.). Recent initiatives have monitored the availability of language resources for different languages, and highlighted a digital divide between English and other languages (Soria et al., 2012). While the economic potential of English ensures that English LRs are developed and maintained not only in the academic sector but also by commercial players, the involvement of research communities for languages such as Italian is much more crucial to ensure that the necessary instruments (both data and tools) are made available for natural language processing purposes.

At the same time, the production of quality LRs is just a first step; LRs must also be documented and made available to the community in such a way that they are easy to find and to use. This entails the description of every LR with a set of metadata that clarify its typology, its language, its size and licensing scheme, and the means of accessing it. Useful information in this sense can be found in the catalogues of language resources associations, such as ELRA, LDC, NICT Universal Catalogue, ACL Data and Code Repository, OLAC, LT World. These catalogues adopt a top-down approach to documenting resources and typically list resources that have reached a high level of maturity - in term of validation, documentation, clearing of IPR issues, etc. As an alternative to this approach, recent projects have been carried out within the LR community to create open, bottom-up repositories where LRs - even those under development - can be duly documented and searched. Such initiatives are for instance the META-SHARE platform (Gavrilidou et al., 2012), the CLARIN VLO (Broeder et al., 2010) and the LRE Map (Calzolari et al., 2012; Del Gratta et al., 2014b; Del Gratta et al., 2014a), with their sets of metadata. In particular the LRE Map was launched as an initiative at LREC2010 in order to crowdsource reliable and accurate documentation for the largest possible set of resources. Au-

thors submitting to that conference were asked to document the resources they used in their paper, both the resources they created and the ones created by others. This initiative has continued and been extended to other conferences[1], and is now a unique source of information on existing language resources and their use in current research. The work in this paper can be set against the background of the major projects in which CNR ILC is currently involved and the aim of setting up a documentation center for language and textual resources within the framework of the CLARIN and DARIAH research infrastructures. As a CLARIN and DARIAH node, CNR ILC has the task of collecting and harmonizing metadata description of LRs at a national level, making Italian resources more visible to national and international research groups, both to the NLP and to the digital humanities communities. To this purpose, our team has inspected the panorama of LR descriptions available in the aforementioned catalogues, and in particular the LRE Map which allows us to monitor how communities build around LR use. Our hypothesis is that many of the resources that the Italian community uses and produces are not as well documented as they should be. As a consequence, many researchers may not be aware of the existence of resources that could be of use for them, and limit themselves to those they know best. In order to verify this, we carried out a cross-analysis of Italian LRs and their uses by Italian researchers, exploiting the data found in the LRE Map from the LREC2014 dataset, which is currently available in LOD format (Del Gratta et al., 2014a). Such data is compared with similar evidence gathered from the proceedings of the CLiC-it 2014 conference, which are available online. CLiC-it 2014 did not adhere to the LRE Map initiative, but comparable information has been collected by manually inspecting the papers. In what follows we will provide a brief description of the set of metadata that we used to monitor the situation with respect to Italian LRs and their use; then some results will be analyzed and discussed by means of graph-like visualizations; finally some conclusions are drawn and perspectives for future work outlined.

[1]Such as COLING, EMNLP, ACL-HLT, RANLP, Interspeech, Oriental-Cocosda, IJCNLP, LTC, NA-ACL

2 Metadata description

The set of metadata used for documenting language resources can vary from repository to repository. Some harmonization initiatives are currently being carried out in order to make diverse datasets interoperable, e.g. (McCrae et al., 2015). Nevertheless a common core has been broadly agreed upon by all; this includes type of resource (corpus, lexicon, tool), modality, language(s), use, availability. To this core set of metadata, the LRE Map adds other metadata that are linked not to the resource itself, but to its use in the paper that is being submitted: thus information about the conference, the paper, the authors and their affiliations is available for each entry in the LRE Map. This also means that any given resource can have more than just one entry in the LRE Map, one for each paper that has used it. Sometimes the resource is marked as new, and in that case we can assume that the authors of the paper are also the producers of this new resource; in most cases the resource is a well known one. So for instance some of the most used resources according to the LRE Map are Princeton WordNet and the British National Corpus. For the purposes of this paper we only took into consideration the following metadata for each entry in the LREC2014 LRE Map: resource name, language, authors and affiliations. We extracted all used LRs with Italian as one of the languages and authors with an Italian affiliation. We then analysed the proceedings of CLiC-it 2014 and manually extracted the same type of information for each paper [2]. We thus obtained two datasets:

Table 1: LRs use - the Italian panorama.

	Authors	LRs	Institutions	Papers
LREC '14	91	25	41	24
CLiC-it '14	107	54	28	42
Total '14	166	74	57	66

[2]One of the most interesting features of the LRE Map is the fact that it provides a user's perspective on language resources. So for instance Princeton WordNet may be defined by some as a lexicon and by others as an ontology; moreover the declared use may vary from paper to paper. In the case of the CLiC-it dataset the data was collected by just one person, and thus this precious information is not available. For this reason this data cannot be inserted into the LRE Map and has to be considered as a simulation.

3 People and Resources: visualising networks

Data visualisation is a method that enables the exploration, filtering and searching of data, skipping the interaction with databases. Data can be mainly visualised for presentation or exploration but in well designed projects there is a continuum between these two modalities (Cairo, 2013).
In this paper we propose two visualisation modalities to discover the interrelations between authors from different institutions and the convergence of authors on the usage of the same resource. In comparing these two conferences the aim was to portray the Italian NLP community highlighting collaborations between people through resources used.
The implementation of the visualisation is based on a well known tool, D3.js, a JavaScript library designed to display digital data in a dynamic graphical form. The two visualisations are:

- a force-directed graph (see a detail in Figure 1)[3] where each author is a node; the links between author-nodes stand for co-authorship in a paper. Different institutions are assigned different colours; in this way people belonging to the same institution are visually identifiable and collaborations among institutions are clear because of the links connecting coauthors of different colours: for example Cristina Bosco from the University of Turin is connected to co-authors from the same institution (purple dots) but also to Maria Simi from the University of Pisa and Simonetta Montemagni from ILC CNR (orange and brown dot, respectively).

- a force-directed graph where each author is a node connected to other persons only through the resources they use, depicted as boxes. Here too, the colour of the person depends on the institution. People are connected to the same resource (1) when they co-authored a paper that uses it, (2) because they use the same resource in independent research works. In the first case, co-author groups are still somewhat identifiable, as they create an island effect (as shown in Figure 2). In the other case heterogeneous people get connected because they use the same resources. As a result, networks of researchers are gathered around LR uses (see Figure 3).

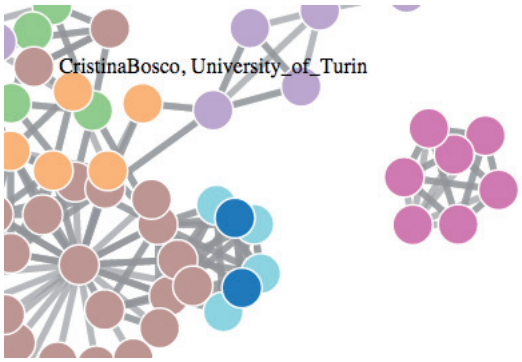


Figure 1: Cross institution co-author networks.

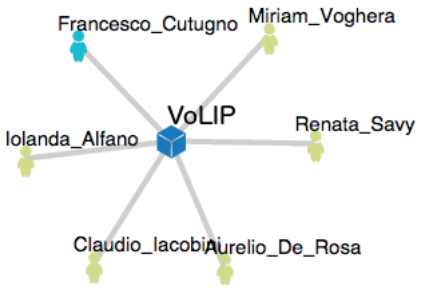


Figure 2: Same resource used by co-authored paper.

Graph-based visualisations pave the way for a social network analysis of the data that we plan as future work. For the moment, thanks to these two graphical devices, some interesting phenomena are now visually evident; we concentrate in particular on how research collaborations gather around LRs. The first phenomenon is that at the LREC2014 there are more international collaborations between Italian and foreign groups. The first edition of CLiC-it instead presents less involvement of foreign co-authors and more collaborations between different Italian institutions. This is clearly due to the fact that CLiC-it is a national conference, while LREC an international one. The second fact is that at LREC2014 we find a smaller number of Italian LRs, as typically papers use the best known ones. CLiC-it instead presents us with a broader panorama: in addition to the best known resources we find a plethora of minor resources -

[3]The interactive visualisations are available online at http://www.clarin-it.it/jvis

in particular corpora - that are not mentioned in the LREC2014 dataset and are mostly used in a single paper. In many cases the user of the resource is also its creator: these resources need documentations to foster future collaborations. Graph-based

4 Conclusions and future works

In this work we use visualisations to show how the Italian NLP community uses LRs in the works presented at two recent conferences of the sector (LREC2014 and CLiC-it 2014). We highlight how collaborations cluster around the use of major resources, and how networks are created by users of the same resource. From the comparison of the two datasets we can infer that the Italian panorama of language resources is rich and varied. We also confirm the prior hypothesis that Italian LRs are rather under-documented and that some positive action is needed in the direction of enhancing their visibility. As a consequence the creation of an observatory of Italian language resources, which is meant to be the nucleus of a newly established CLARIN-IT center, is more than justified. Such an observatory will actively promote the Italian LR community (both creators and users), help in improving the documentation of LRs thus making them more widely known to others and finally ensure their visibility in an international context by using all current standard metadata framework and platforms. This latter point shall involve also an active contribution to the de-fragmentation of the current situation in metadata and description practices, as well as the porting of LR descriptions to emerging channels and formats (LINGhub[4], RDF-LOD).

Acknowledgments

The research carried out in this paper was partly funded by SM@RTINFRA (MIUR Progetto premiale) and PARTHENOS (H2020 INFRADEV-4).

References

Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A data category registry-and component-based metadata framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 43–47. European Language Resources Association (ELRA).

Alberto Cairo. 2013. *L'arte funzionale: Infografica e visualizzazione delle informazion*. Pearson Italia Spa.

Nicoletta Calzolari, Riccardo Del Gratta, Gil Francopoulo, Joseph Mariani, Francesco Rubino, Irene Russo, and Claudia Soria. 2012. The LRE Map. Harmonising Community Descriptions of Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1084–1089. European Language Resources Association (ELRA).

Riccardo Del Gratta, Francesca Frontini, A Fahad Khan, Joseph Mariani, and Claudia Soria. 2014a. The LRE Map for under-resourced languages. In *Workshop Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era, Satellite Workshop of LREC'14*.

Riccardo Del Gratta, F Khan, Sara Goggi, and G Pardelli. 2014b. LRE Map disclosed. In *Proceedings of the ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Harris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, et al. 2012. The META-SHARE Metadata Schema for the Description of Language Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1090–1097. European Language Resources Association (ELRA).

John McCrae, Penny Labropoulou, Jorge Gracia, Marta Villegas, Victor Rodriguez-Doncel, and Philipp Cimiano. 2015. One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. In *Proceedings of the 4th Workshop on the Multilingual Semantic Web*.

Claudia Soria, Nria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, Nicoletta Calzolari, and others. 2012. The FLaReNet Strategic Language Resource Agenda. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1379–1386. European Language Resources Association (ELRA).

[4] http://linghub.lider-project.eu/

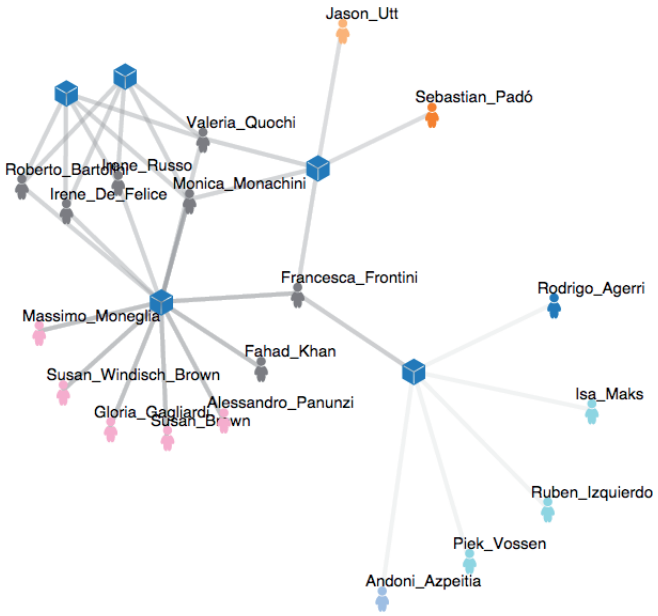


Figure 3: Same resource used in different papers (LREC2014).

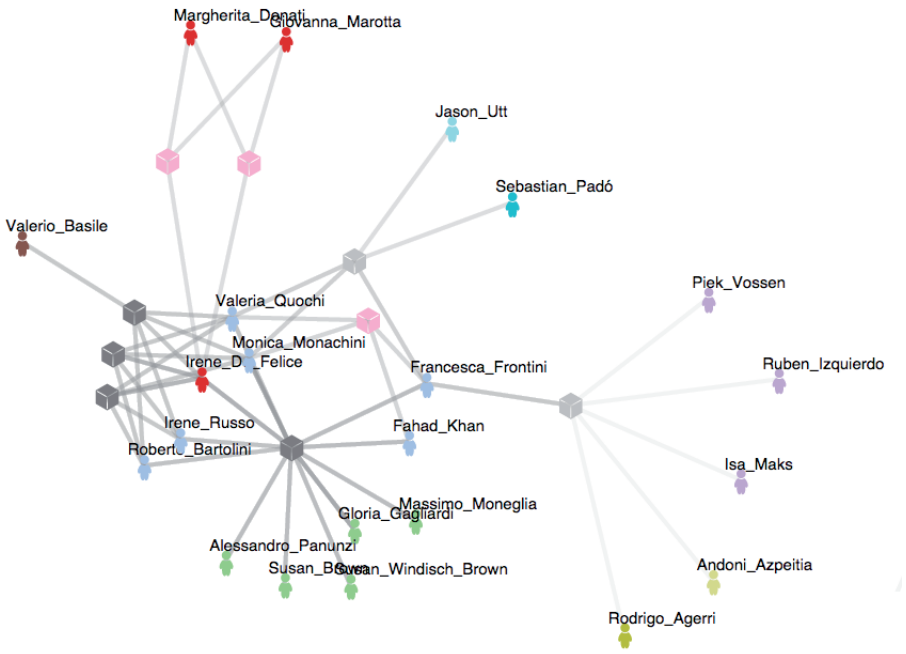


Figure 4: Both conferences together.

A manually-annotated Italian corpus for fine-grained sentiment analysis

Marilena Di Bari, Serge Sharoff, Martin Thomas
University of Leeds
School of Languages, Cultures and Societies
LS2 9JT, Leeds (UK)
{mlmdb,s.sharoff,m.thomas}@leeds.ac.uk

Abstract

English. This paper presents the results of the annotation carried out on the Italian section of the *SentiML corpus*, consisting of both originally-produced and translated texts of different types. The two main advantages are that: (i) the work relies on the linguistically-motivated assumption that, by encapsulating opinions in pairs (called *appraisal groups*), it is possible to annotate (and automatically extract) their sentiment in context; (ii) it is possible to compare Italian to its English and Russian counterparts, as well as to extend the annotation to other languages.

Italiano. *Questo lavoro presenta i risultati dell'annotazione effettuata sulla sezione italiana del corpus "SentiML", che consiste di testi sia originali che tradotti appartenenti a diversi tipi. I due vantaggi principali sono che: (i) il lavoro si fonda sull'assunzione motivata linguisticamente che, codificando le opinioni in coppie (chiamate* appraisal groups*), è possibile annotare (ed estrarre automaticamente) il loro sentiment tenendo in considerazione il contesto; (ii) è possibile confrontare l'italiano con le sue controparti inglese e russa, ed estendere l'annotazione ad altre lingue.*

1 Introduction

Overall, the field of Sentiment Analysis (SA) aims at automatically classifying opinions as positive, negative or neutral (Liu, 2012). While at first the focus of SA was on the document level (coarse-grained) classification, with the years it has become more and more at the sentence level or below the sentence (fine-grained). This shift has been due to both linguistic and application reasons. Linguistic reasons arise because sentiment is often expressed over specific entities rather than an overall document. As for practical reasons, SA tasks are often aimed at discriminating between more specific aspects of these entities. For example, if an opinion is supposed to be on the plot of a movie, it is not unusual that the user also evaluates actors' performance or director's choices (Shastri et al., 2010). For SA applications these opinions need to be assessed separately. Also opinions are not expressed as simple and direct assertions, but by using a number of stylistic devices such as pronominal references, abbreviations, idioms and metaphors. Finally, the automatic identification of sarcasm, irony and humour is even more challenging (Carvalho et al., 2009).

For all these reasons, fine-grained sentiment analysis is looking at entities that are usually chains of words such as "noun+verb+adjective" (e.g. the house is beautiful) or "adverb+adjective+noun" (e.g. very nice car) (Yi et al., 2003; Popescu and Etzioni, 2005; Choi et al., 2006; Wilson, 2008; Liu and Seneff, 2009; Qu et al., 2010; Johansson and Moschitti, 2013).

In addition to the multitude of approaches to fine-grained SA, there is also shortage of multilingual comparable studies and available resources. To close this gap, we designed the *SentiML* annotation scheme (Di Bari et al., 2013) and applied it to texts in three languages, English, Italian and Russian. The proposed annotation scheme extends previous works (Argamon et al., 2007; Bloom and Argamon, 2009) and allows multi-level annotations of three categories: *target* (T) (expression the sentiment refers to), *modifier* (M) (expression conveying the sentiment) and *appraisal group* (AG) (couple of modifier and target). For example in:

> "Gli uomini hanno il potere di [[sradicare]$_{\mathrm{M}}$ la [povertà]$_{\mathrm{T}}$]$_{\mathrm{AG}}$, ma anche di [[sradicare]$_{\mathrm{M}}$ le

[tradizioni]$_T$]$_{AG}$".

(Men have the power to eradicate poverty, but also to eradicate traditions)

the groups "sradicare povertà" (eradicate poverty) and "sradicare tradizioni" (eradicate traditions) have an opposite sentiment despite including the same word *sradicare* (to eradicate).

This scheme has been developed in order to facilitate the annotation of the sentiment and other advanced linguistic features that contribute to it, but also the appraisal type according to the *Appraisal Framework* (AF) (Martin and White, 2005) in a multilingual perspective (Italian, English and Russian). The AF is the development of the *Systemic Functional Linguistics* (Halliday, 1994) specifically concerned with the study of the language of evaluation, attitude and emotion. It consists of *attitude*, *engagement* and *graduation*. Of these, *attitude* is sub-divided into *affect*, which deals with personal emotions and opinions (e.g. *excited*, *lucky*); *judgement*, which concerns author's attitude towards people's behaviour (e.g. *nasty*, *blame*); *appreciation*, which considers the evaluation of things (e.g. *unsuitable*, *comfortable*). The *engagement* system considers the positioning of oneself with respect to the opinions of others, while *graduation* investigates how the use of language amplifies or diminishes attitude and engagement. In particular, *force* is related to intensity, quantity and temporality. To the best of our knowledge the AF has only been applied in the case of Italian for purposes not related to computation (Pounds, 2010; Manfredi, 2011).

This paper is organized as follows: Section 2 describes the annotation scheme and the annotated Italian corpus, Section 3 reports the results and finally Section 4 our conclusions.

2 Annotation scheme and corpus

The scheme, described in (Di Bari et al., 2013), specifies different attributes for the categories *target*, *modifier* and *appraisal group*.

A target is usually a noun. Targets have 2 attributes: *type* ('person', 'thing', 'place', 'action' and 'other'), and prior (out-of-context) *orientation* ('positive', 'negative', 'neutral' and 'ambiguous').

A modifier is what *modifies* the target. It can be an adjective, a verb, an adverb or a noun in the case of two nous linked by a preposition, e.g. "libertà di parola" (freedom of speech). Modifiers have 4 attributes: *attitude* ('affect', 'judgement' and 'appreciation'); *force* referring to the intensity of the modifier, i.e. high like in the case of "molto bella" (very beautiful), 'low' like in the case of "poco elegante" (little elegant), 'reverse' like in the case of "contro la guerra" (against the war) or 'normal'; *polarity* if there is a negation ('marked') or not ('unmarked'), and prior (out-of-context) *orientation* ('positive', 'negative', 'neutral' and 'ambiguous').

Appraisal groups have 1 attribute: contextual *orientation* ('positive', 'negative', 'neutral' and 'ambiguous').

In the example sentence shown in Section 1, the modifier *sradicare* would thus have attitude 'judgement', force 'normal' , polarity 'unmarked', orientation 'ambiguous'; the target *povertà* would have type 'thing' and orientation 'negative', whereas the target *tradizioni* would have type 'thing' and orientation 'positive'; the appraisal group "sradicare povertà" would have orientation 'positive', while the appraisal group "sradicare tradizioni" would have orientation 'negative'.

SentiML has been applied to the text types different from those taken into account in previous works in Italian (Casoto et al., 2008; Basile and Nissim, 2013; Bosco et al., 2013; Sorgente et al., 2014):

- **Political speeches**. Translations of American presidents' addresses.
- **Talks**. Translations of TED (Technology, Entertainment, Design) talks (Cettolo et al., 2012).
- **News**. Belonging to the newspaper *Sole24ore*.

The corpora have been annotated by using MAE (Stubbs, 2011), a freely available software annotation environment. The Italian corpus contains 328 sentences for a total of 9080 tokens. To deal with the limitation of having only one annotator, different confidence-rated machine learning classifiers were used to spot inconsistencies and thus revise the annotations accordingly ((Di Bari et al., 2014)).

3 Results of the annotation

In Table 1 details about the number of the appraisal groups, targets and modifiers are shown,

Language	Text type	Appraisal groups	Targets	Modifiers	% of words included in appraisal groups
ITA	Political	486	411	437	25%
	News	254	203	244	22%
	TED	341	292	323	24%
	tot	1081	906	1004	24%

Table 1: Statistics on the annotated data. A different amount of appraisal groups has been annotated according to the text type, but on average the 24% of words are sentiment-loaded.

along with the percentages of words embedded in appraisal groups for each text type.

Figure 1 shows that 'positive' orientation is the predominant one for appraisal groups with 67%, followed by 'negative' with 32%. These data are consistent with the assumption that appraisal groups should not be 'neutral' nor 'ambiguous' because they carry appraisal and their orientation should be clear in context. At the same time, targets and modifiers can be 'ambiguous' because their orientation depends on the context and 'neutral' in case they are not the element carrying appraisal in the group.

Figure 2 shows the statistics on the other attributes: 'appreciation' is the most common attitude, which is consistent with the fact that this value is associate to 'thing' in the AF (see Section 1), which is the most common target type; polarity, which indicates that a negation has been encountered, has been 'marked' 4% times; force, an important feature for a more accurate prediction of the sentiment, is 'reverse' 4% of times.

We have also compared the contextual orientation manually annotated by us with the prior orientation included in the translation of the 'positive' and 'negative' values in the *NRC Word-Emotion Association Lexicon* (Mohammad, 2011), whose English annotations were manually done through *Amazon's Mechanical Turk*, and the *Roget Thesaurus* and it has entries for about 14200 word types. We calculated that, in the case of Italian, only 29.39% of the words belonging to the appraisal groups were present in the sentiment dictionary, with higher percentage for political speeches (33.54%), followed by news (27.66%) and TED talks (26.98%). As previously found in the case of English, most of these are nouns reasonably not carrying sentiment on their own, but still part of an appraisal group (e.g., *brevetti* (patents), *computer*, *confini* (borders), *nostro* (our)). There are also cases, adjectives in particular, that should probably be included in a dictionary with prior orientation (e.g., *necessario* (necessary), *negativo* (negative), *oberato* (overburdened), *ideale* (ideal)).

In line with our previous experiments in English (Di Bari et al., 2013), we used the following categories for the comparison:

Agreeing words: words whose dictionary orientation agrees with that of the appraisal group they belong to. They cover 69.63% of the total times words were found in the dictionary. This means that we can rely to a certain extent to the dictionary orientation, but not if we aim at more accuracy. The list includes reasonable out-of-context positive words (e.g., *alleati* (allies), *comprensione* (comprehension), *dotato* (gifted), *felicità* (happiness)), as well as out-of-context negative words (e.g., *debolezza* (weakness), *malattia* (sickness), *stagnante* (stagnant), *violenza* (violence)).

Disagreeing words: words whose dictionary orientation does not agree with that of the appraisal group they belong to. They cover 28.18% of the total times words were found in the dictionary, a percentage that demonstrates how crucial the context is. For example reversals such as *abolire* (abolish) and *diminuire* (diminish), and *sfida* (challenge), *sopportare* (to bear), *tendenza* (trend). However, it was interesting to notice that also words normally considered positive (e.g. *prosperare* (to prosper) and *risorse* (resources)) or negative (e.g. and *tensione* (tension) and *rischio* (risk)) became included in groups with opposite orientation.

Ambiguous words: words which already have both positive and negative values in the dictionary. They are *resta* (stays), *rivoluzione* (revolution), *sciogliere* (to unleash), *umile* (humble), and they cover 1.07%.

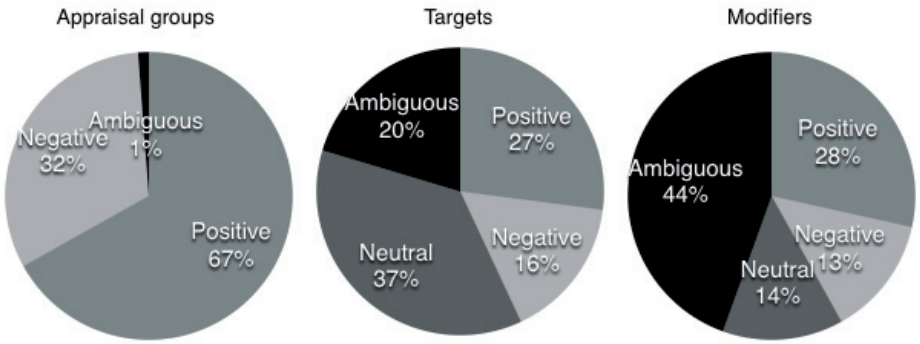


Figure 1: Values for the attribute orientation for appraisal groups, targets and modifiers. In the case of appraisal groups, positive is the most common value, followed by negative.

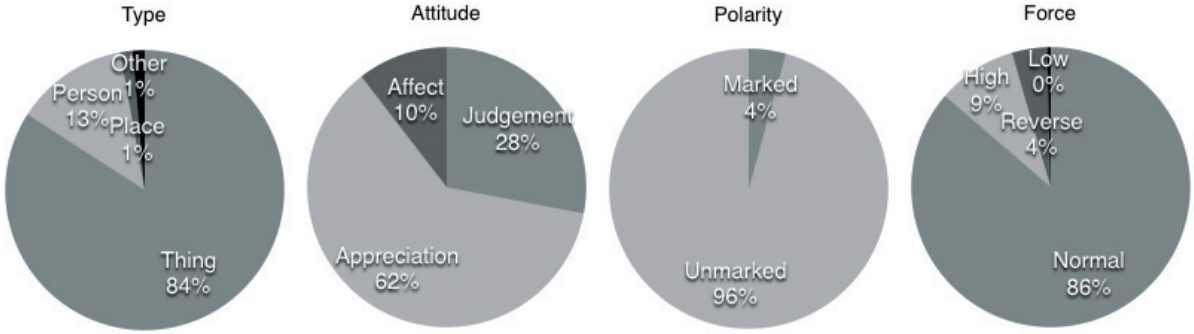


Figure 2: Values for the attributes attitude, polarity, force and type.

4 Conclusions

In this paper we have described a manually-annotated corpus of Italian for fine-grained sentiment analysis. The manual annotation has been done in order to include important linguistic features. Apart from extracting statistics related to the annotations, we have also compared the manual annotations to a sentiment dictionary and demonstrated that (i) the dictionary includes only 29.29% of the annotated words, and (ii) the prior orientation given in the dictionary is different from the correct one given by the context in 28.18% of the cases.

The original and annotated texts in Italian (along with English and Russian) and the Document Type Definition (DTD) of *SentiML* to be used with MAE are publicly available[1].

In the meanwhile, the authors are already working on an automatic system to identify and classify appraisal groups multilingually.

[1] `http://corpus.leeds.ac.uk/marilena/SentiML`

Acknowledgments

The first author would like to thank Michele Filannino (The University of Manchester) for his insights throughout the research.

References

S. Argamon, K. Bloom, A. Esuli, and F. Sebastiani. 2007. Automatically Determining Attitude Type and Force for Sentiment Analysis. In *Proceedings of the 3rd Language and Technology Conference (LTC'07)*, pages 369–373, Poznan, Poland.

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.

Kenneth Bloom and Shlomo Argamon. 2009. Automated learning of appraisal extraction patterns. *Language and Computers*, 71(1):249–260.

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.

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" ;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, TSA '09, pages 53–56, New York, NY, USA. ACM.

Paolo Casoto, Antonina Dattolo, and Carlo Tasso. 2008. Sentiment classification for the italian language: A case study on movie reviews. *Journal of Internet Technology*, 9(4):365–373.

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.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 431–439, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marilena Di Bari, Serge Sharoff, and Martin Thomas. 2013. SentiML: Functional annotation for multilingual sentiment analysis. In *DH-case 2013: Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities*, ACM International Conference Proceedings.

Marilena Di Bari, Serge Sharoff, and Martin Thomas. 2014. Multiple views as aid to linguistic annotation error analysis. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII)*. ACL SIGANN Workshop held in conjunction with Coling 2014.

M.A.K. Halliday. 1994. *An Introduction to Systemic Functional Linguistics*. London:Arnold, 2 edition.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3).

Jingjing Liu and Stephanie Seneff. 2009. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 161–169, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Marina Manfredi. 2011. Systemic functional linguistics as a tool for translation teaching: towards a meaningful practice. *Rivista Internazionale di Tecnica della Traduzione*, 13:49 – 62.

James R Martin and Peter RR White. 2005. *The language of evaluation*. Palgrave Macmillan, Basingstoke and New York.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA, June.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gabrina Pounds. 2010. Attitude and subjectivity in italian and british hard-news reporting: The construction of a culture-specific 'reporter'voice. *Discourse Studies*, 12(1):106–137.

Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 913–921, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lokendra Shastri, Anju G. Parvathy, Abhishek Kumar, John Wesley, and Rajesh Balakrishnan. 2010. Sentiment extraction: Integrating statistical parsing, semantic analysis, and common sense reasoning. In *IAAI*.

Antonio Sorgente, Via Campi Flegrei, Giuseppe Vettigli, and Francesco Mele. 2014. An italian corpus for aspect based sentiment analysis of movie reviews. In *First Italian Conference on Computational Linguistics CLiC-it*.

Amber Stubbs. 2011. Mae and mai: Lightweight annotation and adjudication tools. In *Linguistic Annotation Workshop*, pages 129–133.

Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, University of Pittsburgh.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE.

From a Lexical to a Semantic Distributional Hypothesis

Luigi Di Caro[†], **Guido Boella**[†], **Alice Ruggeri**[†],
Loredana Cupi[†], **John Adebayo Kolawole**[‡], **Livio Robaldo**[⋆]
[†] University of Torino
[⋆] University of Luxembourg
[‡] University of Bologna
{dicaro, boella,ruggeri}@di.unito.it
loredana.cupi@unito.it,kolawolejohn.adebayo@unibo.it
livio.robaldo@uni.lu

Abstract

English. Distributional Semantics is based on the idea of extracting *semantic* information from *lexical* information in (multilingual) corpora using statistical algorithms. This paper presents the challenging aim of the *SemBurst* research project[1] which applies distributional methods not only to words, but to sets of semantic information taken from existing semantic resources and associated with words in syntactic contexts. The idea is to inject semantics into vector space models to find correlations between statements (rather than between words). The proposal may have strong impact on key applications such as Word Sense Disambiguation, Textual Entailment, and others.

Italiano. *La semantica distribuzionale si basa sull'idea di estrarre automaticamente informazione semantica attraverso algoritmi statistici applicati ad occorrenze lessicali in grandi corpora. Questo articolo presenta l'idea del progetto SemBurst che applica metodi distribuzionali non solo alle parole, ma ad insiemi di informazioni semantiche tratte da risorse semantiche disponibili e associate alle parole nei relativi contesti sintattici. Lo scopo e' quello infatti di iniettare semantica negli spazi vettoriali per trovare correlazioni tra informazioni semantiche (piuttosto che tra elementi lessicali). Questo nuovo approccio potra' avere un alto impatto su applicazioni chiave come Word Sense Disambiguation, Textual Entailment, e altri.*

[1]Semantic Burst: Embodying Semantic Resources in Vector Space Models, financed by Compagnia di San Paolo - cod. 2014_L1_272.

1 Introduction and Background

One of the main current research frontiers in Computational Linguistics is represented by studies and techniques usually associated with the label Distributional Semantics (DS), which are focused on the exploitation of distributional analyses of words in syntactic compositions. Their importance is demonstrated by recent ERC projects (COMPOSES and DisCoTex[2]) and by a growing research interest in the scientific community[3]. The proposal presented in this paper is about going far beyond this state of the art.

DS uses traditional Data Mining (DM) techniques on text, considering language as a grammar-based type of data, instead of simple unstructured sequences of tokens. It quantifies semantic (in truth lexical) similarities between linguistically-refined tokens (words, lemmas, parts-of-speech, etc.), based on their distributional properties in large corpora. DM relies on Vector Space Models (VSMs), a representation of textual information as vectors of numeric values (Salton et al., 1975). DM techniques such as Latent Semantic Analysis (LSA) have been successfully applied to text for information indexing and extraction tasks, using matrix decompositions such as Singular Value Decomposition (SVD) to reconstruct the latent structure behind the distributional hypothesis (Deerwester et al., 1990). It usually works by evaluating the relatedness of different terms, forming word clusters sharing similar contexts. Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) and Salient Semantic Analysis (SSA) (Hassan and Mihalcea, 2011) revisits these methods in the way they define the conceptual layer. With LSA a word's hidden

[2]European Research Council projects nr. 283554 (COMPOSES) and nr. 306920 (DisCoTex).

[3]Clark, S. Vector space models of lexical meaning. A draft chapter of the Wiley-Blackwell Handbook of Contemporary Semantics second edition.

concept is based on its surrounding words, with ESA it is based on Wikipedia entries, and with SSA it is based on hyperlinked words in Wikipedia entries. These approaches represent only a partial step towards the use of semantic information as input for Distributional Analysis.

While distributional representations excel at modelling lexical semantic phenomena such as word similarity and categorization (conceptual aspect), Formal Semantics in Computational Linguistics focuses on the representation of the meaning in a set theoretic way (functional aspect), providing a systematic treatment of compositionality and reasoning. Recent interest in the combination of Formal Semantics and Distributional Semantics have been proposed (Lewis and Steedman, 2013) (Turney, 2012) (Garrette et al., 2014), that employ approaches based on the lexical level. However, 1) the problem of compositionality of lexical distributional vectors is still open and the proposed solutions are limited to combination of vectors, 2) reasoning on classic distributional representations is not possible, since they are VSMs at the lexical level only, 3) the connection of DS with traditional Formal Semantics is not straightforward (Turney, 2012) (Garrette et al., 2014) since DS is limited to a semantics of similarity which is able to support retrieval but not other aspects such as reasoning; and 4) DS does not scale up to phrases and sentences due to data sparseness and growth in model size (Turney, 2012), restraining the use of tensors.

2 A *Semantic* Distributional Hypothesis

This proposal is based on the idea of applying distributional analysis not only to words but also to sets of semantic features taken from semantic resources. The idea is that the semantic information injected into an input text corpus will act as a catalyst to facilitate the creation of further semantic information and to find correlations with semantic features of other words in their syntactic context. For instance, the word "*cat*" in "*the cat bites the mouse*" will be replaced by physical facts (it has claws, paws, eyes, whiskers, etc.), behavioural information (it chases mice, it is capable of climbing up a tree, etc.), taxonomical information (it is a feline, it is a predator, etc.), habitats, etc. This will create a new multi-dimensional semantic search space where distributional analysis will be used to clean up and correlate statements rather than words, for example, finding the relation between a carnivore-subject and a meat-object in the sentence "*The cat bites the mouse*" or between a cat's claws and the act of climbing in the sentence "*The cat climbs the tree*".

2.1 Feasibility

The proposed shift to *semantics* as input for distributional analysis is now feasible due to the large number of semantic resources available such as BabelNet (Navigli and Ponzetto, 2010), ConceptNet (Speer and Havasi, 2012), FrameNet (Baker et al., 1998), DBPedia, etc. However, these resources are sometimes incomplete, contradictory, ambiguous, and difficult to integrate together, so they cannot be used in Formal Semantics. Formal Semantics handles reasoning, quantification, and compositionality of meaning using set-theoretic models, and therefore requires data consistency. The aim of this proposal is to overcome these problems by applying the distributional hypothesis to the partial and contradictory semantic information that can be associated with words contained in large corpora and structured in syntactic contexts, in the same way it has been successfully applied to words in the last few decades. For example, if a corpus contains ambiguities and other noise, this does not prevent distributional analysis on words, because the calculations use the most significant data. Analogously, in case of a few ambiguities and contradictions in the semantic resources, a distributional approach using several resources and advances in Data Mining will manage to derive the most probable relations between statements.

2.2 Research Objectives

The presented approach is intended to reduce the existing gap between Distributional Semantics and Formal Semantics by creating a novel type of semantics, still distributional, but working on a semantic rather than lexical input. The idea is articulated in the following sub-objectives: 1) to acquire and integrate semantic information from different resources; 2) to create not only distributional word representations, but also distributional representations of semantic features with tensors. By moving to the semantic level, it will help overcome the problem of sparseness in classic word-based tensors. Since semantic information represents knowledge shared by multiple words, this proposal will allow to consider more complex syntactic structures to be considered than currently practiced. Then, it aims to 3) deal with compo-

sitionality at a more appropriate level - no longer as a fusion of lexical distribution vectors, but as a fusion of semantic features and 4) will enable reasoning on the semantic representation built via distributional vectors of semantic features. Further semantic resources will be created, which can be re-injected several times as input into the distributional analysis, thus 5) creating a positive loop of expanding knowledge. The proposal can also 6) consider multilingual contexts where semantic resources are not available, 7) finally reframing tasks as later described in Section 3.6.

3 Project Architecture

3.1 Data Acquisition

The first step required by this proposal is to aggregate linguistic and semantic resources such as ConceptNet, FrameNet, WordNet, BabelNet, etc. The result will be a semantic database (SDB) of lexical and semantic information. This will require integration of data from different sources with problems such as alignment, conflict resolution, and granularity mismatch. The second step regards the expansion of an input corpus (result of a selection from existing available corpora) with the semantic information contained in SDB for each of its words. Let us assume a word w_i in SDB can be associated with a set $\sigma_i = \{< rel_a, c_1 >, < rel_b, c_2 >, ..., < rel_k, c_n >\}$ of semantic features of the type $< rel, c_j >$ to mean that word w_i has a relation rel with concept c_j in some semantic resource (e.g., w_i=*cat* and $\sigma_i = \{< isA$, MAMMAL$>, < capableOf$, JUMP$>, ...\}$). This word-by-facts replacement can be iterated multiple times over the concepts in σ_i (e.g., c_j=$MAMMAL$ in σ_i can enrich σ_i itself to σ_k with σ_{MAMMAL} such that σ_k = $\sigma_i \bigcup \{< isA$, ANIMAL$>, < capableOf$, BREATH$>, ...\}$. Given a sentence S, the idea is to enrich each w_i with σ_i so that we build a different and richer input for distributional analysis than in traditional approaches (see Figure 1).

3.2 Distributional Analysis of Semantics

The second part of the proposal concerns the use of advanced DM techniques such as tensor-based representations (of semantics, rather than words) by embodying syntactic roles (subjects, modifiers, verbs, and arguments) into its dimensions (see Figure 1). The complexity of algorithms for tensors is a major challenge in this level, although recent research has shown that background information can improve this issue (Schifanella et al., 2014). Advanced data analysis techniques on tensors allow operations that are suitable for the aim of this ongoing research project. In particular, the problem of correlating lexical items will be reframed as the problem of correlating sets of semantic features within syntactic structures, using similarity and correlation measures over tensors to align, merge and filter data items.

3.3 Compositionality and reasoning

The proposal allows to address the compositionality problem at a semantic level. Let us consider the adjective-noun collocation "*dead parrot*". Parrots are pets, but dead parrots are not. This is an example of complicated compositionality (Kruszewski and Baroni, 2014). Unlike e.g., "*blue parrot*", the adjective overrides typical features of the noun it is associated with. Currently, Distributional Semantics uses to model compositionality by merging word distribution vectors (Mitchell and Lapata, 2008; Grefenstette and Sadrzadeh, 2011), hopefully lowering the frequency of collocations where the phrase "*dead parrot*" occurs as a pet. In our approach, instead, we reframe the problem as: how can distributional analysis handle the fact that the semantic feature $< hasProperty$, NOT-ALIVE$>$ associated with the word "*dead*" overrides the necessary feature of the role of pet ($< isA$, PET$>$), i.e., $< hasProperty$, ALIVE$>$ played by "*parrot*"? Moreover, we can apply reasoning on the resulting semantic representation of "*dead parrot*": since the property NOT-ALIVE in semantic resources is associated with $< hasProperty$, NO-MOVE$>$, we can also predict that, for example, "*the dead parrot flies*" is not a proper sentence since FLY in $< capableOf$, FLY$>$ is associated with $< isA$, MOVE$>$.

3.4 Extension of Semantic Resources

A distributional analysis over the acquired semantic information can create novel semantic resources with the following radically new aspects. First, semantics will assume the form of combinations of statements within syntactic contexts, thus generalizing over concepts which could not be found even in very large corpora. Assume, for instance, that "*cat*" is not associated with the semantic feature $< has$, CLAWS$>$: we can add this feature to the word "*cat*" if it occurs in contexts where the distinguishing feature for climb-

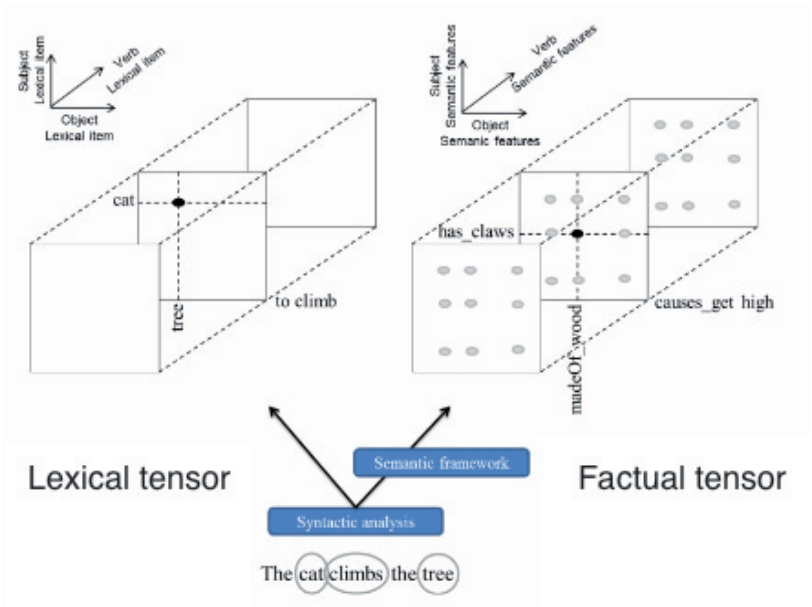


Figure 1: Distributional representation of natural language based on statements rather than lexical items.

ing is using claws ("*the * climbs the mast*", "*the * climbs the curtains*", etc.); moreover, the extended resources will be used again thus creating a positive loop of semantic feedback.

3.5 Multilingual Mapping

Multilingualism can be better managed since semantic features represent conceptual rather than lexical information units. When semantic resources are missing in one language, the proposed approach will use those of the English language, using automated translation from the target language to English. Ambiguities and errors will be introduced, but analyses on large numbers will hopefully manage the situation, allowing the creation of semantic knowledge for new languages.

3.6 Exploitation

Word Sense Disambiguation (WSD). Instead of linking words to word senses (a priori defined in resources such as WordNet) by exploring word-based contexts, we will replace each word with all the semantic features of all its uses in the corpora, clustering features and disambiguating by matching the word features with those of other words in the syntactic structure using the result of the semantic analysis (see Section 3.2).

Parsing. Syntactic parsing is a procedure that requires semantic information (e.g., to understand which phrase in the parse tree a modifier should be associated with). This approach will alleviate ambiguity problems at syntactic level by using the semantics extracted by the distributional approach over the semantic features.

Information Retrieval (IR). By using the proposed approach, computational systems can process complex queries and improve precision and recall of relevant documents. The aim is to go beyond the state of the art in query expansion by combining similar semantic features in accordance with the syntactic structure, rather than using bag-of-words approach, synonyms and paraphrases.

Textual Entailment (TE). Current research in TE attempts to solve the problem of implicit meaning in texts by lexical inference (e.g., *selling* implies *owning*), using resources (e.g., WordNet), distributional semantics and similarity measures. However, these techniques still operate at lexical level. This proposal operates at a semantic rather than lexical level which brings out the implicit meanings sought by other means in TE research.

Generation and Summarization. This proposal will enable the generation of lexical compositions reflecting plausible combinations of semantic features instead of lexical substitutions. This will open a completely new horizon of summarization results.

4 Conclusions

This paper presents a recently-funded project on a research frontier in Computational Linguistics. It includes a brief survey on the topic and the essential keys of the proposal with its impact.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.

Dan Garrette, Katrin Erk, and Raymond Mooney. 2014. A formal approach to linking logical form and vector-space lexical semantics. In *Computing Meaning*, pages 27–48. Springer.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.

Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *AAAI*.

Germán Kruszewski and Marco Baroni. 2014. Dead parrots make bad pets: Exploring modifier effects in noun phrases. *Lexical and Computational Semantics (* SEM 2014)*, page 171.

Mike Lewis and Mark Steedman. 2013. Combining distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Claudio Schifanella, K Selçuk Candan, and Maria Luisa Sapino. 2014. Multiresolution tensor decompositions with mode hierarchies. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2):10.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Peter D Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, pages 533–585.

An Active Learning Approach to the Classification of Non-Sentential Utterances

Paolo Dragone[1], **Pierre Lison**[2]
[1]Sapienza University of Rome
[2] University of Oslo
dragone.paolo@gmail.com, plison@ifi.uio.no

Abstract

English. This paper addresses the problem of classification of non-sentential utterances (NSUs). NSUs are utterances that do not have a complete sentential form but convey a full clausal meaning given the dialogue context. We extend the approach of Fernández et al. (2007), which provide a taxonomy of NSUs and a small annotated corpus extracted from dialogue transcripts. This paper demonstrates how the combination of new linguistic features and active learning techniques can mitigate the scarcity of labelled data. The results show a significant improvement in the classification accuracy over the state-of-the-art.

Italiano. *Questo articolo affronta il problema della classificazione delle non-sentential utterances (NSUs). Le NSUs sono espressioni che, pur avendo una forma incompleta, esprimono un significato completo dato il contesto del dialogo. Estendiamo l'approccio di Fernández et al. (2007), il quale fornisce una tassonomia per NSUs ed un piccolo corpus estratto da transcript di dialoghi. Questo articolo dimostra come, tramite l'utilizzo di nuove feature linguistiche in combinazione con tecniche di active learning, si riesce ad attenuare la sarsità di dati annotati. I risultati mostrano un miglioramento significativo dell'accuratezza rispetto allo stato dell'arte.*

1 Introduction

In dialogue, utterances do not always take the form of complete, well-formed sentences with a subject, a verb and complements. Many utterances – often called *non-sentential utterances*, or NSUs for short – are fragmentary and lack an overt predicate. Consider the following examples from the British National Corpus:

A: How do you actually feel about that?
B: **Not too happy.** [BNC: JK8 168-169]

A: They wouldn't do it, no.
B: **Why?** [BNC: H5H 202-203]

A: [...] then across from there to there.
B: **From side to side.** [BNC: HDH 377-378]

Despite their ubiquity, the semantic content of NSUs is often difficult to extract automatically. Non-sentential utterances are indeed intrinsically dependent on the dialogue context for their interpretation – for instance, the meaning of "why" in the example above is impossible to decipher without knowing what precedes it.

This paper describes a new approach to the classification of NSUs. The approach builds upon the work of Fernández et al. (2007), which present a corpus of NSUs along with a taxonomy and a classifier based on simple features. In particular, we show that the inclusion of new linguistic features and the use of active learning provide a modest but significant improvement in the classification accuracy compared to their approach.

The next section presents the corpus used in this work and its associated taxonomy of NSUs. Section 3 describes our classification approach (extracted features and learning algorithm). Section 4 finally presents the empirical results and their comparison with the baseline.

2 Background

Fernández et al. (2007) provide a taxonomy of NSUs based on 15 classes, reflecting both the form and pragmatic function fulfilled by the utterance.

The aforementioned paper also presents a small corpus of annotated NSUs extracted from dialogue transcripts of the British National Corpus

NSU Class	Example		Frequency
Plain Ack. (Ack)	*A: ...*	*B: mmh*	599
Short Answer (ShortAns)	*A: Who left?*	*B: Bo*	188
Affirmative Answer (AffAns)	*A: Did Bo leave?*	*B: Yes*	105
Repeated Ack. (RepAck)	*A: Did Bo leave?*	*B: Bo, hmm.*	86
Clarification Ellipsis (CE)	*A: Did Bo leave?*	*B: Bo?*	82
Rejection (Reject)	*A: Did Bo leave?*	*B: No.*	49
Factual Modifier (FactMod)	*A: Bo left.*	*B: Great!*	27
Repeated Aff. Ans. (RepAffAns)	*A: Did Bo leave?*	*B: Bo, yes.*	26
Helpful Rejection (HelpReject)	*A: Did Bo leave?*	*B: No, Max.*	24
Check Question (CheckQu)	*A: Bo isn't here. okay?*		22
Sluice	*A: Someone left.*	*B: Who?*	21
Filler	*A: Did Bo ...*	*B: leave?*	18
Bare Modifier Phrase (BareModPh)	*A: Max left.*	*B: Yesterday.*	15
Propositional Modifier (PropMod)	*A: Did Bo leave?*	*B: Maybe.*	11
Conjunct (Conj)	*A: Bo left.*	*B: And Max.*	10
Total			1283

Table 1: Taxonomy of NSUs with examples and frequencies in the corpus of Fernández et al. (2007).

(Burnard, 2000). Each instance of NSU is annotated with its corresponding class and its antecedent (which is often but not always the preceding utterance). Table 1 provides an overview of the taxonomy, along the frequency of each class in the corpus and prototypical examples taken from Ginzburg (2012). See also e.g. Schlangen (2003) for related NSU taxonomies. Due to space constraints, we do not provide here an exhaustive description of each class, which can be found in (Fernández, 2006; Fernández et al., 2007).

3 Approach

In addition to their corpus and taxonomy of NSUs, Fernández et al. (2007) also described a simple machine learning approach to determine the NSU class from simple features. Their approach will constitute the baseline for our experiments. We then show how to extend their feature set and rely on active learning to improve the classification.

3.1 Baseline

The feature set of Fernández et al. (2007) is composed of 9 features. Four features capture some key syntactic and lexical properties of the NSU itself, such as the presence of yes/no words or wh-words in the NSU. In addition, three features are extracted from the antecedent utterance, capturing properties such as the mood or the presence of a marker indicating whether the utterance is complete. Finally, two features encode similarity measures between the NSU and its antecedent, such as the number of repeated words and POS tag sequences common to the NSU and its antecedent.

The classification performance of our replicated classifier (see Table 2) are in line with the results presented in Fernández et al. (2007) – with the exception of the accuracy scores, which were not provided in the original article.

3.2 Extending the feature set

In order to improve the classification accuracy, we extended the baseline features described above with a set of 23 additional features, summing up to a total of 32 features:

- **POS-level features**: 7 features capturing shallow syntactic properties of the NSUs, such as the initial POS tags and the presence of pauses and unclear fragments.
- **Phrase-level features**: 7 features indicating the presence of specific syntactic structures in the NSU and the antecedent, for instance the type of clause-level tags (eg. S, SQ, SBAR) in the antecedent or the initial phrase-level tag (eg. ADVP, FRAG, NP) in the NSU.
- **Dependency features**: 2 features signaling the presence of certain dependency patterns in the antecedent, for example the occurrence of a *neg* dependency in the antecedent.
- **Turn-taking features**: one feature indicating

whether the NSU and its antecedent are uttered by the same speaker.

- **Similarity features**: 6 features measuring the parallelism between the NSU and its antecedent, such as the local (character-level) alignment based on Smith and Waterman (1981) and the longest common subsequence at the word- and POS-levels, using Needleman and Wunsch (1970).

The phrase-level and dependency features were extracted with the PCFG and Dependency Parsers (Klein and Manning, 2003; Chen and Manning, 2014) from the Stanford CoreNLP API.

3.3 Active learning

The objective of active learning (AL) (Settles, 2010) is to interactively query the user to annotate new data by selecting the most informative instances (that is, the ones that are most difficult to classify). Active learning is typically employed to cope with the scarcity of labelled data. In our case, the lack of sufficient training data is especially problematic due to the strong class imbalance between the NSU classes (as exemplified in Table 1). Furthermore, the most infrequent classes are often the most difficult ones to discriminate. Fortunately, the dialogue transcripts from the BNC also contain a large amount of unlabelled NSUs that can be extracted from the raw transcripts using simple heuristics (syntactic patterns to select utterances that are most likely non-sentential).

The active learning algorithm we employed in this work is a pool-based method with uncertainty sampling (Lewis and Catlett, 1994). The sampling relies on *entropy* (Shannon, 1948) as measure of uncertainty. Given a particular (unlabelled) instance with a vector of feature values $\mathbf{f}$, we use the existing classifier to derive the probability distribution $P(C = c_i|\mathbf{f})$ for each possible output class c_i. We can then determine the corresponding entropy of the class C:

$$H(C) = -\sum_i P(C=c_i|\mathbf{f}) \log P(C=c_i|\mathbf{f})$$

A high entropy indicates the "unpredictability" of the instance. The most informative instances to label are therefore the ones with high entropy. As argued in Settles (2010), entropy sampling is especially useful when there are more than two classes, as in our setting. We applied the JCLAL active learning library[1] to extract and annotate 100 new instances of NSUs, which were then added to the training data. The distribution of NSU classes for these instances is shown in Table 5.

NSU Class	Instances
Helpful Rejection	21
Repeated Acknowledgment	17
Clarification Ellipsis	17
Acknowledgment	11
Propositional Modifier	9
Filler	9
Sluice	3
Repeated Affirmative Answer	3
Factual Modifier	3
Conjunct Fragment	3
Short Answer	2
Check Question	2

Table 5: Class frequencies of the 100 additional NSUs extracted via active learning.

4 Evaluation

We compared the classification results between the baseline and the new approach which includes the extended feature set and the additional data extracted via active learning. All the experiments were conducted using the Weka package (Hall et al., 2009). Table 2 presents the results using the J48 classifier, an implementation of the C4.5 algorithm for decision trees (Quinlan, 1993), while Table 3 presents the results using Weka's SMO classifier, a type of SVM trained using sequential minimal optimization (Platt, 1998). In all experiments, we follow Fernández et al. (2007) and remove from the classification task the NSUs whose antecedents are not the preceding utterance, thus leaving a total of 1123 utterances.

All empirical results were computed with 10-fold cross validation over the full dataset. The active learning (AL) results refer to the classifiers trained after the inclusion of the 100 additional instances. The results show a significant improvement of the classification performance between the baseline and the final approach using the SVM and the data extracted via active learning. Using a paired t-test with a 95% confidence interval between the baseline and the final results, the improvement in accuracy is statistically significant with a p-value of 6.9×10^{-3}. The SVM does

[1] cf. http://sourceforge.net/projects/jclal.

Experimental setting	Accuracy	Precision	Recall	F_1-Score
Train-set (baseline feature set)	0.885	0.888	0.885	0.879
Train-set (extended feature set)	0.889	0.904	0.889	0.889
Train-set + AL (baseline feature set)	0.890	0.896	0.890	0.885
Train-set + AL (extended feature set)	0.896	0.914	0.896	0.897

Table 2: Accuracy, precision, recall and F_1 scores for each experiment, based on the J48 classifier.

Experimental setting	Accuracy	Precision	Recall	F_1-Score
Train-set (baseline feature set)	0.881	0.884	0.881	0.875
Train-set (extended feature set)	0.899	0.904	0.899	0.896
Train-set + AL (baseline feature set)	0.883	0.893	0.883	0.880
Train-set + AL (extended feature set)	**0.907**	**0.913**	**0.907**	**0.905**

Table 3: Accuracy, precision, recall and F_1 scores for each experiment, based on the SMO classifier.

	Baseline			Final approach		
NSU Class	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Plain Acknowledgment	0.97	0.97	0.97	0.97	0.98	0.97
Affirmative Answer	0.89	0.84	0.86	0.81	0.90	0.85
Bare Modifier Phrase	0.63	0.65	0.62	0.77	0.75	0.75
Clarification Ellipsis	0.87	0.89	0.87	0.88	0.92	0.89
Check Question	0.85	0.90	0.87	1.00	1.00	1.00
Conjunct Fragment	0.80	0.80	0.80	1.00	1.00	1.00
Factual Modifier	1.00	1.00	1.00	1.00	1.00	1.00
Filler	0.77	0.70	0.71	0.82	0.83	0.78
Helpful Rejection	0.13	0.14	0.14	0.31	0.43	0.33
Propositional Modifier	0.92	0.97	0.93	0.92	1.00	0.95
Rejection	0.76	0.95	0.83	0.90	0.90	0.89
Repeated Ack.	0.74	0.75	0.70	0.77	0.77	0.77
Repeated Aff. Ans.	0.67	0.71	0.68	0.72	0.55	0.58
Short Answer	0.86	0.80	0.81	0.92	0.86	0.89
Sluice	0.67	0.77	0.71	0.80	0.84	0.81

Table 4: Precision, recall and F_1 score per class between the baseline (initial feature set and J48 classifier) and the final approach (extended feature set with active learning and SMO classifier).

not perform particularly well on the baseline features but scales better than the J48 classifier after the inclusion of the additional features. Overall, the results demonstrate that the classification can be improved using a modest amount of additional training data combined with an extended feature set. However, we can observe from Table 4 that some NSU classes remain difficult to classify. Distinguishing between e.g. *Helpful Rejections* and *Short Answers* indeed requires a deeper semantic analysis of the NSUs and their antecedents than cannot be captured by morpho-syntactic features alone. Designing appropriate semantic features for this classification task constitutes an interesting question for future work.

5 Conclusion

This paper presented the results of an experiment in the classification of non-sentential utterances, extending the work of Fernández et al. (2007). The approach relied on an extended feature set and active learning techniques to address the scarcity of labelled data and the class imbalance. The evaluation results demonstrated a significant improvement in the classification accuracy.

The presented results also highlight the need for a larger annotated corpus of NSUs. In our view, the development of such a corpus, including new dialogue domains and a broader range of conversational phenomena, could contribute to a better understanding of NSUs and their interpretation.

Furthermore, the classification of NSUs according to their type only constitutes the first step in their semantic interpretation. Dragone and Lison (2015) focuses on integrating the NSU classification outputs for natural language understanding of conversational data, building upon Ginzburg (2012)'s formal theory of conversation.

References

L. Burnard. 2000. Reference guide for the british national corpus (world edition).

D. Chen and C. D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1, pages 740–750.

P. Dragone and P. Lison. 2015. Non-sentential utterances in dialogue: Experiments in classification and interpretation. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, page 170.

R. Fernández, J. Ginzburg, and S. Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

R. Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College London.

J. Ginzburg. 2012. *The Interactive Stance*. Oxford University Press.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

D. D. Lewis and J. Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156.

S. B. Needleman and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

J. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research, April.

R. J. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

D. Schlangen. 2003. *A coherence-based approach to the interpretation of non-sentential utterances in dialogue*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

B. Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal, The*, 27(3):379–423, July.

T. F. Smith and M. S. Waterman. 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

The CompWHoB Corpus: Computational Construction, Annotation and Linguistic Analysis of the White House Press Briefings Corpus

Fabrizio Esposito[†], **Pierpaolo Basile**[‡], **Francesco Cutugno**[⋆], **Marco Venuti**[*]
[†] Dept. of Humanities, University of Naples Federico II
[‡] Dept. of Computer Science, University of Bari Aldo Moro,
[⋆] DIETI, University of Naples Federico II,
[*] Dept. of Humanities, University of Catania
fabrizio.esposito3@unina.it, pierpaolo.basile@uniba.it,
francesco.cutugno@unina.it, mvenuti@unict.it

Abstract

English. The CompWHoB (Computational White House press Briefings) Corpus, currently being developed at the University of Naples Federico II, is a corpus of spoken American English focusing on political and media communication. It represents a large collection of the White House Press Briefings, namely, the daily meetings held by the White House Press Secretary and the news media. At the time of writing, the corpus amounts to more than 20 million words, covers a period of time of twenty-one years spanning from 1993 to 2014 and it is planned to be extended to the end of the second term of President Barack Obama. The aim of the present article is to describe the composition of the corpus and the techniques used to extract, process and annotate it. Moreover, attention is paid to the use of the Temporal Random Indexing *(TRI)* on the corpus as a tool for linguistic analysis.

Italiano. *Il CompWHoB Corpus, in sviluppo presso l'Università di Napoli Federico II, è un corpus di parlato inglese-americano comprendente le conferenze condotte dai segretari statunitensi per i rapporti con la stampa, definite come Press Briefings. Allo stato attuale il corpus è composto da più di 20 milioni di parole e si estende dal 1993 sino a fine 2014. L'obiettivo di questo articolo è di descrivere la composizione del corpus, le tecniche utilizzate per estrarre ed annotare i testi, e mostrare come possa fungere da fonte di analisi linguistica attraverso l'utilizzo del Temporal Random Indexing (TRI).*

1 Introduction

As political speech has been gaining more and more attention over recent years in the analysis of communication strategies, political corpora have become of paramount importance for the fulfilment of this objective. The CompWHoB Corpus, a spoken American English corpus currently being developed at the University of Naples Federico II, wants to meet the need for political language data, as it focuses on the political and media communication genre. This resource is a large collection of the transcripts of the White House Press Briefings, namely, the daily meetings held by the White House Press Secretary and the news media. As one of the main official channels of communication for the White House, briefings play indeed a crucial role in the administration communication strategies (Kumar, 2007). The corpus currently amounts to more than 20 million words and spans from 1993 to 2014, thus covering a period of time of twenty-one years and five presidencies. Work is underway to extend the corpus so as to reach the end of the second term of President Barack Obama. Unlike other political corpora such as CORPS (Guerini et al., 2008; Guerini et al., 2013) and the Political Speech Corpus of Bulgarian (Osenova and Simov, 2012), the CompWHoB does not include monological situations, due to the inherent dialogical characteristics of the briefings. As other web corpora (Baroni and Kilgarriff, 2006; Baroni et al., 2009; Lyding et al., 2014), the CompWHoB can be considered a web corpus (Kilgarriff and Grefenstette, 2003; Hundt et al., 2007), since its texts are directly extracted from *The American Presidency Project* website. Moreover, it should be pointed out that WHoB is a pre-existing specialized corpus (Spinzi and Venuti, 2013) annotated by using XML mark-up and mainly employed in the field of corpus lin-

CompWHoB Corpus							
Presidency	**texts**	**tokens**	**tokens mean**	**types**	**TTR**	**turn-takings**	**WHo-s**
Bill Clinton_1	1,072	4,581,665	4,274	79,129	36.97	116,437	497
Bill Clinton_2	1,066	4,658,054	4,370	81,789	37.89	102,160	525
George W. Bush_1	777	3,660,600	4,711	65,635	34.30	78,992	133
George W. Bush_2	1,057	4,536,616	4,292	73,809	34.65	82,702	286
Barack Obama_1	804	4,470,070	5,560	76,604	36.23	87,432	299
Barack Obama_2	463	3,344,567	7,224	48,493	26.51	44,982	74
TOTAL	**5,239**	**25,251,572**		**426,458**		**512,651**	**1,814**

Table 1: Composition of the CompWHoB Corpus in its current stage (July 2015); _1 and _2 stands for the first term and second term of each presidency, respectively; type-token ratio was calculated using Guiraud's (Guiraud, 1954) index of lexical richness; WHo-s stands for White House staff, namely, personnel identified as belonging or related to the White House presidential staff.

guistics. Thus, the aim of the present article is to describe how the corpus can be used as a future resource in different research fields such as computational linguistics, (political) linguistics, political science, etc.

The paper is structured as follows: Section 2 gives an overview of the corpus. Section 3 describes the details of the corpus construction and annotation. The use of TRI on the corpus is then discussed in Section 4. Lastly, Section 5 concludes the paper.

2 Corpus Overview

The CompWHoB Corpus consists of the transcripts of the press conferences held by the White House Press Secretaries and/or other administration officials and the news media. The texts that form the corpus were all extracted from the American Presidency Project website `www.presidency.ucsb.edu`, where the Press Briefings document archive section can be freely consulted. Data was collected and formatted into a standardized XML encoding, according to the TEI Guidelines (Sperberg-McQueen and Burnard, 2007). In some cases, texts were subsequently split to mark the beginning of the new president first term. Six are the presidencies represented in the CompWHoB Corpus: both Bill Clinton and George W. Bush eight-year term are included, while the second term of the incumbent US President, Barack Obama, is not complete since he is currently in office. Thus, at the current stage (July 2015) the corpus contains a total of 5,239 texts comprising 25,251,572 tokens and 422,891 types, and spans from January 27, 1993 until December 18, 2014. Given the inherent dialogical characteristics of press conferences, a total number of 512,651 turn-takings has been calculated so far. Across the time span covered by the corpus, 1,814 are the speakers individually identified as press secretaries, presidential staff members or administration officials. See Table 1 for more details.

3 Corpus Construction and Annotation

3.1 Construction and Structural Annotation

Data extracted comes in a standardized format. Each briefing consists of a transcript where every turn-taking is signalled by the use of the capital letters to identify the speaker. Two are the main roles found in the transcriptions: the podium, namely, the White House Press Secretary or any other administration official, always identified by their surnames; the press corps, identified by the use of the capital letter Q. Information about the date of the event was extracted and then added to the beginning of every press conference. As first step after data extraction, the resulting texts were encoded in XML format in a semi-automatic way by using regular expressions and manual checking. Transcripts were then mapped to XML files according to a calendar year division. Metatextual information contained in the data was encoded as well so as to enrich the corpus and make it easily navigable. Thus, the CompWHoB Corpus is structured as follows: every year forming part of the corpus is diachronically structured. A *div* tag was created to mark the beginning and the end of every transcript. An attribute value shows the date of that specific event in a `yyyy-mm-dd` format. Every *div* contains the dialogical situation of the press conference, where each speaker is identified by the use of a *u* tag. In order to provide

an in-depth description of the sociolinguistic characteristics of the speakers, every *u* tag consists of self-explanatory multiple attributes: *role*, *sex* and *who*. Since in the transcripts press corps are only identified by the capital letter Q, it was impossible neither to recover information about the gender nor the name. Thus, for every media member the attribute value *sex* is always *u*, namely, unknown, and both *role* and *who* attribute values are always *journalist*. Conversely, since information about Press Secretaries and members related to the presidential administration staff was available in the transcripts, attribute values contain information about the role, gender and name of the speaker. This operation had to be made manually, but one of the main objectives of this work is to make it semi-automatic querying an existent political database that will make the process less burdensome. As many are the White House members involved in the press conferences, we decided to categorize them by role. Thus, Press Secretaries are the only ones identified as *podium*, due to their function of conducting the briefing. Administration officials and presidential staff members can be instead recognized by the role value *podium_* plus the position held by them (e.g. military, administration, etc.). The beginning and the end of every speech is marked by the use of *p* tags. As original transcripts contained also meta-textual information enclosed in brackets about audience reactions and speech events descriptions (e.g. (Laughter), (Applause), etc.), we decided to keep it so as to broaden and vary future analysis approaches. See Table 2 for a summary of these tags. See Table 3 for the description of the corpus press conference structure.

Tag
{event type="laughter"}
{event desc="applause"}
{event desc="inaudible"}

Table 2: Meta-textual speech events tags

{div1} # date of the press conference
{u} # identification of the speaker
{p} # speech of the identified speaker
{self-closing tag} # extra-textual speech events

Table 3: CompWHoB briefing structure

3.2 Linguistic Annotation

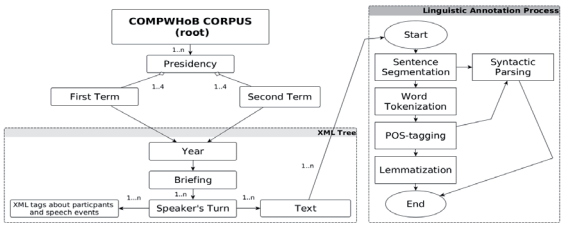


Figure 1: CompWHoB structure and linguistic annotation process

As regards the NLP aspect (Figure 1), we chose to adopt Python (3.4 version) as programming language, using the Natural Language Toolkit (NLTK) platform (Bird et al., 2009), since it provides a large suite of libraries for natural language processing. As first step, sentence segmentation and word tokenization were carried out. POS-tagging was then performed employing the Penn Treebank tag set (Marcus et al., 1993), trained on the Treebank Corpus. We made this choice to have immediately a first grasp on the linguistic data. Being at the early stages of our work, we decided to test NLTK POS tagger by comparing the output with a human-labeled Gold Standard test set consisting of 24 sections randomly selected from the corpus, amounting to over 500 tokens. Since at the current stage POS tagging achieves an accuracy of 92%, our future aim is to improve the performance of NLTK POS tagger once the corpus is complete, providing it with a syntactic parsing as well. As for the lemmatization of the resulting texts, we decided to use the WordNet lemmatizer provided by the NLTK platform. During this task we had to map the part-of-speech tags to the WordNet part-of-speech names in order to get a more accurate output. Texts processing tasks were always performed taking into account each turn-taking. This means that, at the current status, one of the main advantages of the CompWHoB Corpus is the possibility to retrieve linguistic information by specifying the name and/or the role of the speaker, allowing an in-depth analysis of the acquired information. This is why our primary objective in the near future is to provide the means to query the corpus. We plan to reach this goal by employing the Corpus Workbench (CWB) architecture and the *Corpus Query Processor* (Christ et al., 1999; Evert and Hardie, 2011).

4 TRI on the CompWHoB Corpus

Our intention was to perform a linguistic analysis with the aim of finding some variation in word usage across several presidential and political mandates. We chose to model word usage exploiting distributional semantic models (Sahlgren, 2006). In a distributional semantic model, words are represented as mathematical points in a geometric space. Similar words are represented close in that space. The space is built taking into account words co-occurrences in a large corpus. One drawback of this kind of approach is that geometric spaces built on different corpora are not comparable. Moreover the temporal feature is not included in these models. Considering the peculiarities of the CompWHoB Corpus such as temporal information and different speakers, a technique able to manage these kind of features is needed. Recently, a technique called TRI based on Random Indexing (Sahlgren, 2005) able to manage temporal information has been proposed in (Basile et al., 2014). TRI can build different word spaces for several time periods allowing the analysis of how words change their meaning over time. Relying on TRI, we build six separate word spaces, one space for each presidency. The first goal of our analysis is to find interesting words that change their meaning across time. Since word vectors in each word space are made comparable thanks to the TRI tool, it is possible to compare the similarity of a word vector in each word space. In particular, given a word w and two time periods t_1 and t_2 is possible to compare the cosine similarity between the word vector of w in t_1 and word vector of w in t_2. A low level of similarity between vectors indicates a high word usage variation across the two time periods. Exploiting this technique we discovered some words that significantly change their usage. In this case, it is worth paying attention to the words resulting from the time periods representing the end of a presidency second term and the beginning of a new one. For example, investigating the neighbourhood of the word *Guatemala* in `Clinton_2/Bush_1`, we note that in `Clinton_2` words such as *donors*, *accord* and *workable* appear, while in `Bush_1` the word *Guatemala* is near to other geo-political entities, for example: *honduras* and *slovak*. Investigating historical events in that period we found that in 1999 President Clinton finally apologized for America's role in almost a half-century of repression in Guatemala.

The second analysis concerns how a particular topic is treated. We selected the topic of the American debate on guns. The idea was to analyse how each presidency discusses this subject. We selected the word *gun* as the representative word of the topic. Moreover, we expanded the topic employing semantic frames in which the word *gun* had been previously used. We adopted FrameNet to extract relevant frames. Following this methodology we identified other relevant words: *firearm, handgun, machine-gun, shooter, shotgun* as nouns; and *discharge, fire, hit, shoot* as verbs.

In order to represent the gun topic in the word space we adopted the vector sum operator. For each word space a vector was built, representing the vector sum of words belonging to the topic. The sum vector is used to retrieve the most similar vectors using cosine similarity. This operation was repeated for each administration. The idea was to analyse the neighbourhood of the gun topic in each presidency. Results show a clear evolution in how the different administrations dealt with this subject. While in Bill Clinton and George W. Bush presidencies the first fifteen most similar vectors mainly denote the semantic field of weapons, it is only from the Obama administration that adjectives and nouns appealing to emotions make their appearance (e.g. *heartening, suffer, grassroots, darn*), marking a new era in the White House communication strategies about the gun issue.

5 Conclusions

At the time of writing, the CompWHoB Corpus is probably one of the largest political corpora mainly based on spontaneous spoken language. This feature represents one of its strongest points, as the linguistic analysis performed by employing the TRI has proved. As for the near future, two are our main goals: the first one is to make the process of structural annotation as much computational as possible by retrieving information from available political databases; the second one is to provide the corpus with syntactic parsing and improve the overall performance of the linguistic annotation process. In terms of accessibility, we intend to make the CompWHoB Corpus available via the CPQ web interface (Hardie, 2012) by the end of next year. For now, the fully annotated corpus is accessible and available on request.

References

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 87–90. East Stroudsbourg.

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*, 43:209–226, September.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing Word Meaning over Time by Exploiting Temporal Random Indexing. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *First Italian Conference on Computational Linguistics CLiC-it 2014*. Pisa University Press.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Oliver Christ, Bruno M. Schulze, and Esther Knig, 1999. *Corpus Query Processor (CQP). User's Manual*. Institut fr Maschinelle Sprachverarbeitung, Universitt Stuttgart, Stuttgart, Germany.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, Birmingham, UK. University of Birmingham.

Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2008. Corps: A corpus of tagged political speeches for persuasive communication processing. 5(1):19–32.

Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava, 2013. *The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions*, volume 7688 of *Lecture Notes in Computer Science*, pages 86–98. Springer Berlin Heidelberg.

Paul Guiraud. 1954. *Les Charactres Statistiques du Vocabulaire. Essai de mthodologie*. Presses Universitaires de France, Paris.

Andrew Hardie. 2012. Cqpweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17:380–409.

Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer. 2007. *Corpus linguistics and the web*. Rodopi, Amsterdam and New York.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.

Martha J. Kumar. 2007. *Managing the Presidents Message: the White House Communications Operation*. The John Hopkins University Press.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden, April. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Petya Osenova and Kiril Simov. 2012. The political speech corpus of bulgarian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Magnus Sahlgren. 2005. An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm: Stockholm University, Faculty of Humanities, Department of Linguistics.

C. Michael Sperberg-McQueen and Lou Burnard. 2007. TEI P5:Guidelines for Electronic Text Encoding and Interchange.

Cinzia Spinzi and Marco Venuti, 2013. *Tracking the change in an Institutional Genre: A Diachronic Corpus-based study of White House Press Briefings*, pages 182–197. Cambridge Scholars Publishing, Newcastle upon Tyne.

Costituzione di un corpus giuridico parallelo italiano-arabo

Fathi Fawi
Dipartimento di Studi linguistici e culturali comparati
Università Ca' Foscari – Venezia
fathi_fawi@yahoo.com

Abstract

English. Parallel corpora are an important resource for many applications of computational linguistics, such as machine translation, terminology extraction, semantic disambiguation, etc. In this paper we present our attempt to build an Italian-Arabic parallel corpus in the legal domain, aligned at the sentence level and tagged at the POS level.

Italiano. *I corpora paralleli rappresentano un'importanza assoluta per tante applicazioni della linguistica computazionale, come la traduzione automatica, l'estrazione delle terminologie, o la disambiguazione semantica, ecc. In questo lavoro presentiamo il nostro tentativo di creare un corpus giuridico parallelo italiano-arabo allineato a livello di frase e annotato a livello morfosintattico.*

1 Introduzione

Con il crescente sviluppo delle tecnologie informatiche che consentono di raccogliere, gestire ed esplorare enormi quantità di dati linguistici, l'interesse alla creazione di corpora linguistici è cresciuto recentemente in una maniera esponenziale. È indubbio che oggigiorno l'enorme disponibilità dei dati sul web ha agevolato significativamente la costituzione e la distribuzione dei corpora linguistici sia i corpora monolingui che quelli multilingui. In effetti, i corpora costituiscono una risorsa essenziale per il campo linguistico soprattutto per le analisi contrastive tra due o più lingue, per la didattica delle lingue straniere e per gli studi lessicografici e di traduzione. Nell'ambito della linguistica computazionale i corpora linguistici, e in particolare quelli paralleli, acquistano un'importanza assoluta, soprattutto per applicazioni come la traduzione automatica, l'estrazione di terminologie o la disambiguazione semantica.

Tuttavia, non tutte le lingue prendono ugualmente parte a corpora paralleli bilingui o multilingui. In effetti, l'arabo è una lingua che presenta una limitata partecipazione a corpora paralleli, soprattutto a quelli specialistici. È un fenomeno che si può considerare come un possibile effetto della modesta disponibilità sul web di testi paralleli in lingua araba e in altre lingue, nonché della complessità del sistema morfologico arabo.

In questo contributo cerchiamo di esporre la nostra esperienza con la creazione di un corpus giuridico parallelo italiano-arabo specializzato nel diritto internazionale. È un corpus allineato a livello di frase e annotato morfosintatticamente. Una versione del corpus bilingue allineato a livello di frase sarà disponibile gratuitamente per la comunità scientifica al sito del Laboratorio di Linguistica Computazionale dell'Università di Ca' Foscari, Venezia[1].

2 Stato dell'arte

A nostra conoscenza, fino al tempo di questo lavoro non esiste un corpus parallelo italiano-arabo nel dominio giuridico. Nell'ambito del progetto *L'arabo per la 488* (Picchi et al., 1999) è stato creato un corpus parallelo italiano-arabo di testi generici: si tratta di progetto finalizzato allo sviluppo di strumenti e risorse tanto per la lingua italiana quanto per la lingua araba, con particolare cura per l'aspetto contrastivo. Se invece guardiamo allo stato dell'arte delle nostre due lingue come partecipi insieme ad altre lingue di corpora paralleli, troviamo che l'italiano prende parte a risorse testuali multilingue in misura maggiore rispetto all'arabo.

Dei corpora paralleli in italiano e altre lingue ricordiamo *Bononia Legal Corpus* (Rossini Favretti et al., 2007), che è un corpus inglese-

1 http://project.cgm.unive.it

italiano di testi giuridici paralleli e comparabili, sviluppato presso l'università di Bologna. Il progetto è costituito in due fasi: nella prima fase si è costruito un corpus pilota, costituito da corpora paralleli in inglese e in italiano; mentre nella fase successiva vengono aggiunti corpora comparabili nelle due lingue riguardanti testi nell'ambito legislativo, giudiziario e amministrativo per analizzare le caratteristiche linguistiche dei due sistemi legali. Inoltre, nell'ambito del progetto CATEX (*Computer Assisted Terminology Extraction*) presso l'Accademia Europea di Bolzano è stato realizzato un corpus giuridico parallelo italiano-tedesco (Gamper, 1998). Questo corpus comprende una raccolta di leggi italiane con la relativa traduzione in tedesco con una dimensione di quasi 5 milioni di tokens, ed è allineato a livello di frase.

Per quanto concerne, invece, i corpora paralleli in arabo e altre lingue si rammenta EAPCOUNT (Hammouda, 2010), che è un corpus parallelo inglese-arabo con 341 testi delle Nazioni Unite allineati a livello di paragrafo. Inoltre, si menziona il corpus creato presso il laboratorio di linguistica computazionale dell'università autonoma di Madrid (Samy et al., 2006). Si tratta di un corpus parallelo multilingue (inglese-spagnolo- arabo) che contiene una collezione dei documenti delle Nazioni Unite, allineati a livello di frase e annotati morfosintatticamente.

3 Progettazione del corpus

Come dominio tematico del corpus abbiamo scelto il diritto internazionale e in particolare i diritti umani nel mondo. La scelta di questo genere testuale ha le seguenti motivazioni:

- Il linguaggio giuridico è uno dei linguaggi settoriali che presentano molte peculiarità sui diversi livelli di analisi linguistica, il che rende indifferibilmente necessario fornire e sviluppare corpora di testi giuridici;
- Per quanto riguarda la lingua araba, la maggior parte dei corpora giuridici disponibili sul web riguarda il codice di famiglia dei paesi arabi, che, ispirato ai principi della Shariah Islamica, contiene tante terminologie islamiche che non hanno corrispondenti in italiano. Per il problema dell'intraducibilità dei termini giuridici islamici tra l'arabo e l'italiano, abbiamo pensato quindi al diritto internazionale, dove risulta limitata l'influenza della dimensione religiosa dei termini;
- L'accuratezza della traduzione dei testi paralleli è un fattore essenziale soprattutto trattandosi di terminologie giuridiche, e nei documenti dell'Organizzazione delle Nazioni Unite (ONU) abbiamo trovato un livello di traduzione tanto accurato, visto il carattere ufficiale dei documenti.

4 Descrizione del corpus

I documenti del corpus sono dell'ONU. Si tratta di una grande raccolta di accordi, convenzioni, protocolli internazionali sempre nell'ambito del diritto internazionale in generale e dei diritti umani in particolare. La lingua originale dei documenti del corpus parallelo è l'inglese e sia i testi italiani che i testi arabi sono una traduzione dall'inglese. I testi del corpus si dividono in due categorie: la prima comprende un insieme di convenzioni e accordi internazionali nell'ambito dei diritti umani nel mondo, mentre la seconda contiene le convenzioni dell'Organizzazione Internazionale del Lavoro (ILO). In totale il corpus comprende all'incirca 1,1 milione di parole. Tabella 1 indica i dettagli del corpus.

language	n.parole	n.frasi	lunghezza media delle frasi	type/token ratio
Italiano	545682	18675	30	0.028
Arabo	615947	18391	39	0.068

Tabella 1. Dati statistici del corpus

5 5 Costituzione e preparazione del corpus

Per i testi del corpus il web rappresenta la fonte principale sia per i testi arabi che per quelli italiani. Il risultato di questa fase è un insieme di documenti in formato PDF in entrambe le lingue. Il formato PDF non consente, tuttavia, un trattamento automatico dei testi, quindi bisogna convertire i testi nel formato "Plain text format" che è adeguato a qualsiasi trattamento computazionale del corpus, e poi salvare i testi in UNICODE che è adeguato nel nostro caso dato che i sistemi di scrittura delle due lingue di interesse sono diversi.

Il processo della conversione non è, tuttavia, banale come sembra, soprattutto per la lingua araba. Fra le notevoli osservazioni individuate durante la conversione dei testi arabi ricordiamo: la perdita di alcuni caratteri, lo scambio tra certi

caratteri (soprattutto tra "ا" e "ل"), l'inversione della direzione di scrittura (soprattutto i numeri), la perdita del formato del testo originale, ecc. Tutto questo richiede un grande sforzo per rimuovere ogni forma di "rumore" e restituire la normalità dei testi. Nel caso dei testi italiani gli errori derivati dalla conversione riguardano maggiormente il cambiamento del formato del testo originale.

6 Trattamento del corpus

Fino al passo precedente, lo stato del corpus è grezzo, cioè senza nessuna annotazione linguistica utile per esplorare ed interrogare il corpus in modo migliore. L'importanza dei corpora annotati consiste non solo nella possibilità di esplorare ed estrarre informazioni dal testo, ma anche nel fornire "training e valutazione di algoritmi specifici in sistemi automatici." (Zotti, 2013).
Il trattamento automatico del nostro corpus comprende le seguenti fasi:

6.1 Segmentazione

La segmentazione dei testi è stata effettuata nelle due lingue a livello di frase. Per segmentare i testi abbiamo utilizzato un algoritmo nel pacchetto NLTK basato sulla punteggiatura (".", "?", "!"). Tuttavia, non mancano gli errori anche in questa fase; soprattutto per la mancanza dell'uso delle lettere maiuscole in arabo.
Vista la natura giuridica dei testi, si sono registrate alcune peculiarità riguardanti i confini di frase nei testi del corpus. In questo caso il segno della fine frase non è solo il punto finale come è il caso dei testi generali, ma i segni ":", ";" si possono considerare anche confine di frase, soprattutto quando iniziano una lista di clausole o commi. Il risultato di questa fase è un testo segmentato a livello di una sola frase per riga.

6.2 Tokenizzazione

Tokenizzare un testo significa ridurlo nelle sue unità ortografiche minime, dette tokens, che sono unità di base per ogni successivo livello di trattamento automatico. La complessità di questo compito dipende maggiormente dal tipo di lingua umana in trattamento nonché dal suo sistema di scrittura.
Nell'ambito del trattamento automatico della lingua araba riconoscere l'unità ortografica di base delle parole arabe appare un compito particolarmente complicato per effetto della complessità della morfologia araba, basata su un sistema flessionale e pronominale molto ricco (Habash, 2010). Ne consegue che per disambiguare al meglio le unità lessicali di un testo arabo ogni sistema di tokenizzazione necessita di un analizzatore morfologico. Per tokenizzare i testi arabi del corpus abbiamo utilizzato il sistema MADA+TOKAN[2] (Habash et al., 2009) che nel nostro caso ha avuto un'accuratezza all'incirca 98%. Nel caso dei documenti italiani si è utilizzato il tokenizzatore disponibile al sito di ItaliaNLP Lab[3].

6.3 Allineamento

Per il processo di allineamento si intende rendere due testi, o due unità testuali (nel nostro caso due frasi) allineati l'uno di fronte all'altro. Questa fase si configura come un processo essenziale lavorando sui corpora paralleli. L'allineamento viene effettuato normalmente da appositi programmi che si servono di metodi statistici e linguistici per mettere in corrispondenza due unità di testo l'una è traduzione dell'altra. Nel caso dei metodi statistici si utilizzano i calcoli probabilistici della lunghezza delle unità (frasi, parole, caratteri) dei due testi paralleli per stabilire una adeguata equivalenza tra i due testi in esame. Inoltre, il metodo statistico si può arricchire di repertori lessicali derivati da dizionari o corrispondenze traduttive prestabilite. Non c'è dubbio che l'utilizzo del metodo ibrido appare più conveniente soprattutto quando si tratta di lingue che hanno sistemi di scrittura tanto diversi tra loro, come per es. le lingue del nostro corpus.

Per allineare i nostri testi, abbiamo utilizzato *LogiTerm* che fa parte di *Terminotix*[4]. Questo programma segmenta e allinea automaticamente due testi creando il risultato in formati diversi (HTML, XML, TMX). L'accuratezza dell'allineamento nel nostro caso è all'incirca 95%, quindi non mancava un intervento manuale per correggere alcuni errori dovuti in generale alle caratteristiche linguistiche delle due lingue in questione. La maggior parte degli errori individuati durante l'allineamento riguarda la lunghezza della frase araba. Come si può osservare dal numero totale delle frasi nella

2 We used version 3.2 of MADA+TOKAN
3 http://www.italianlp.it/
4 http://www.terminotix.com/index.asp?lang=en

Tabella 1, la lingua araba tende a congiungere le frasi, quindi non è raro di trovare un livello di allineamento 2 a 1. Dopo la verifica manuale dei risultati di questa fase, i testi allineati sono salvati in due formati XML e TMX.

<prop type="ltattr-match">1-1</prop>
<prop type="ltattr-id">17</prop>
<tuv xml:lang="it">
<seg>Ogni persona ha diritto al godimento dei diritti e delle libertà riconosciuti e garantiti nella presente Carta senza alcuna distinzione, in particolare senza distinzione di razza, sesso, etnia, colore, lingua, religione, opinione politica o qualsiasi altra opinione, di origine nazionale o sociale, di fortuna, di nascita o di qualsiasi altra situazione.</seg>
</tuv>
<tuv xml:lang="ar">
<seg> يتمتع كل شخص بالحقوق والحريات المعترف بها والمكفولة فى هذا الميثاق دون تمييز خاصة إذا كان قائما على العنصر أو العرق أو اللون أو الجنس أو اللغة أو الدين أو على الرأى السياسى أو أى رأي آخر، أو المنشأ الوطني أو الاجتماعي أو الثروة أو المولد أو أى وضع آخر.</seg>
</tuv>
</tu>
<tu>

Tabella 2. Estratto del corpus allineato in TMX

<seg match="1-1" id="17">
<src>Ogni persona ha diritto al godimento dei diritti e delle libertà riconosciuti e garantiti nella presente Carta senza alcuna distinzione, in particolare senza distinzione di razza, sesso, etnia, colore, lingua, religione, opinione politica o qualsiasi altra opinione, di origine nazionale o sociale, di fortuna, di nascita o di qualsiasi altra situazione.</src>
<tgt> يتمتع كل شخص بالحقوق والحريات المعترف بها والمكفولة فى هذا الميثاق دون تمييز خاصة إذا كان قائما على العنصر أو العرق أو اللون أو الجنس أو اللغة أو الدين أو على الرأى السياسى أو أى رأي آخر، أو المنشأ الوطني أو الاجتماعي أو الثروة أو المولد أو أى وضع آخر.</tgt>
</seg>

Tabella 3. Estratto del corpus allineato in XML

6.4 Annotazione del corpus

Per l'annotazione o l'etichettatura linguistica di un corpus si intende associare alle porzioni del testo informazioni linguistiche in forma di etichetta (tag o mark-up), sia per rendere esplicito il contenuto del testo sia per ottenerne una conoscenza approfondita. Il tipo di annotazione più conosciuto è quello morfosintattico o il cosiddetto POS (part-of-speech tagging), che consiste nell'attribuire ad ogni parola nel testo la sua categoria grammaticale. Il POS tagging possiede un'importanza rilevante nel trattamento automatico del linguaggio, in quanto rappresenta il primo passo nell'annotazione automatica dei testi, quindi gli errori riscontrabili durante questa fase potrebbero incidere sulle successive analisi.

Per taggare i testi arabi del nostro corpus, abbiamo utilizzato il pacchetto Amira 2.1 (Diab, 2009). Amira è un sistema di POS tagging basato sull'apprendimento supervisionato che utilizza le macchine a vettori di supporto (SVM). Questo sistema comprende tre moduli per il trattamento automatico della lingua araba: tokenizzazione, POS tagging, e base-phrase chunked. Nel nostro caso il sistema PoS Tagging di Amira raggiunge un'accuratezza all'incirca 94%.

Per i testi italiani si è usato Felice-POS-Tagger (Dell'Orletta, 2009). Felice-POS-Tagger è una combinazione di sei tagger, con tre algoritmi diversi. Ognuno dei tre algoritmi viene utilizzato per costruire un *left-to-right* (LR) tagger e un *right-to-left* (RL) tagger. L'accuratezza del Felice-POS-Tagger nel taggare i testi del nostro corpus è all'incirca 97%.

Le/RD organizzazioni/S dei/EA lavoratori/S e/CC dei/EA datori/S di/E lavoro/S hanno/V il/RD diritto/S di/E elaborare/V i/RD propri/AP statuti/S e/CC regolamenti/S amministrativi/A ,/FF di/E eleggere/V liberamente/B i/RD propri/AP rappresentanti/S ,/FF di/E organizzare/V la/RD propria/AP gestione/S e/CC la/RD propria/AP attività/S ,/FF e/CC di/E formulare/V il/RD proprio/AP programma/S di/E azione/S ./FS

Tabella 3. Estratto del corpus italiano annotato a livello PoS Tagging

/منظمات IN/ل CC/و DET_NN/العمال NNS_FP/منظمات RP/ل
NNS_FP/أصحاب NN/العمل DET_NN/الحق DET_NN
/لوائحها CC/و NN_PRP_FS3 /دساتيرها NN/وضع IN/في
/IN/في CC/و PUNC/، DET_JJ_FS /الإدارية NN_PRP_FS3
/حرية IN/ب NNS_MP_PRP_FS3 /مهمتها NN/نتخابا
NN_FS /كاملة JJ_FS ،/PUNC و/CC في/IN تنظيم/NN
/NN_PRP_FS3 نشاطها CC/و إدارتها/NN_FS_PRP_FS3
/NN برامج IN/إعداد CC/و في PUNC/،
/NN_PRP_FS3 عملها ./PUNC

Tabella 4. Estratto del corpus arabo annotato a livello PoS Tagging

7 Conclusione

In questo lavoro abbiamo cercato di dare una descrizione del nostro progetto di creare un corpus parallelo italiano-arabo nel campo del diritto internazionale. La costruzione di tale corpus risponde allo scopo generale di fornire risorse linguistiche utili alle applicazioni della linguistica computazionale, soprattutto considerando la mancanza visibile dei corpora paralleli italiano-arabo di testi specialistici. Il

trattamento computazionale del corpus è arrivato fino al PoS tagging, estendibile nel futuro ad altri livelli di annotazione e di arricchimento. Nel futuro intendiamo estendere questo corpus in due sensi: verticale e orizzontale. L'estensione orizzontale riguarda l'aggiunta di altri testi giuridici, mentre quella verticale ha a che fare con il trattamento automatico del corpus a livelli più avanzati.

Bibliografia

Dell'Orletta F. 2009. Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian.* Reggio Emilia, Italy.

Delmonte R. 2007. *VEST* - Venice Symbolic Tagger. In *Intelligenza Artificiale*, Anno IV, N° 2, pp. 26-27.

Diab, M. 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt

Gamper, J. 1998. CATEX– A Project Proposal. In Academia, 14, 10-12

Hammouda S. 2010. Small Parallel Corpora in an English-Arabic Translation Classroom: No Need to Reinvent the Wheel in the Era of Globalization. In Shiyab, S., Rose, M., House, J., Duval J.,(eds.), *Globalization and Aspects of Translation,* UK: Cambridge Scholars Publishing

Lenci A., Montemagni S., Pirrelli V. 2012. *Testo e computer: elementi di linguistica computazionale*, Carocci editore, Roma

Rossini Favretti R., Tamburini F., Martelli E. 2007. Words from Bononia Legal Corpus. In *Text Corpora and Multilingual Lexicography* (W.Teubert ed.), John Benjamins

Samy, D., Moreno-Sandoval, A., Guirao, J.M., Alfonseca, E. 2006. Building a Multilingual Parallel Corpus Arabic-Spanish-English. In *Proceedings of International Conference on Language Resources and Evaluation LREC-06*, Genoa, Italy

Zotti, P. 2013. Costruire un corpus parallelo Giapponese-Italiano. Metodologie di compilazione e applicazioni. In Casari, M., Scrolavezza, P. (eds), *Giappone, storie plurali,* I libri di Emil-Odoya Edizioni. Bologna

Habash, N., Rambow, O., Roth, R. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Choukri, K., Maegaard, B., editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. The MEDAR Consortium, April.

Habash, N. 2010. Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers.

Picchi E. , Sassolini E. , Nahli O. , Cucurullo S. 1999. Risorse monolingui e multilingui. Corpus bilingue italiano-arabo. In *Linguistica computazionale*, XVIII/XIX, Pisa.

Italian-Arabic domain terminology extraction from parallel corpora

Fathi Fawi, Rodolfo Delmonte
Department of Linguistic Studies and Comparative Cultures
Università Ca' Foscari – Venezia
fathi_fawi@yahoo.com, delmonte@unive.it

Abstract

English. In this paper we present our approach to extract multi-word terms (MWTs) from an Italian-Arabic parallel corpus of legal texts. Our approach is a hybrid model which combines linguistic and statistical knowledge. The linguistic approach includes Part Of Speech (POS) tagging of the corpus texts in the two languages in order to formulate syntactic patterns to identify candidate terms. After that, the candidate terms will be ranked by statistical association measures which here represent the statistical knowledge. After the creation of two MWTs lists, one for each language, the parallel corpus will be used to validate and identify translation equivalents.

Italiano. *In questo lavoro presentiamo il nostro approccio all'estrazione di termini composti da un corpus giuridico parallelo italiano-arabo. In una prima fase vengono estratti termini composti dai corpora monolingui tramite un approccio ibrido che combina le annotazioni linguistiche fornite dal POS tagging con le informazioni statistiche offerte dalle misure di associazione lessicale. In una seconda fase viene utilizzato il corpus parallelo per estrarre equivalenti di traduzione.*

1 Introduction

The development of robust approaches aiming at terminology extraction from corpora plays a key role in a lot of applications related to NLP, such as information retrieval, ontology construction, machine translation, etc. The main approaches adopted to terms extractions are linguistic-based, statistical-based, and hybrid-based. While the linguistic approach tries to identify terms by capturing their syntactic properties, called *synaptic compositions* (Pazienza et al., 2005), the statistical one uses different association measures (Church et al., 1989) to determine the degree of association or cohesiveness between the multi-word terms (MWTs) components. There is no doubt that the use of a hybrid approach, which combines linguistic and statistical information to identify candidate terms, can guarantee best results rather than relying basically on one approach (Frantzi et al., 1999).

In this paper we present our approach to extract MWTs from an Italian-Arabic parallel corpus of legal texts. The rest of this paper is organized as follows: in Section 2 we present related works; Section 3 describes our proposed approach to extract MWTs from parallel corpora; Section 4 presents the experiments and the results; and Section 5 explains the Conclusion and future works.

2 Related works

There are a lot of efforts that have been done to extract MWTs from monolingual corpora both in Italian (Bonin et al., 2010, Basili et al., 2001) and Arabic (El Mahdaouy et al.,2013, Al Khatib et al., 2010, Abed et al, 2013). The literature of terms extraction from parallel corpora reveals a high dependence on the heuristic methods which calculate the translation probability of terms in the source and target languages. NATools (Simões et al., 2003) uses co-occurrences count of terms in the parallel corpus for building a sparse matrix which will be processed to create a probabilistic translation dictionary for the words of the corpus.

Regarding the domain terminology extraction from parallel texts including Arabic, we can find only rare works, and this may be because of two reasons: a) Arabic is one of those languages which lack specialized parallel corpora in electronic format; b) Arabic is a complex language and its morphosyntactic features affect the overall performance of NLP tasks, especially the bitext word alignment. In (Lahbib et al. 2014) an approach to extract Arabic-English domain

terminology from aligned corpora was presented. The approach consists of the following steps: 1) morphological analysis and disambiguation of the corpus words; 2) extraction of relevant Arabic terms using POS to filter some words, and TF-IDF (Term Frequency- Inverse Document Frequency) to measure the relevance toward one domain; 3) alignment of the texts at the word level, using GIZA++; 4) translations extraction, based on a translation matrix generated from the alignment process, which consists of extracting, for each Arabic word in the corpus, the most likely corresponding translation. To evaluate the approach, a vocalized version of hadith corpus[1] has been used, gaving accuracy rates close to 90%. Here we can note some observations: firstly the approach relies on a probabilistic tool to align the texts at word level. This does not give good results with languages like Arabic which has its own syntactic and morphological features. Secondly, the corpus of evaluation is an Islamic corpus which contains a lot of Islamic terminologies which do not have a translation in other languages, but just transliteration.

Regarding the domain terminology extraction from parallel corpora including the Italian language, we can mention the CLE project (Streiter et al., 2004), where a trilingual corpus with legal texts in Ladin, German and Italian has been created. CLE is stored in a relational database and is accessible via the Internet through BISTRO[2], the Juridical Terminology Information System of Bolzano. Furthermore, there is the LexALP project (Lyding et al., 2006), where sophisticated tools have been developed for the collection, description and harmonization of the legal terminology of spatial planning and sustainable development in four languages, namely French, German, Italian and Slovene.

3 The proposed approach

In this paper we propose a corpus-based approach to extract MWTs from bilingual corpora. It is a hybrid approach which combines statistical methods with linguistic knowledge. Providing the presence of a parallel corpus, the approach consists of the following phases:
1. using POS tagging to create candidate terms in each language;
2. applying statistical methods to rank the candidate terms in order to create a terminology list in each language;
3. using the parallel corpus for identifying translation equivalents of MWTs.

3.1 Morphological analysis

In this phase all the texts of the corpus are tagged at the POS level. The tagging task is done at monolingual level, given its dependency on the language. Regarding the Arabic texts we used the Amira tagger (Diab, 2009), which is based on a supervised learning approach. Amira system uses Support Vector Machine (SVM) for the processing of Modern Standard Arabic texts. In our case the POS tagging accuracy is close to 94%.

Regarding the Italian texts we used the VEST tagger (Delmonte, 2007). Vest is a symbolic rule tagger that uses little quantitative and statistical information. It is based on tagged lexical information and uses a morphological analyzer for derivational nouns, cliticized verbs and some adjectives. Vest has achieved around 95,7% of accuracy.

3.2 Create candidate terms

In this step we use the POS tagging and sequence identifier to form syntactic patterns in order to extract monolingual candidate terms which fit the rules of the grammar. For Arabic, we used the patterns proposed by El Mahdaouy et al.(2013):
–(Noun + (Noun|ADJ) + |(Noun|ADJ) + |(Noun|ADJ))
–Noun Prep Noun
For the Italian texts, we used the following set of POS patterns, proposed by Bonin et al. (2010):
Noun+(Prep+(Noun|ADJ)+|Noun|ADJ)+

3.3 Statistical filter

To rank the candidate MWTs and separate terms from non-terms, we used two statistical methods: Log-Likelihood Ratio (LLR) (Dunning, 1993) as *unithood* measure to rank the candidate terms extracted in the last phase; and C-NC value method as described in Frantzi et al., (1999) as the measure of *termhood*, i.e., for extracting relevant terms from those ranked by LLR.

3.3.1 Likelihood ratio

1 http://library.islamweb.net/hadith/index.php
2 http://www.eurac.edu/bistro

LLR is a widely used statistical test for hypothesis testing. LLR is a more suitable hypothesis testing method for low-frequency terms. For bi-grams the LLR is defined as the following:

LLR $(w_1, w_2) = Nw_1;w_2 \log(Nw_1;w_2) + Nw_1;\text{-}w_2 \log(w_1;\text{-}w_2) + N\text{-}w_1;w_2 \log(N\text{-}w_1;w_2) + N\text{-}w_1;\text{-}w_2 \log(N\text{-}w_1;\text{-}w_2) - (Nw_1;w_2 + w_1;\text{-}w_2) \log(Nw_1;w_2 + w_1;\text{-}w_2) - (Nw_1;w_2 + N\text{-}w_1;w_2) \log(Nw_1;w_2 + N\text{-}w_1;w_2) - (w_1;\text{-}w_2 + N\text{-}w_1;\text{-}w_2) \log(w_1;\text{-}w_2 + N\text{-}w_1;\text{-}w_2) - (N\text{-}w_1;w_2 + N\text{-}w_1;\text{-}w_2) \log(N\text{-}w_1;w_2 + N\text{-}w_1;\text{-}w_2) + N \log(N)$,

where $Nw_1;w_2$ is the number of terms in which w_1 *and* w_2 co-occur; $Nw_1;\text{-}w_2$ is the number of terms in which only w_1 occurs; $N\text{-}w_1;w_2$ is the number of terms in which only w_2 occurs; $N\text{-}w_1;\text{-}w_2$ is the number of terms in which neither w_1 nor w_2 occurs; and N is the number of extracted terms.

3.3.2 C-NC value

The method C-NC value combines linguistic and statistical information (Frantzi et al.,1999). The first component, C-value measures the *termhood* of a candidate string using its statistical characteristics which are: number of occurrence; term nesting, which means the frequency of the candidate string as part of other longer candidate terms; the number of these longer candidate terms; and the length of the candidate string. It is defined as:

$$\text{C-value(a)} = \begin{cases} \log_2 (|a|) \cdot f(a) & \text{if } a \text{ is not nested,} \\ \log_2 \left((a() . \left(f(a) - \frac{1}{p(T_a)} \sum f(b) \right) \right) & \text{otherwise,} \end{cases}$$

where a is the candidate string; $|a|$ is the length in words of a; $f(a)$ is its frequency of occurrence in the corpus; T_a is the set of extracted candidate terms that contain a; $p(T_a)$ is the number of these candidate terms, and $\sum f(b)$ are the sum of frequency by which a appears in longer strings. As we can see if the candidate string is not nested, its *termhood* score will be based on its total frequency in the corpus and its length. If it is nested, the *termhood* will consider its frequency as a nested string and the number of the longer strings into which it appears.

The NC-value component combines the C-value of a candidate string together with the contextual information. By *term context words* we mean the words which appear in vicinity of the extracted candidate terms in the text. A word can be defined as a term context word on the basis of the number of terms into which it appears. The criterion is that the higher the number of terms in which a word appears, the higher the likelihood that the word is a context word and that it will occur with other terms. So the weight of a context word will be calculated in this way: $weight(w) = \frac{a(w)}{n}$, where w is the context word; $a(w)$ is the number of terms into which w appears; n is the total number of candidate terms. So the N-value $(a) = \sum f_a(w) \times weight(w)$, where $f_a(w)$ is the frequency of w as a context word of the term a and C_a is the set of context words of a. This measure is combined with the C-value to provide the C-NC value:

$$\text{C-NC value(a)} = 0.5 \times C - value(a) + 0.5 \times N - value(a)$$

In our case the C-NC value receive as input the output of the *unithood* measures, namely LLR.

3.4 Identification of translation equivalents

The MWTs lists extracted by the C-NC-value in both languages will be recovered in the parallel corpus. The terms in their context will receive a marked format, using square brackets, to be distinguished from the rest of the words in the corpus. Then we used another algorithm to identify translation equivalents of terms from the parallel corpus. In every translation unit, which contains a source sentence with its target translation, created in TMX format, the system searches the terms between square brackets in both source and target languages. Primarily the system collects in a dictionary the bilingual terms for every translation unit present in the parallel corpus. Afterwards the system will validate the real translation equivalents in the dictionary. The relations types in the bilingual terms dictionary will be as follows:

- one2one
- many2many
- many2one
- one2many

} positive relations

- one2null
- many2null
- null2one
- null2many

} negative relations

After excluding the negative relations, since they will not produce translation equivalents, the

system uses the following method for validating relevant equivalents of translation:
a) We use the LLR test, as described above, for estimating the association degree between the bilingual MWTs. In this case the system uses the statistical features of every bilingual MWTs pair in the parallel contest for calculating its LLR value.
b) As a second step the system uses a SMT, namely Google Translate: the idea here is that by means of the translation of the MWTs components the system can identify valid translation equivalents.
c) For the translation pairs, which the LLR test and SMT system failed to identify, the system can use the MWTs index in the parallel context. This last choice relies on the idea that for our language pair the index of the words in the context can be considered a good indicator of translation relation. Within every translation unit, the code combines the words with the closest index in the bilingual context, with distance threshold value = 4.

4 Experiments and Results

4.1 The Corpus

We applied the approach to an Italian-Arabic parallel corpus specialized in the domain of international law (Fawi, 2015). The corpus comprises approximately one million words and is aligned at sentence level.

Italian MWTs	Arabic MWTs
1- camera d'appello 2- mandato d'arresto 3- responsabilità penale individuale 4- diritto internazionale umanitario 5- tenta di commettere il reato	1-دائرة الاستئناف 2- أمر بالقبض 3- المسؤولية الجنائية الفردية 4- القانون الإنساني الدولي 5- الشروع في ارتكاب الجريمة

Table 1. Italian-Arabic equivalent MWTs

4.2 Evaluation

The evaluation process of the term recognition system is a very complex task, not only because there is no specific gold standard for evaluating and comparing different MWTs extraction approaches, but also for the intrinsic nature of the *term* for which it is difficult to give a precise linguistic definition (Pazienza et al., 2005). Since there is no reference list against which we can measure the performance of our approach, we decided to carry out the evaluation mainly by manual validation. The approach validation consists of two parts: MWTs extraction from monolingual corpus (Table 2, 3) and MWTs extraction from parallel corpus (Table 4).

Measure	Arabic		Italian	
	precision	recall	precision	recall
LLR	84%	74%	89%	80%

Table 2. Evaluation of the *unithood* measure

Measure	Arabic			Italian		
	n-best 100	n-best 300	n-best 500	n-best 100	n-best 300	n-best 500
C-NC value	84%	75%	69%	85%	80%	77%

Table 3. Precision of the C-NC value applied on the output of LLR with n-best = 100, 300, 500

measures	recall	precision
LLR	70 %	86 %
SMT system	51 %	88 %
Context Index	50 %	70 %

Table 4. Evaluation of the translation equivalents extraction

5 Conclusion

In this paper we presented our proposed approach to extract multi-word terms from parallel corpora in the legal domain. Regarding the monolingual extraction, we can observe that the results in Italian are a little higher than those in Arabic and this is due to the morphological complexity of the Arabic language which has an impact on the POS tagging performance and therefore on the MWTs extraction. Regarding the bilingual extraction we note that the mediocre recall in SMT system is due to the legal peculiarity of the corpus terms which do not always correspond to the Google translation, while the low recall in the method based on the MWTs index can be attributable to the limited reordering between the two languages. We believe that our attempt can be considered the first one of its type in the Arabic-Italian bilingual domain terminology extraction, and that the results are encouraging. Future work will focus on improving the performance of the approach.

References

Abed, A. M, Tiun, S., and Albared, M., 2013. Arabic Term Extraction Using Combined Approach On Islamic Document. In *Journal of Theoretical & Applied Information Technology*, vol. 58, no. 3, pp. 601 – 608.

Al Khatib, K., Badarneh, A. 2010. Automatic extraction of arabic multi-word terms. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pp. 411-418.

Attia M., Tounsi L., Pecina P., van Genabith J., Toral A. Automatic Extraction of Arabic Multiword Expressions. In *COLING 2010 Workshop on Multiword Expressions: from Theory to Applications*. Beijing, China

Basili, R., Moschitti, A., Pazienza, M., Zanzotto, F. 2001. A contrastive approach to term extraction. In *Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA)*, France, pp.119-128

Bonin, F., Dell'Orletta, F., Venturi, G., and Montemagni, S. 2010. A contrastive approach to multi word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, La Valletta,Malta, 19–21 May, pp. 3222–3229

Boulaknadel, S., Daille, B., Aboutajdine, D. 2008. A multi-word term extraction program for arabic language. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, LREC, pp. 1485-1488.

Church K.W., Hanks P. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the Association of Computational Linguistics*, pp.76–83

Delmonte R. 2007. VEST - Venice Symbolic Tagger. In *Intelligenza Artificiale*, Anno IV, N° 2, pp. 26-27

Diab, M. 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egyp

Dunning T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. In *Computational Linguistics,* vol.19, No.1, pp. 61-74.

El Mahdaouy A., Ouatik S., Gaussier E. 2013. A Study of Association Measures and their Combination for Arabic MWT Extraction. In *10th International Conference on Terminology and Artificial Intelligence*, Paris, France

Fawi, F, 2015. Costituzione di un corpus giuridico parallelo italiano-arabo. To appear in *Second Italian Conference on computational Linguistics CliC-it 2015*, 3-4 December 2015, Trento.

Frantzi K., Ananiadou S. 1999. The C–value / NC Value domain independent method for multi–word term extraction. In *Journal of Natural Language Processing*, 6(3), pp.145–179

Lahbib W., Bounhasm I., Elayed, B. 2014. Arabic - English domain terminology extraction from aligned corpora. In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences. Lecture Notes in Computer Science*, Vol. 8841, Springer, pp. 745-759

Lyding, V., Chiocchetti, E., Sérasset, G., Brunet-Manquat, F. 2006. The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, Sydney, pp. 25-31

Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge Mining*, Springer Verlag, pp.255-279

Simões, A. and Almeida, J.J. 2003. NATools: A Statistical Word Aligner Workbench. In *Procesamiento del Lenguaje Natural*, 31, pp.217-224

Streiter, O., Stuflesser M., Ties, I. 2004: CLE, an aligned. Tri-lingual Ladin-Italian-German Corpus. Corpus Design and Interface. In *LREC 2004, Workshop on First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*, May 24

Annotating opposition among verb senses: a crowdsourcing experiment

Anna Feltracco[1-2], Elisabetta Jezek[2], Bernardo Magnini[1], Simone Magnolini[1-3]
[1]Fondazione Bruno Kessler, Povo-Trento
[2] University of Pavia, Pavia
[2] University of Brescia, Brescia
{feltracco,magnini,magnolini}@fbk.eu, jezek@unipv.it

Abstract

English. We describe the acquisition, based on crowdsourcing, of opposition relations among Italian verb senses in the T-PAS resource. The annotation suggests the feasibility of a large-scale enrichment.

Italiano. *Descriviamo l'acquisizione, basata su crowdsourcing, di relazioni di opposizione tra sensi di verbi italiani nella risorsa T-PAS. L'annotazione mostra la fattibilità di un arricchimento su larga scala.*

1 Introduction

Several studies have been carried out on the definition of opposition in linguistics, philosophy, cognitive science and psychology. Our notion of opposition is based on lexical semantic studies by Lyons (1977), Cruse (1986; 2002; 2011), and Pustejovsky (2000), as synthesized in Jezek (2015).

The category of opposites can be said to include pairs of terms that contrast each other with respect to one key aspect of their meaning, such that together they exhaust this aspect completely. Examples include the following pairs: *to open* / *to close*, *to rise* / *to fall*. Paradoxically, the first step in the process of identifying a relationship of opposition often consists in identifying something that the meanings of the words under examination have in common. A second step is to identify a key aspect in which the two meanings oppose each other[1]. Opposites cannot be true simultaneously of the same entity, for example a *price* cannot be said *to rise* and *to fall* at exactly the same point in time.

It is an open discussion whether opposition is a semantic or a lexical relation (Murphy, 2010; Fellbaum, 1998); what is clear is that the predicate that is considered opposite of another predicate does not activate this relation for all its senses. For example, the Italian verb *abbattere* is considered opposite to *costruire* as far as the former is considered in its sense of *to destroy (a building)*, and the latter in its sense of *to build (a building*). The opposition relation does not hold if *abbattere* is considered in its sense of *to kill (an animal)*.

Oppositions between verbs senses are poorly encoded in lexical resources. English WordNet 3.1 (Miller et al., 1990) tags oppositions among verb senses using the label *antonymy*; for example, increase#1 is in antonymy relation with decrease#1, diminish#1, lessen#1, fall#11. In VerbOcean (Chklovski and Pantel, 2004), opposition (*antonymy*) is considered a symmetric relation between verbs, which includes several subtypes; the relation is extracted at verb level (not at sense level). FrameNet (Ruppenhofer et al., 2010), on the other hand, has no tag for the opposition relation, although a subset of cases can be traced via the *perspective on* relation. As regards Italian, in MultiWordNet (Pianta et al., 2002) the opposition relation (labeled: *antonymy relation*) is considered a lexical relation and is represented in the currently available version for English, but not for Italian. In SIMPLE (Lenci et al., 2000), the opposition relation (*antonymy*) is considered a relation between word senses and it has been defined for adjectives (e.g., *dead/alive* and *hot/cold*), although the authors specify it can possibly be extended also to other parts of speech. In Senso Comune (Oltramari et al., 2013) the annotation of the opposition relation appears not to be implemented, even if the tag for the relation (*antonimia*) is present.

The experiment described in the paper focuses

[1]According to Cruse, opposites indicate the relation in which two terms typically differ along only one dimension of meaning: in respect of all other features they are identical (Cruse, 1986, p.197).

on the annotation of opposition relations among verb senses. We annotate these relations in the lexical resource T-PAS (Jezek et al., 2014), an inventory of typed predicate argument structures for Italian manually acquired from corpora through inspection and annotation of actual uses of the analyzed verbs. The corpus instances associated to each T-PAS represent a rich set of grounded information not available in other resources and facilitate the interpretation of the different senses of the verbs [2].

We collected data using crowdsourcing, a methodology already used in other NLP tasks, such as Frame Semantic Role annotation (Fossati et al., 2013; Feizabadi and Padó, 2014), Lexical Substitution (Kremer et al., 2014), Contradictory Event Pairs Acquisition (Takabatake et al., 2015).

2 The T-PAS Resource

The T-PAS resource (Jezek et al., 2014) is a repository of Typed Predicate Argument Structures for Italian acquired from corpora by manual clustering of distributional information about Italian verbs. T-PASs are corpus-derived verb patterns with specification of the expected semantic type (ST) for each argument slot, such as [[Human]] *guida* [[Vehicle]]. T-PAS is the first resource for Italian in which semantic selection properties and sense-in-context distinctions of verbal predicates are characterized fully on empirical ground. We discover the most salient T-PASs using a lexicographic procedure called Corpus Pattern Analysis (CPA) (Hanks, 2004), which relies on the analysis of co-occurrence statistics of syntactic slots in concrete examples found in corpora.

The resource consists of three components[3]:

1. a repository of corpus-derived T-PASs linked to lexical units (verbs);
2. an inventory of about 230 corpus-derived semantic types for nouns (HUMAN, EVENT, BUILDING, etc.), relevant for the disambiguation of the verb in context (see Table 1)
3. a corpus of sentences instantiating the T-PASs[4].

[2]The experiment is part of a broader project consisting in enriching the T-PAS resource with the annotation of different types of opposition relation not present in other resources.

[3]The first release of T-PAS contains 1000 analyzed average polysemy verbs. T-PAS is freely available under a Creative Common Attribution 3.0 license at tpas.fbk.eu.

[4]The reference corpus is a reduced version of ItWAC (Baroni and Kilgarriff, 2006).

Verb: *abbattere*
- ▷ **T-PAS 1:**
  [[Human]] *abbattere* [[Animate]]
  - ▷ **Annotated Corpus:**
    ..Kenai, il più giovane, *abbatte* l' orso..
    ..un bracconiere *abbatteva* un coniglio..
    ...
- ▷ **T-PAS 2:**
  [[Human | Event]] *abbattere* [[Building]]
  - ▷ ...
- ▷ **T-PAS n°**
  - ▷ ...

Table 1: T-PAS Resource Structure.

At present, T-PASs are not linked by any semantic relation. Our experiment extends the resource by adding opposition relations among T-PASs following a pilot experiment described in Feltracco et al. (2015). In the following sections, we illustrate the annotation tasks we elaborated and the new experiments we performed, together with their evaluation.

3 Annotation Tasks

In this section we define the annotation tasks (Section 3.1), how such tasks have been implemented using T-PAS (Section 3.2) and the crowdsourcing platform we used to collect the data (Section 3.3).

3.1 Tasks Definition

In order to annotate the opposition relation among T-PASs, we have set the experiment in two steps. In the first step (Task A) we want to determine if there is an opposition relation between a certain sense of a verb (the *source verb*) and another verb (the *target verb*); in the second step (Task B) we want to identify which sense of the Target Verb holds the opposition relation with the source verb, if identified in Task A.

As for Task A (see an example in Table 2), we showed annotators a pair of sentences: S1 (i.e. Frase 1 in Table 2) is a sentence, extracted from the annotated corpus of T-PAS, that contains the source verb (in bold), while S2 (i.e. Frase 2 in Table 2) is identical to S1, with the exception of the source verb, which is substituted with the target verb (in bold). Annotators were asked to compare the two sentences and choose one among the following options: A1) S2 makes sense and holds an opposition relation with S1, or A2) S2 makes sense but it does not hold an opposition relation with S1, or A3) S2 does not make sense [5].

[5]Task A is comparable to a Lexical Substitution task. For

Confronta le seguenti frasi.	
Frase 1: L' appello va, pertanto, **respinto**. (annotated example of T-PAS 1 of the source verb = *respingere*)	
Frase 2: L' appello va, pertanto, **approvato**. (Target Verb = *approvare*)	
Task A: "Diresti che la Frase 2 ha un senso compiuto? Se sì, diresti che c'é una relazione di opposizione tra le Frasi?	
A1: Frase 2 ha senso e si oppone alla Frase 1 (30.9%)	Inter-Annotator Agreement
A2: Frase 2 ha senso ma non si oppone alla Frase 1 (5.1%)	Ao: 72.4%
A3: Frase 2 non ha senso (64%)	Fleiss's coefficient: 0.44
Task B: "Leggi le seguenti frasi. In quali frasi **approvare** ha lo stesso senso della Frase 2?"	
B1: La Commissione **approva** l'emendamento 2.15 del relatore.	
B2: Gli astronauti hanno **approvato** l' uso del TVIS in questa configurazione.	Average Agreement
B3: In ogni caso , non verranno **approvati** i candidati che abbiano registrato assenze superiori a un terzo del numero complessivo di ore di lezione previste.	Normalized Ao: 71.7%
	M. A. Fleiss's coefficient: 0.32
B4: Nessuna delle precedenti	

Table 2: Example and Results for Task A and Task B (Ao: Observed Agreement, M.A.: Macro Average).

If a relation of opposition was identified, we asked annotators to complete Task B. In Task B, annotators had to consider the target verb in S2 and select among a list of sentences containing that verb, those in which the target verb has the "same" meaning as in S2 (Table 2) [6].

3.2 Tasks Implementation

Tasks implementation required: (i) the selection of the source verb and target verb, (ii) the extraction of the examples (for S1), and (iii) the substitution of the verb in the examples (for S2).

Verbs Selection. For the annotation of the T-PAS resource, we selected pairs of verbs (source verb and target verb) according to three conditions: (i) both verbs are present in the T-PAS resource; (ii) both verbs appear in the Dizionario dei Sinonimi e dei Contrari - Rizzoli Editore[7] as lemmas; (iii) the target verb is annotated as *contrary* for the source verb and viceversa in the Dizionario dei Sinonimi e dei Contrari; thus, for each pair source verb A - target verb B, also the pair source verb B - target verb A has been considered. The total number of verb pairs extracted according to these criteria is 436. Since our aim is to annotate opposition among verb senses, we implemented Task A for each of the T-PASs of the source verbs (i.e. for T-PAS 1 of the source verb *abbattere*, for T-PAS 2, for T-PAS 3, ..), for a total of 2263 T-PASs.

Examples Extraction. In order to increase the reliability of the annotation, we extracted up to three examples for each sense of the verbs from the T-PAS resource, according to their availability in the resource (i.e., we extracted up to three examples for the T-PAS 1 of the source verb *abbattere*, up to three examples for the T-PAS 2, ..). We discarded examples annotated as "non regular" such as metonymical uses and, to simplify the task, we selected the shortest examples, composed by at least 5 tokens. The extracted examples for a verb have been used both as the S1 in Task A (when it is the source verb), and as the answers proposed in Task B (when it is the target verb).

Verb Substitution. We generated S2 from S1 substituting the source verb with the target verb automatically conjugated accordingly, using the library: `italian-nlp-library`[8]. The library analyzes only the verb and not the whole sentence and does not manage all the suffixes; to solve this we added some simple rules. This system grants a quick implementation avoiding parsing or deeper analysis of the sentence.

3.3 Crowdflower Platform Settings

For crowdsourcing we used the Crowdflower platform[9], with the following parameter setting. We

instance, in McCarthy and Navigli (2009) and Kremer et al. (2014), annotators are asked to provide a synonym for a word in a sentence that would not change the meaning of the sentence. In our case we asked annotators to validate the sense of a sentence in which a word is substituted with a supposed opposite.

[6] In other Word Sense Disambiguation (WSD) tasks, e.g. in Mihalcea (2004), annotators are asked to select among a sense inventory. By contrast, we showed non-expert annotators in the crowd the target verb in context, taking advantage of the availability of examples of the verb in the resource.

[7] http://dizionari.corriere.it/dizionario_sinonimi_contrari

[8] https://github.com/jacopofar/italian-nlp-library

[9] http://www.crowdflower.com

initially set the payment to 0.04 USD, then to 0.05 USD for each page and the number of sentence pairs for page to 5[10]. One out of these 5 pairs was a Test Question (TQ), i.e. a question for which we already know the answer. If an annotator misses many TQs s/he is not permitted to continue the annotation and his/her judgments are rejected: we set the threshold of this accuracy to 71%. We selected the TQs among the total sentence pairs and we annotated them before lunching the tasks. We also set parameters in order to have annotators with Italian Language skills.

4 Results and Discussion

A total of 712 pairs of sentences has been annotated with 3 judgments in almost a month, for a total of 2136 judgments (plus judgments for TQs).

For Task A, the overall inter-annotator agreement (IAA) calculated using *Fleiss's coefficient* (Artstein and Poesio, 2008) is 0.44, with an Observed Agreement (Ao) of 72.4%. Overall, answer A1 was chosen 30.9% of the times, answer A2 5.1% and answer A3 64%.

We observe many cases in which a mismatch between the verb (in any of its meanings) and the new context in which the verb is inserted invalidates the sense of the sentence in its entirety. For instance, in Example 1, where "ridare" is the source verb and "trattenere" the target verb, the relation between the target verb and the direct object argument produces a sentence which has no sense. The pair has been judged as "Frase 2 non ha senso" by the three annotators, since you can "ridare un esame" ("take an exam again") but not "trattenere un esame" (*"to hold, to keep an exam").

(1) S1: Posso **ridare** un esame già sostenuto?
S2: Posso **trattenere** un esame già sostenuto?

Other cases in which the three annotators chose "Frase 2 non ha senso" depend on the relation between the verb and other elements of the sentence. Example 2 shows a case with a coordinative structure between two events: in S1 somebody has been "imprisoned *and* deported", in S2 somebody has been "released *and* deported". We believe that annotators judged the two events in S2 as incompatible.

(2) S1: Era stato **incarcerato** e deportato.
S2: Era stato **liberato** e deportato.

Task B was proposed to annotators only if an opposition had been identified in Task A (i.e. answer A1). Results are calculated for the pairs which collected a minimum of two (out of three) answers A1, for a total of 211 pairs. We calculated the IAA for each sense, considering a match when annotators agree both on selecting and not selecting a sentence (i.e. a sense). The overall average Ao, normalized by the number of annotators, is 71.7%. In addiction, we calculated a *Macro Average-Fleiss'coefficient* (Mihalcea et al., 2004), where also the Expected Agreement (Ae) and the Fleiss'coefficient were determined for each pair, and then combined in an overall average. We calculated Ae *a posteriori*, considering the distribution of judgments of annotators, resulting in a *Macro Average-Fleiss'coefficient* of 0.32 [11].

As regards the crowdsourcing methodology, although the use of examples in place of sense definition simplified the annotation, the tasks were considered rather difficult by many annotators and most of them were discarded for low accuracy in the initial page (which has only TQs), especially for missing TQs for Task B.

5 Conclusion and Further work

In this paper we have presented a crowdsourcing experiment for the annotation of the opposition relation among verb senses in the Italian T-PAS resource.

The annotation experiment has shown the feasibility of collecting opposition relation among Italian verb senses through crowdsourcing. We propose a methodology based on the automatic substitution of a verb with a candidate opposite and show that the IAA obtained using sense examples is comparable with the IAA obtained by other annotations based on sense definitions.

Ongoing work includes further annotation of the opposition relations in T-PAS using crowd answers and a deep examination of the causes which lead to the generation of sentences with no sense.

[10] In addition to Task A and B, annotators were asked another question concerning the relation among the two verbs. In this paper we are not discussing this further question.

[11] These values are similar to the rates reported in other WSD tasks using definitions of senses and not examples; e.g. IAA in Senseval-2 Verb Lexical Sample by expert annotators (Palmer et al., 2006) is 71% and in Senseval-3 by Contributors over the Web (Mihalcea et al., 2004) IAA is 67.3% with a Macro Average-K of 0.35. However in these tasks the IAA was computed somewhat differently (Palmer et al., 2006).

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, volume 2004, pages 33–40, Barcelona, Spain, July.

D. Alan Cruse. 1986. *Lexical semantics*. Cambridge University Press.

D. Alan Cruse. 2002. Paradigmatic relations of exclusion and opposition II: Reversivity. *Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen: Lexicology: An international handbook on the nature and structure of words and vocabularies*, 1:507–510.

D. Alan Cruse. 2011. *Meaning In Language: An Introduction To Semantics And Pragmatics*. Oxford University Press, USA.

Parvin Sadat Feizabadi and Sebastian Padó. 2014. Crowdsourcing annotation of non-local semantic roles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 226–230, Gothenburg, Sweden, April. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Anna Feltracco, Elisabetta Jezek, and Bernardo Magnini. 2015. Opposition relations among verb frames. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 16–24, Denver, Colorado, June. Association for Computational Linguistics.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing FrameNet to the crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 742–747. Association for Computational Linguistics.

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 Typed 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.

Elisabetta Jezek. 2015. *The Lexicon. An Introduction.* Oxford: Oxford University Press.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden, April. Association for Computational Linguistics.

Alessandro Lenci, Federica Busa, Nilda Ruimy, Elisabetta Gola, Monica Monachini, Nicoletta Calzolari, and Antonio Zampolli. 2000. Linguistic specifications deliverable d2 , Technical report, University of Pisa and Institute of Computational Linguistics of CNR, Pis.

John Lyons. 1977. Semantics, Vol. I. *Cambridge: Cambridge*.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language resources and evaluation*, 43(2):139–159.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *In Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain, July.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244.

M. Lynne Murphy. 2010. *Lexical meaning*. Cambridge University Press.

Alessandro Oltramari, Guido Vetere, Isabella Chiari, Elisabetta Jezek, Fabio Massimo Zanzotto, Malvina Nissim, and Aldo Gangemi. 2013. Senso Comune: a collaborative knowledge resource for Italian. In *The People's Web Meets NLP*, pages 45–67. Springer.

Martha Palmer, Hwee Tou Ng, and Hoa Trang Dang. 2006. Evaluation of WSD systems. In *Word Sense Disambiguation: Algorithms and Applications*, pages 75–106. Springer.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, volume 152, pages 55–63.

James Pustejovsky. 2000. Events and the semantics of opposition. In *Events as grammatical objects*, pages 445–482. Stanford, CA: CSLI Publications.

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended theory and practice. retrieved November 12, 2013.

Yu Takabatake, Hajime Morita, Daisuke Kawahara, Sadao Kurohashi, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2015. Classification and acquisition of contradictory event pairs using crowdsourcing. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 99–107, Denver, Colorado, June. Association for Computational Linguistics.

Gold standard vs. silver standard: the case of dependency parsing for Italian

Michele Filannino
The University of Manchester
School of Computer Science
M13 PL, Manchester (UK)
filannim@cs.man.ac.uk

Marilena Di Bari
University of Leeds
School of Languages, Cultures and Societies
LS2 9JT, Leeds (UK)
mlmdb@leeds.ac.uk

Abstract

English. Collecting and manually annotating gold standards in NLP has become so expensive that in the last years the question of whether we can satisfactorily replace them with automatically annotated data (silver standards) is arising more and more interest. We focus on the case of dependency parsing for Italian and we investigate whether such strategy is convenient and to what extent. Our experiments, conducted on very large sizes of silver data, show that quantity does not win over quality.

Italiano. *Raccogliere e annotare manualmente dati linguistici gold standard sta diventando oneroso al punto che, negli ultimi anni, la possibilita' di sostituirli con dati annotati automaticamente (silver) sta riscuotendo sempre piu' interesse. In questo articolo indaghiamo la convenienza di tale strategia nel caso dei dependency parser per l'italiano. Gli esperimenti, condotti su dati silver di grandissime dimensioni, dimostrano the la quantita' non vince sulla qualita'.*

1 Introduction

Collecting and manually annotating linguistic data (typically referred to as *gold* standard) is a very expensive activity, both in terms of time and effort (Tomanek et al., 2007). For this reason, in the last years the question of whether we can train good Natural Language Processing (NLP) models by using just automatically annotated data (called *silver* standard) is arising interest (Hahn et al., 2010; Chowdhury and Lavelli, 2011).

In this case, human annotations are replaced by those generated by pre-existing state-of-the-art systems. The annotations are then merged by using a committee approach specifically tailored on the data (Rebholz-Schuhmann et al., 2010a). The key advantage of such approach is the possibility to drastically reduce both time and effort, therefore generating considerably larger data sets in a fraction of the time. This is particularly true for text data in different fields such as temporal information extraction (Filannino et al., 2013), text chunking (Kang et al., 2012) and named entity recognition (Rebholz-Schuhmann et al., 2010b; Nothman et al., 2013) to cite just a few, and for non-textual data like in in medical imaging recognition (Langs et al., 2013).

In this paper we focus on the case of dependency parsing for the Italian language. Dependency parsers are systems that automatically generate the linguistic dependency structure of a given sentence (Nivre, 2005). An example is given in Figure 1 for the sentence "Essenziale per l'innesco delle reazioni è la presenza di radiazione solare." (The presence of solar radiation is essential for triggering the reactions). We investigate whether the use of very large silver standard corpora leads to train good dependency parsers, in order to address the following question: *Which characteristic is more important for a training set: quantity or quality?*

The paper is organised as follows: Section 2 presents some background works on dependency parsers for Italian; Section 3 presents the silver standard corpus used for the experiments and its linguistic features, with Section 4 describing the experimental settings and Section 5 describing the results of the comparison between the trained parsers (considering different sizes of data) and two test sets: gold and silver. Finally, the paper's contributions are summed up in Section 6.

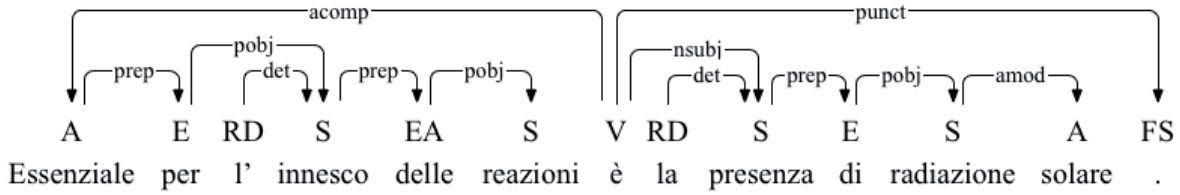


Figure 1: An example of dependency tree for an Italian sentence.

2 Background

Since dependency parsing systems play a pivotal role in NLP, their quality is crucial in fostering the development of novel applications. Nowadays dependency parsers are mostly data-driven, and mainly designed around machine learning classifiers. Such systems "train classifiers that predict the next action of a deterministic parser constructing unlabelled dependency structures" (Nivre, 2005).

Like in the case of other languages, in Italian ad-hoc cross-lingual and mono-lingual shared tasks are organised every year to push the boundaries of such technologies (Buchholz and Marsi, 2006; Bosco et al., 2009; Bosco and Mazzei, 2011; Bosco et al., 2014). The most important shared task about dependency parsing systems for Italian is hosted by the EVALITA series, in which participants are provided with manually annotated training data and the evaluation of their system is performed on a non disclosed portion of the data. Since the different systems presented so far have reached an overall performance close to 90% (Lavelli, 2014), we believe that the question of whether we can start using silver standards is a relevant one.

3 The corpus

The silver standard data comes from a freely available corpus created as part of the project PAISÀ (*Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati*) (Lyding et al., 2014). The project was aimed at "overcoming the technological barriers currently preventing web users from having interactive access to and use of large quantities of data of contemporary Italian to improve their language skills".

The PAISÀ corpus[1] is a set of about 380,000 Italian texts collected by systematically harvesting the web looking for frequent Italian collocations. It consists of about 13M sentences and 265M tokens fully annotated in CoNLL format. The average length of the sentences is about 20 tokens.

The Part-of-Speech tags have been automatically annotated by using ILC-POS-TAGGER (Dell'Orletta, 2009) and the dependency structure by using DeSR Dependency Parser (Attardi et al., 2007), the top performer system at the EVALITA shared task. The POS-tags are annotated according to the TANL tagset[2], whereas the dependency relations follow the ISST-TANL tagset[3]. These automatic annotations have been successively revised and manually corrected on different stages: text cleaning, annotation corrections and tools alignment.

Unfortunately we found out that The PAISÀ corpus includes some sentences which cannot be used for training purposes due to invalid CONLL representations (i.e. duplicated or missing IDs, and invalid dependency relations). These sentences represent the 6.04% of the corpus, yet only the 0.10% of the tokens. This difference shows the presence of many small invalid sentences.

Thus we have created a filtered corpus with the working sentences to which we will refer from now on with the name of *silver* as opposed to the EVALITA corpus as *gold*. In the latter, for training purposes we merged training and development test sets, whereas we did not modify the official test set.

4 Experiments

4.1 Test corpora

We quantitatively measured the performance of the proposed parsers with respect to two test sets: gold and silver.

[1] http://www.corpusitaliano.it/it/contents/description.html

[2] http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset

[3] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

	original	filtered	Δ%
Sentences	13.1M	12.3M	93.96%
Tokens	264.9M	264.6M	99.90%
Sentence length	20.3	21.5	-

Table 1: PAISÀ corpus' statistics. The figures show the presence of many short and invalid sentences.

The gold test set corresponds to the official benchmark test set for the EVALITA 2014 dependency parsing task. It contains 344 sentences manually annotated with 9066 tokens (∼26 tokens per sentence). The silver test set, instead, is composed of 1,000 randomly selected sentences from the silver data, which have not been used for training purposes in the experiments.

4.2 Experimental setting

The experiments have been carried out using eight different sizes of training set from the silver data: 500, 1K, 5K, 25K, 75K, 125K, 250K and 500K sentences. A limitation of the learning algorithm prevented us to consider even larger training sets[4].

We used the Unlabelled Attachment Score (UAS) measure which studies the structure of a dependency tree and assesses whether the output has the correct head and dependency arcs. The choice of UAS measure is justified by the fact that the gold and silver label sets are not compatible.

We trained the models with *MaltParser*[5] v.1.8.1 by using the default parameters.

The overall set of experiments took about a month with 16 CPU cores and 128Gb of RAM.

5 Results

The complete results are presented in Table 2. The 8 parsers trained on silver data perform poorly when tested against the gold test set (∼32%). The same happens for the opposite setting: the parser trained on the gold data and tested on the silver test set (last column of Table 2). By training on one set and testing on another (gold vs. silver), performance immediately drops of about 35%.

When the parser is trained on and tested against the gold data the performance is 85.85%. Such configuration corresponds to the EVALITA14 setting and provides results comparable with the one obtained by the afore-mentioned challenge's participants.

Training set		UAS against	
corpus	size	gold test	silver test
silver	500	30.14	66.11
	1.000	30.95	67.00
	5.000	32.21	69.11
	10.000	32.44	69.56
	25.000	32.83	69.92
	75.000	33.22	69.79
	125.000	33.47	70.27
	250.000	**33.58**	70.23
	500.000	33.20	**71.17**
gold	7.978	**85.85**	**48.30**

Table 2: Parsers' performance against silver and gold test sets. Silver data refers to PAISÀ corpus, whereas gold refers to EVALITA14 training and development set. Silver data have been used for training purposes in different sizes. Sizes are expressed in number of sentences.

The interesting result lies in the fact that providing a dataset 1000 times bigger does not significantly enhance the performance. This is true regardless of the type of test set used: gold (3.06% variance) and silver (4.89% variance). Moreover, training a parser on a data set smaller than its test set does not negatively affect the final performance.

Figure 2 depicts the performance curves for the models trained on silver data only.

In order to allow for the reproducibility of this research and the possibility of using these new resources, we make the dependency parsing models and the used data sets publicly available at `http://www.cs.man.ac.uk/~filannim/projects/dp_italian/`.

6 Conclusions

We presented a set of experiments to investigate the contribution of silver standards when used as substitution of gold standard data. Similar investigations are arising interesting in any NLP subcommunities due to the high cost of generating gold data.

The results presented in this paper highlight two important facts:

[4] The instance×feature matrix exceeds the maximum size allowed by the `liblinear` implementation used.

[5] `www.maltparser.org`

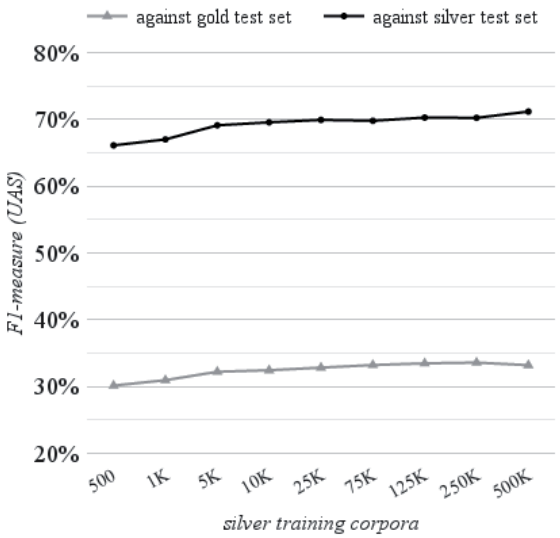


Figure 2: Parsers' performance against silver and gold test sets. In both cases, the models exhibit an asymptotic behaviour. Figures are presented in Table 2. Silver data sizes express the number of sentences. 'K' stands for 1.000.

- The size increase of the training corpus does not provide any sensible difference in terms of performance. In both test sets, a number of sentences between 5.000 and 10.000 seem to be enough to obtain a reliable training. We note that the size of the EVALITA training set lies in such boundary.

- The annotations between gold and silver corpora may be different. This is suggested by the fact that none of the parsers achieved a satisfactory performance when trained and tested on different sources.

We also note that the gold and silver test data sets have different characteristics (average sentence length, lexicon and type of annotation), which may partially justify the gap. On the other hand, the fact that a parser re-trained on annotations produced by a state-of-the-art system (DeSR) in the EVALITA task performs poorly on the very same gold set sheds light on the possibility that such official benchmark test set may not be representative enough.

The main limitation of this study lays in the fact that the experiments have not been repeated multiple times, therefore we have no information about the variance of the figures (UAS column in Table 2). On the other hand, the large size of the data sets involved and the absence of any outlier figure suggest that the overall trends should not change. With the computational facilities available to us for this research, a full analysis of that sort would have required years to be completed.

The results presented in the paper shed light on a recent research question about the employability of automatically annotated data. In the context of dependency parsing for Italian, we provided evidences to support the fact that the quality of the annotation is a far better characteristic to take into account when compared to quantity.

A similar study on languages other than Italian would constitute an interesting future work of the research hereby presented.

Acknowledgements

The authors would like to thank Maria Simi and Roberta Montefusco for providing the EVALITA14 gold standard set, and the two anonymous reviewers who contributed with their valuable feedback. MF would also like to thank the EPSRC for its support in the form of a doctoral training grant.

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Atanas Chanev, and Massimiliano Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using DeSR. In *EMNLP-CoNLLPAISA*, pages 1112–1118.

Cristina Bosco and Alessandro Mazzei. 2011. The EVALITA 2011 parsing task: the dependency track. *Working Notes of EVALITA*, 2011:24–25.

Cristina Bosco, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice DellâOrletta, and Alessandro Lenci. 2009. EVALITA'09 parsing task: comparing dependency parsers and treebanks. *Proceedings of EVALITA*, 9.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. *Proceedings of EVALITA*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Faisal Mahbub Chowdhury and Alberto Lavelli. 2011. Assessing the practical usability of an automatically annotated corpus. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 101–109. Association for Computational Linguistics.

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. *Proceedings of EVALITA*, 9.

Michele Filannino, Gavin Brown, and Goran Nenadic. 2013. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 53–57, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Udo Hahn, Katrin Tomanek, Elena Beisswanger, and Erik Faessler. 2010. A proposal for a configurable silver standard. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 235–242. Association for Computational Linguistics.

Ning Kang, Erik M van Mulligen, and Jan A Kors. 2012. Training text chunkers on a silver standard corpus: can silver replace gold? *BMC bioinformatics*, 13(1):17.

Georg Langs, Allan Hanbury, Bjoern Menze, and Henning Müller. 2013. Visceral: Towards large data in medical imaging—challenges and directions. In *Medical Content-Based Retrieval for Clinical Decision Support*, pages 92–98. Springer.

Alberto Lavelli. 2014. Comparing state-of-the-art dependency parsers for the EVALITA 2014 dependency parsing task. *Proceedings of EVALITA*.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, 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.

Joakim Nivre. 2005. Dependency grammar and dependency parsing. Technical report, Växjö University: School of Mathematics and Systems Engineering.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno-Yepes, Erik M van Mulligen, Ning Kang, Jan A Kors, David Milward, Peter T Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, et al. 2010a. The CALBC silver standard corpus for biomedical named entities-a study in harmonizing the contributions from four independent named entity taggers. In *LREC*.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M Van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010b. CALBC silver standard corpus. *Journal of bioinformatics and computational biology*, 8(01):163–179.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *EMNLP-CoNLL*, pages 486–495.

Phrase Structure and Ancient Anatolian languages
Methodology and challenges for a Luwian syntactic annotation

Federico Giusfredi
Dipartimento di Filologia, Letteratura e Linguistica
University of Verona
federico.giusfredi@gmail.com

Abstract

English For the Marie Skłodowska Curie (MSCA) funded project "SLUW – A computer aided study of the (morpho)-syntax of Luwian" a collection of phrase structure trees from the Luwian corpus is currently being prepared. Luwian is a language belonging to the Anatolian branch of Indo-European; its structures are different from those of English and the language itself is partly obscure. The present paper will describe some special needs, open challenges and methodologies relevant for the annotation of phrase-structure of Luwian.

Italiano *Per il progetto Marie Skłodowska Curie "SLUW – A computer aided study of the (morpho)-syntax of Luwian", è in preparazione un'ampia collezione di alberi sintattici a costituenti per il corpus luvio. Il luvio era una lingua del ceppo anatolico dell'indoeuropeo; la sua struttura è diversa da quella dell'inglese, e la sua decifrazione è in parte incompleta. In questo articolo, saranno discusse alcune necessità, problemi e metodi rilevanti per l'annotazione della sintassi dei costituenti del luvio.*

1 Introduction

Annotating a dead language, especially if lacunae and obscure sequences occur frequently in the corpus, is a challenging task. In the case of phrase-structure trees, those challenges complicate the usual issues represented by "trapping" (an element nested within the boundaries of a phrase it does not belong to) and standard discontinuous phrases.

The language under investigation is Luwian, an ancient member of the Anatolian branch of Indo-European, the second largest one after Hittite by number of documents. It was written using two different writing systems (the cuneiform script and the Anatolian hieroglyphs). The attestations cover a time span of almost one millennium, between the 16th and the 8th centuries BCE (cf. Melchert, 2003).

Syntactically speaking, it features a rather strict SOV word-order as far as some classes of constituents are concerned (Wackernagel particles, inflected verb at the end, left-branching of genitives and attributes); while a few elements can move with relative freedom (for instance adverbs, indirect case NPs and PPs with respect to the position of a direct object).

The final goal of the SLUW project, a Horizon2020 MSCA funded two-year research plan hosted by the University of Verona (2015-2017) is to produce a general study of the syntax (and morpho-syntax) of the language; in order to do so, a significant selection of sentences (about 30% to 50% of the corpus) will be collected and annotated in order to produce phrase-structure trees that will help highlight syntactic patterns. Theory-free phrase structure annotation is more suitable than Universal Dependencies for this kind of approach, as the boundaries of linear and non-linear phrases as well as their canonical or non-canonical position within the sentence are more easily identified.

Since the structure of Luwian is very different from the one of English – Anatolian languages had peculiar features that must be accounted for – the starting point for the development of a POS tagset, the "label-tag" context-sensitive system of the Penn Treebank II, requires to be modified in order to better match the object of study.

2 Expanding the tagset

Different languages have different features, and some of them may be especially relevant for the understanding of the syntax (or of any other aspects of its nature that may be of interest). In the case of Luwian, the Penn POS system (Taylor, Marcus and Santorini, 2003) needs to be expanded on both the phrase and the word level. The following addenda represent the state of the Luwian tagset as of September 2015; other modifications will certainly occur during the future analysis of the corpus.

On the phrase level, the preliminary analysis indicated that the following elements need to be added to the POS labels:

CLP	Clitic "Phrase"
INTR	Introductory particle
QUOT	Direct speech marker

CLP is a pseudo-node (it does not represent a real constituent). In Luwian, a large set of particles with different functions is bound to P2 (2nd word position) – some belonging to the VP, some working on the sentence or inter-phrasal level. While "movement" may be assumed for argumental elements, a proper analysis of some of these clitics has not yet been attempted. They will therefore be analyzed in the position that they actually occupy in the phrase structure, at least during the theory-free phase of annotation.

INTR is a typical element of the Anatolian syntax: an accented particle that works as a coordinating conjunction, but may also open any sentence in which no other accented elements occur before the Wackernagel particles.

Finally, QUOT is a direct speech marker that quite frequently occurs in Wackernagel position.

On the word level, most of the special features of the Anatolian languages can be dealt with by wisely using a functional architecture (matching case endings, verbal inflection; cfr. Taylor, Marcus and Santorini, 2003; also Marcus et al., 1994). Formal markers for nominal elements will include case(-like) specifiers, such as:

-NOM	Nominative
-ACC	Accusative
-GEN	Genitive
-DAT	Dative
-ABL	Ablative
-VOC	Vocative
-NAN	Nom./Acc. (neutra)
-ANT	-*ant*- form (ergative-like)

For verbs, marking endings, time, mood, and voice is also of the utmost importance:

-#S/P	#th person singular/plural
(-)T	Past tense
(-)I	Imperative
(-)MP	Medio-Passive

The case-attributes are important because simply co-indexing elements belonging to the same phrase would make it difficult to assess the cases in which the agreement between two or more elements is not perfect.

This happens in some cases with certain Anatolian modifiers (numerals and nouns do not always agree in number) and with some types of syntactic alignment ("ergative"-like *ant*-forms are modified by attributes in common-gender nominative, and can be anaphorically recalled by neutral pronouns).

Apart from these functional tags, on the word-level specific POS tags also need to be added. For instance, as far as adjectives are concerned:

GJJ	Genitival adjective
PJJ	Possessive adjective
REL	Relative "pronoun"

GJJ represents a peculiar type of synchronically productive adjective that was used to replace the genitive case (cf. Bauer, 2014, 147ff.), an example being *mayas(s)a/i-* "of the adult(s)". It implied a genitival relationship to *maya-* "adult"; it was inflected and agreed with the regens, thus we may have ablative (instrumental):

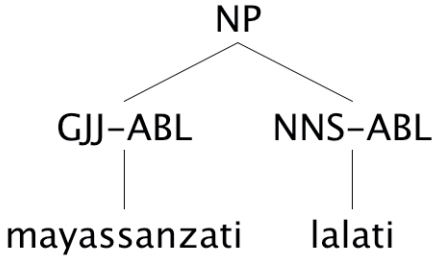
[1] *mayassanzati lalati*

adult=gen.adj.=pl.=abl. tongue=abl.

"The tongues of the adults"

(text KUB 35.24 i 4)

which results in the constituent-structure represented in the following tree.



In case of more complex genitival chains, the nesting of the constituents disambiguates different levels of possession, for instance:

[2] *sasaliya Maritis Zwarimis* FILIUS-*muwiyaya*

sasali=n/a=pl. PN_1=gen. PN_2=gen. son=gen.adj.

sasali's of Maritis, son of Zwarimis

(text Malatya 3, §1)

Tags must therefore be available in order to mark the structure of the phrases and disambiguate from other genitival strategies. PJJ are possessive adjectives similar to English *my*, but they also require inflection and agreement, as in the case of GJJ.

2.1 Subordination and relative clauses

A preliminary analysis has shown that, in some cases, Anatolian subordinate clauses contain a complex set of candidate "nodes" on the level of the SBAR element of the POS tagset, that would roughly correspond to the CP node of a transformational tree: the so-called Anatolian "connectives" (INTR) and subordinating conjunctions may co-occur, and this calls for caution as far as the syntactic representation is concerned.

Consider for instance the following example, in which the syntactic status of the first INTR-element *a* is problematic, because the "complementizer"-slot in the subordinate is already taken by the subordinating conjunction *kuman*, and the "complementizer"-slot of the main clause is occupied by another INTR-element, which makes the intepretation of the subordinate as embedded impossible (or at least very difficult).

[3] **[$_{INTR}$ a]** [$_S$ [$_{SBAR}$ **t-1** [$_{QUOT}$ wa] [$_{VP}$ [$_{NP\text{-}OBJ}$ kummaya DEUS.DOMUS-sa] **[$_{IN\text{-}1}$ kuman]** [$_V$ tamaha]]] **[$_{INTR}$ a** [$_{QUOT}$ wa] [$_{NP}$ mu] [$_{PTCL}$ tta] [$_{VP}$ [$_{DP\text{-}SBJ}$ zanzi kutassarinzi] [$_V$ appan awinta]]]]

"And, when I built the holy temples, these orthostats followed me."

(text Karkemish A11a §§14f.)

The identification of this problem (that also exists in Hittite) has important theoretical consequences regarding the inter-phrasal syntax of Anatolian: "connectives" like *a* were so far consistently presented as coordinating elements, but apparently this is not always the case (cf. Cotticelli-Kurras and Giusfredi 2015).

As for the REL label, the treatment of relative sentences in Anatolian is rather peculiar. The two clauses formally appear to be coordinated; the relative element in the relative clause is frequently referred to a nominal element (Hoffner and Melchert, 2008, 423-424). In such cases, it is inflected to agree with the noun, and is recalled by a pronoun in the main clause. A pseudo-English example can be the following:

[4] ***to what man** you spoke, **that** is a liar*

what=dat. man=dat. you spoke, that=nom. is a liar

Therefore, the REL element needs to be assigned the range of attributes of an adjective.

3 Lacunae and *cruces*

Lacunae in a text preserved on a clay tablet – or on any other kind of perishable support – may interfere with the parsing of syntactic structures. So does the presence of segments or sequences of segments that have not been fully deciphered.

From the point of view of phrase-structure annotation, these two peculiarities of the corpora of ancient dead languages can occur in two different forms: either the unparsable element is an isolated node on the phrase level, or it belongs to a complex phrase, along with other elements that are analyzable.

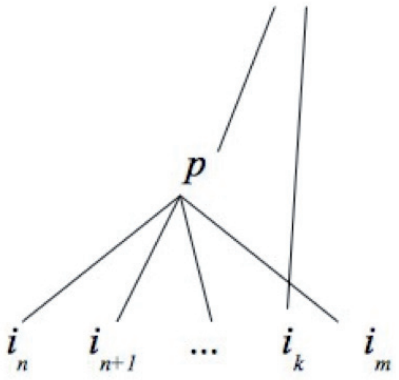
In the first case, the unparsable element can simply be assigned a specific tag – in a way similar to the <damage> XML tag proposed by Korkiakangas and Lassila (2013). A similar problem has also been discussed by Zemánek (2007), in the framework of a treebank of the ancient Semitic Ugaritic language.

When, on the contrary, the unparsable element(s) interrupt(s) a phrase, the problem can be seen as a special case of phrase discontinuity (in other words, it is formally identical to the case in which a dislocation or movement produces discontinous phrases).

3.1 Discontinuous phrases

Discontinuous phrases, both the "sprachwirklich" ones and the ones produced by an unparsable element, can be formally defined as fol-

low. Rephrasing the definition of yield *Y* of a node *p* given by Kallmeyer, Maier and Satta (2009; cf. Maier, 2011) as the set of all the indices $i \in \mathbb{N}$ such that *p* dominates the leaf labeled with the i^{th} terminal, one can generalize the definition of "discontinuous phrase" as follows. A phrase that is mapped at the node *p* with yield *Y* is a discontinuous phrase iff for $I_n \in Y \;\; \exists\, m > n$ such that $I_m \in Y$, $\exists\, k$ such that $n < k < m$. and $i_k \notin Y$.



Discontinuity can, in several cases, be solved employing iterations or recursive strategies; however, from the point of view of linguistic representation, this may, in given circumstances (such as trapping), interfere with the morpho-syntactic notation (nesting NPs will not always solve the problem of a discontinuous NP containing an extraneous element such as a preverb).

In the cases where nesting is not a valid option, using attribute indexing and pointers (Taylor, Marcus and Santorini 2003) in order to co-index the components of a phrase (for a formal definition of component see Kallmeyer, Maier and Satta, 2009) appears to be the best strategy available.

4 Conclusion

The creation of phrase-structure trees for ancient languages with structural peculiarities that make them very different from modern ones may require specific modifications to the usual parsing tagsets. Such modifications may occur both on the phrase and on the word levels. In order to minimize the challenges and maximize flexibility, a context-sensitive syntax with both labels and functional tags is more suitable than a rigid one; for instance functional markers for case inflection may apply to several different categories of labels (all nouns, adjectives and pronouns).

As far as discontinuous phrases are concerned, in the analysis of dead languages they may be natural linguistic phenomena, but they may also be the result of either poor text preservation or limited understanding of given segments. In order to avoid inaccurate nesting, a system of co-indexing appears to be the most advisable solution to guarantee a good degree of accuracy in the linguistic representation and a regular treatment of the linearity issues.

References

Anna Bauer. 2014. *Morphosyntax of the Noun Phrase in Hieroglyphic Luwian*, Brill, Leiden.

Paola Cotticelli-Kurras and Federico Giusfredi. 2015. *On Luwian Syntax: presentation of the SLUW project*, paper presented at the Arbeitstagung der Indogermanischen Gesellschaft, Marburg, 21 September 2015.

J. David Hawkins. 2000. *Corpus of Hieroglyphic Luwian Inscriptions, Volume I, Inscriptions of the Iron Age*. De Gruyter, Berlin/New York.

Harry A. Hoffner and H. Craig Melchert. 2008. *A Grammar of the Hittite Language*. Brill, Leiden.

Laura Kallmeyer, Wolfgang Maier and Giorgio Satta. 2009. *Synchronous rewriting in treebanks*. Proceedings of the 11th International Conference on Parsing Technologies. Paris: 69-72.

Timo Korkiakangas and Matti Lassila. 2013. *Abbreviations, fragmentary words, formulaic language: treebanking mediaeval charter material*. Proceedings of The Third Workshop on Annotation of Corpora for Research in the Humanities. Sofia.

KUB 35 = *Keilschrifturkunden aus Boghazköi*, Band 35, 1993. Gebr. Mann, Berlin.

Wolfgang Maier. 2011. *Characterizing Discontinuity in Constituent Treebanks.* Formal Grammar Lecture Notes in Computer Science Volume 5591: pp 167-182

Mitchell Marcus, Grace Kim, Mary Ann Mrcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Jatz, Britta Schasberger. 1994. *The Penn Treebank: Annotating Predicate Argument Structure*. University of Pennsylvania, Philadelphia.

H. Craig Melchert. 2003. *The Luwians*. Brill, Leiden.

Ann Taylor, Mitchell Marcus and Beatrice Santorini. 2003. *The Penn Treebank: An Overview.* University of York. Heslington, York.

Petr Zemánek. 2007. *A Treebank of Ugaritic. Annotating Fragmentary Attested Languages*. Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories. Bergen.

Linking dei contenuti multimediali tra ontologie multilingui: i verbi di azione tra IMAGACT e BabelNet

Lorenzo Gregori, Andrea Amelio Ravelli, Alessandro Panunzi
Università di Firenze
{lorenzo.gregori,alessandro.panunzi}@unifi.it,
aramelior@gmail.com

Abstract

English. We present a study dealing with the linking between two multilingual and multimedial resources, BabelNet and IMAGACT. The task aims to connect the videos contained in the IMAGACT Ontology of Actions and the related verb entries in BabelNet. The linking experiment is based on an algorithm that exploits the lexical information of the two resources. The results show that is possible to achieve an extensive linking between the two ontologies. This linking is highly desirable in order to build a rich multimedial knowledge base that can be exploited for the following complex tasks: the reference disambiguation and the automatic/assisted translation of both the verbs and the sentences which refer to actions.

Italiano. *Lo studio qui presentato riguarda il collegamento tra due risorse multilingui e multimediali, BabelNet e IMAGACT. In particolare, l'esperimento di linking ha come oggetto i video dell'ontologia dell'azione IMAGACT e le rispettive entrate lessicali verbali di BabelNet. Il task è stato eseguito attraverso un algoritmo che opera sulla base delle informazioni lessicali presenti nelle due risorse. I risultati del linking mostrano che è possibile effettuare un collegamento estensivo tra le due ontologie. Tale collegamento è auspicabile nel senso di fornire una base di dati ricca e multimediale per i complessi task di disambiguazione del riferimento dei verbi di azione e di traduzione automatica e assistita delle frasi che li contengono.*

1 Introduzione[1]

Le ontologie sono strumenti ampiamente utilizzati per rappresentare risorse linguistiche sul web e renderle sfruttabili da metodi di elaborazione automatica del linguaggio naturale. La disponibilità di linguaggi formali condivisi, come RDF e OWL, e lo sviluppo di ontologie di alto livello, come *lemon* (McCrae et al., 2011), stanno portando ad una metodologia unificata per la pubblicazione delle risorse linguistiche in forma di *open data*[2].

La rappresentazione delle informazioni attraverso le ontologie non è però sufficiente alla costruzione della rete semantica che è alla base dei nuovi paradigmi del web. L'interconnessione delle informazioni e, di conseguenza, il *mapping* e il *linking* tra ontologie diverse divengono aspetti essenziali per l'accesso alla conoscenza e per il suo arricchimento, come testimoniato dagli sviluppi sempre maggiori della ricerca in questo ambito (Otero-Cerdeira et al., 2015).

L'esigenza di massimizzare le connessioni tra risorse diverse si deve confrontare con il fatto che ogni ontologia è costruita con criteri differenti, che fanno capo a differenti quadri teorici. In questo contesto l'*instance matching* diventa particolarmente rilevante, poiché consente di collegare risorse senza mappare le entità ontologiche (Castano et al., 2008; Nath et al., 2014).

In questo articolo presentiamo un'ipotesi di collegamento tra due ontologie linguistiche, BabelNet (Navigli and Ponzetto, 2012a) e IMAGACT (Moneglia et al., 2014a), entrambe multimediali,

[1] Lorenzo Gregori ha scritto le sezioni 1 e 3 ed ha sviluppato gli algoritmi di *linking*; Andrea Amelio Ravelli ha scritto le sezioni 2 e 4 ed ha realizzato e valutato i dataset; Alessandro Panunzi ha supervisionato il lavoro e la scrittura dell'articolo. La ricerca è stata condotta nellambito del progetto MODELACT, programma Futuro in Ricerca 2012. Project code: RBFR12C6O8 2012-2015.

[2] In questo ambito è particolarmente rilevante l'iniziativa del Linguistic Linked Open Data Cloud (Chiarcos et al., 2011), che raccoglie e collega ontologie linguistiche in RDF e ad oggi contiene più di 500 risorse.

multilingui e sfruttabili per task di traduzione e disambiguazione (Moro and Navigli, 2015; Russo et al., 2013; Moneglia, 2014). Il collegamento tra le ontologie avviene attraverso la componente visuale di IMAGACT, ovvero la rappresentazione delle azioni per mezzo di scene prototipiche.

2 Risorse

2.1 IMAGACT

IMAGACT è un'ontologia dell'azione in cui le entità di riferimento sono identificate attraverso scene (video o animazioni 3D). I verbi che nelle varie lingue si riferiscono allo stesso concetto azionale sono collegati a una stessa scena, che rappresenta prototipicamente tale concetto. La raffigurazione visiva dell'azione veicola l'informazione a livello cross-linguistico: in questo modo si configura come uno strumento di disambiguazione del riferimento azionale, task particolarmente problematico date le differenti strategie di lessicalizzazione dell'azione adottate dalle lingue naturali.

All'interno del sistema di disambiguazione di IMAGACT, le scene prototipiche svolgono la funzione di interlingua. L'organizzazione dell'ontologia non è però regolata da principi lessicografici, ma da criteri di identificazione cognitiva dei concetti azionali categorizzati nelle diverse lingue. Di conseguenza, a partire da una scena è possibile ottenere i corrispettivi lemmi in tutte le lingue presenti in IMAGACT, con la certezza che tali lemmi siano adatti a descrivere l'azione di riferimento, e siano quindi intertraducibili quando applicati al contesto in oggetto (Panunzi et al., 2014).

La risorsa è stata creata a partire dall'analisi di corpora di parlato italiano e inglese attraverso giudizi di categorizzazione espressi da annotatori madrelingua. In questo modo sono stati associati oltre 500 lemmi verbali (per ciascuna lingua) alle stesse 1010 scene prototipiche. I video così ottenuti sono stati in seguito utilizzati per estendere l'ontologia ad altre lingue tramite giudizi di competenza da parte di annotatori madrelingua. Ad oggi, l'interfaccia web di IMAGACT[3] contiene dati completi su quattro lingue (italiano, inglese, spagnolo e cinese mandarino). Il database è in continua espansione (Moneglia et al., 2014b), e attualmente raccoglie dati da 18 lingue con diversi livelli di completezza (Tabella 1).

[3]http://www.imagact.it

Lingue	Verbi BN	Verbi IM
English	57.996	1.299
Spanish	16.832	735
Italian	15.590	1.100
Portuguese	11.517	792
German	5.210	992
Chinese	4.299	1.171
Norwegian	2.227	107
Danish	1.980	73
Polish	1.910	1.145
Hindi	342	189
Urdu	187	73
Bangla	117	120
Serbian	91	1.145
Sanskrit	35	198
Oriya	6	140
Totale	118.339	9.273

Tabella 1: Le 15 lingue comuni di BabelNet (BN) e IMAGACT (IM) con il relativo numero di verbi.

2.2 BabelNet

BabelNet[4] è una rete semantica multilingue, un dizionario enciclopedico strutturato in ontologia. Alla base della risorsa vi è la combinazione di due tra le più importanti basi di conoscenza, una linguistica e una enciclopedica, liberamente disponibili online: WordNet e Wikipedia. Si tratta, ad oggi, della più estesa risorsa multilingue per la disambiguazione semantica, giunta alla versione 3.0, che ricopre 271 lingue.

Informazione semantica e informazione enciclopedica sono state raccolte e collegate attraverso un algoritmo di mappatura automatica, al fine di creare un dizionario di concetti ed entità caratterizzato sia da ricchezza informativa, sia da una fitta rete di rapporti semantici a livello ontologico.

Concetti ed entità sono rappresentati per mezzo di BabelSynset (da qui in avanti, BS), estensione del concetto di *synset* utilizzato in WordNet. Un BS corrisponde a un concetto unitario a cui sono collegate le parole che nelle varie lingue vi si riferiscono, corredate da proprietà semantiche, una glossa ed esempi d'uso[5]. Un ulteriore contributo giunge dalla famiglia di risorse collegate a Wikipedia[6], attraverso cui è stato possibile collegare

[4]http://babelnet.org

[5]Tali collegamenti si basano anche su relazioni ereditate dai vari Multilingual WordNet(s).

[6]Wiktionary, OmegaWiki (versione allineata dei Wiktionary nelle singole lingue), Wikidata (database *document-*

ai BS le immagini archiviate in Wikimedia Commons e offrire così tale informazione a supporto della disambiguazione.

3 L'esperimento di *linking*

L'esperimento si colloca nello scenario più generale in cui viene realizzato un collegamento tra un'ontologia di dominio aperto, che raccoglie un'ampia serie di concetti eterogenei poco specificati, e una specialistica, dove i concetti di un singolo dominio sono rappresentati in modo più dettagliato (Magnini and Speranza, 2002).

Il nostro esperimento sfrutta i verbi equivalenti in traduzione nelle due risorse per effettuare il *linking* tra le ontologie attraverso le scene prototipali presenti in IMAGACT. Il criterio utilizzato per trovare le scene che possono rappresentare i concetti azionali di BabelNet è quello della maggior corrispondenza tra i verbi collegati allo stesso BS e quelli collegati alle scene di IMAGACT [7].

Un diverso tentativo di *mapping* su IMAGACT è stato condotto da De Felice et al. (2014) attraverso l'analisi della collegabilità tra i tipi azionali di IMAGACT e i *synset* di WordNet (per l'inglese) e ItalWordNet (per l'italiano). Anche in quell'esperimento è stato sfruttato l'insieme dei verbi comuni come indice di similarità tra gli oggetti delle diverse ontologie. L'ambito di applicazione del nostro esperimento è però diverso, poiché è allargato ad un contesto multilingue: prima di tutto non si tratta di un *mapping* tra concetti delle due ontologie, ma di un *linking* tra istanze di tipo diverso; in secondo luogo non si usa WordNet, ma BabelNet e, coerentemente, il dato di IMAGACT che viene sfruttato non corrisponde ai tipi (che dipendono dai diversi lemmi delle varie lingue) ma alle scene (oggetti di riferimento interlinguistico).

3.1 Il dataset

L'esperimento che presentiamo è stato condotto su un dataset di riferimento annotato manualmente e composto da 25 scene di IMAGACT e 30 BS; per ciascuna coppia ⟨scena,BS⟩ è stato giudicato se la scena fosse o meno adatta a rappresentare il BS. Il dataset, comprendente 750 giudizi binari (25 scene per 30 BS), è stato utilizzato per la valutazione degli algoritmi automatici[8]. Per selezionare le scene sono stati considerati i prototipi azionali della variazione di 7 verbi inglesi (*put*, *move*, *take*, *insert*, *press*, *give* e *strike*), che proiettano 152 BS totali; di questi, ne sono stati scelti 25. La selezione è stata fatta in modo casuale, ma ha tenuto conto del fatto che non tutte le scene sono collegate allo stesso numero di verbi, sia perché le lingue di IMAGACT non sono egualmente rappresentate (vedi tabella 1), sia perché è diverso il numero di verbi utilizzabili per riferirsi alle azioni nelle diverse lingue. Per creare un *test set* che fosse un campionamento rappresentativo, abbiamo cercato di preservare queste differenze quantitative inserendo scene con un diverso numero di verbi collegati (da un minimo di 7 a un massimo di 18) in modo proporzionale all'intero set di scene di IMAGACT. Anche il numero di verbi contenuti in un BS è molto variabile, per cui la loro selezione ha seguito un criterio simile: ognuno dei 30 BS del dataset ha da un minimo di 4 ad un massimo di 51 verbi collegati. Il dataset pubblicato è derivato da un *agreement*: sono stati utilizzati tre annotatori ed il giudizio inserito è quello di almeno due annotatori su tre. L'*inter-rater agreement* riporta una *k* di Fleiss pari a 0,76.

3.2 Gli algoritmi

L'algoritmo base (Algoritmo 1) utilizzato per il *linking* si avvale di una funzione che calcola la vicinanza tra una scena e un BS misurando la frequenza con cui i verbi collegati alla scena di IMAGACT sono legati anche al BS. L'insieme dei candidati è composto da tutti i BS che sono concetti possibili per ogni verbo collegato alla scena.

Algoritmo 1 Algoritmo base

1: s: scena in ingresso
2: V: set di verbi collegati a s in IMAGACT
3: List⟨BabelSynset⟩ LS: lista vuota
4: per ogni v_i in V
5: List⟨BabelSynset⟩ Syn = lista di BS collegati a v_i
6: aggiungi Syn a LS
7: List⟨BabelSynset⟩ $FLS = freqList(LS)$
8: Collega s ai primi n BabelSynset di FLS

La funzione *freqList* calcola la lista di frequenza dei BS in *LS* e li ordina dal più frequente al meno frequente.

A partire da questa versione di base dell'algoritmo di *linking*, è stata implementata una versione migliorata che sfrutta la rete semantica di BabelNet, in modo da includere nell'analisi anche i BS semanticamente vicini. Anziché estrarre gli equivalenti in traduzione soltanto dai BS collegati di-

oriented che contiene risorse multimediali).

[7] Per il test è stata utilizzata la versione 3.0 di BabelNet; i dati sono stati estratti utilizzando l'API Java (Navigli and Ponzetto, 2012b)

[8] Il dataset è disponibile alla pagina http://bit.ly/1MtZqB9

rettamente ai verbi, vengono considerati anche i verbi appartenenti a *synset* collegati al BS principale tramite le relazioni di BabelNet fino a un certo livello di profondità. Per pesare in modo differenziato i BS collegati al verbo abbiamo utilizzato una funzione ricorsiva w, definita nel modo seguente.

Dato S l'insieme dei BS, $s_0 \in S$ collegato direttamente al verbo e $s', s'' \in S$ collegati tra loro da una relazione $r \in R$, definiamo una funzione $w : S \rightarrow [0,1]$ tale che $w(s_0) = freq(s_0)$ e $w(s'') = w(s') \cdot c \cdot p(r)$, dove: $freq$ calcola la frequenza del BS così come riportato nell'algoritmo base; R è l'insieme delle relazioni tra BS verbali; $p(rel) : R \rightarrow [0,1]$ è una funzione che assegna un peso ad ogni relazione R; $c \in [0,1]$ è un coefficiente di riduzione di peso all'aumentare della distanza dal nodo centrale.

Questa metrica consente di differenziare l'importanza delle diverse relazioni semantiche di BabelNet. Abbiamo infatti verificato che, mentre alcune di esse sono molto rilevanti per il task, altre importano informazione non pertinente e devono quindi essere escluse. La tabella seguente mostra l'elenco delle relazioni tra BS verbali con il relativo valore di rilevanza misurato con *information gain* sul dataset annotato.

Relazioni in BabelNet	**Valore *IG***
Hyponym	0.135
Also See	0.050
Hypernym	0.041
Verb Group	0.039
Entailment	0.009
Gloss Related	0.000
Antonym	0.000
Cause	0.000

Tabella 2: Relazioni tra BS verbali.

4 Valutazione

4.1 Risultati della valutazione

I due algoritmi sono stati eseguiti sulle 25 scene del dataset, quello di base e quello migliorato (considerando soltanto un livello di profondità). È stata quindi verificata l'aderenza dei primi n BS estratti dagli algoritmi alle scene. La tabella 3 riporta la sintesi dei risultati ottenuti[9].

Per entrambi gli algoritmi il primo BS candidato per il *linking* è sempre corretto. I risultati

	Alg. base	**Alg. migliorato**
% corr. $(n = 1)$	100%	100%
% corr. $(n = 2)$	84%	88%
% corr. $(n = 3)$	76%	83%

Tabella 3: Percentuale di assegnazioni corrette di scene a BS con i due algoritmi e al variare di n.

peggiorano progressivamente al crescere di n e, parallelamente, aumenta anche il divario qualitativo tra i due algoritmi. Inoltre, il dataset annotato è stato utilizzato come *training set* per due ulteriori algoritmi di *machine learning*, uno che considera soltanto *features* relative ai BS collegati in modo diretto, l'altro che include anche *features* dei BS collegati indirettamente attraverso relazioni semantiche. Il risultato[10] riporta rispettivamente 81,16% e 86,95% di assegnamenti corretti di scene al BS. Benché il *test set* sia troppo piccolo per avere una stima precisa sull'efficacia, i risultati sono incoraggianti e compatibili con quelli ottenuti dai due algoritmi (semplice e migliorato) eseguiti sul dataset. La differenza tra le due percentuali mostra chiaramente che l'utilizzo dei BS vicini è significativo per questo task.

4.2 Conclusioni

Benché non sia ancora stato fatto un *fine-tuning* dei parametri (per il quale è necessario un *test set* più ampio), i buoni risultati ottenuti da questo esperimento aprono la possibilità di collegare le due ontologie attraverso le scene di IMAGACT, al fine di arricchire entrambe le risorse. Da un lato, i video di IMAGACT potrebbero rappresentare i concetti azionali dei BS; dall'altro, IMAGACT verrebbe arricchita con l'informazione di traduzione presente in BabelNet.

Inoltre, dall'osservazione di Babelfy (Moro et al., 2014), motore di *word sense disambiguation* e *entity linking* derivato da BabelNet, è apparso evidente che l'ipotesi di *linking* qui proposta avrebbe un notevole impatto sull'espressività della rappresentazione visuale delle frasi, con l'associazione di immagini a nomi e di video a verbi.

Infine, è importante notare che sia BabelNet sia IMAGACT sono risorse in espansione: poiché gli algoritmi sfruttano gli equivalenti in traduzione, i risultati potranno essere via via più precisi all'aumentare delle lingue e dei lemmi considerati.

[9] Per i risultati completi consultare http://bit.ly/1MtZqB9

[10] L'algoritmo utilizzato è SVM con *kernel* lineare e la valutazione è fatta con *10-folds cross-validation*.

References

S. Castano, A. Ferrara, D. Lorusso, and S. Montanelli. 2008. On the ontology instance matching problem. In *Database and Expert Systems Application, 2008. DEXA '08. 19th International Workshop on*, pages 180–184, Sept.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, pages 245–275.

Irene De Felice, Roberto Bartolini, Irene Russo, Valeria Quochi, and Monica Monachini. 2014. Evaluating ImagAct-WordNet mapping for English and Italian through videos. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume I, pages 128–131. Pisa University Press.

Bernardo Magnini and Manuela Speranza. 2002. Merging Global and Specialized Linguistic Ontologies. In *Proceedings of the Workshop Ontolex-2002 Ontologies and Lexical Knowledge Bases, LREC-2002*, pages 43–48.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications - Volume Part I*, ESWC'11, pages 245–259, Berlin, Heidelberg. Springer-Verlag.

Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014a. The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Massimo Moneglia, Susan Brown, Aniruddha Kar, Anand Kumar, Atul Kumar Ojha, Heliana Mello, Niharika, Girish Nath Jha, Bhaskar Ray, and Annu Sharma. 2014b. Mapping Indian Languages onto the IMAGACT Visual Ontology of Action. In Girish Nath Jha, Kalika Bali, Sobha L, and Esha Banerjee, editors, *Proceedings of WILDRE2 - 2nd Workshop on Indian Language Data: Resources and Evaluation at LREC'14*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Massimo Moneglia. 2014. Natural Language Ontology of Action: A Gap with Huge Consequences for Natural Language Understanding and Machine Translation. In Zygmunt Vetulani and Joseph Mariani, editors, *Human Language Technology Challenges for Computer Science and Linguistics*, volume 8387 of *Lecture Notes in Computer Science*, pages 379–395. Springer International Publishing.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June. Association for Computational Linguistics.

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.

Rudra Nath, Hanif Seddiqui, and Masaki Aono. 2014. An efficient and scalable approach for ontology instance matching. *Journal of Computers*, 9(8).

Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli and Simone Paolo Ponzetto. 2012b. Multilingual WSD with just a few lines of code: the BabelNet API. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Korea.

Lorena Otero-Cerdeira, Francisco J. Rodrguez-Martnez, and Alma Gmez-Rodrguez. 2015. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949 – 971.

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.

Irene Russo, Francesca Frontini, Irene De Felice, Fahad Khan, and Monica Monachini. 2013. Disambiguation of Basic Action Types through Nouns Telic Qualia. In Roser Saur, Nicoletta Calzolari, Chu-Ren Huang, Alessandro Lenci, Monica Monachini, and James Pustejovsky, editors, *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon. Generative Lexicon and Distributional Semantics*, pages 70–75.

New wine in old wineskins: a morphology-based approach to translate medical terminology

Raffaele Guarasci, Alessandro Maisto
Department of Political, Social and Communication Sciences
University of Salerno
Via Giovanni Paolo II, 132,
84084 Fisciano (SA)
{rguarasci,amaisto}@unisa.it

Abstract

English. In this work we introduce the first steps toward the development of a machine translation system for medical terminology. We explore the possibility of basing a machine translation task in the medical domain on morphology. Starting from neoclassical formative elements, or *confixes*, we started building MedIta, a cross-language ontology of medical morphemes, aiming to offer a standardized medical consistent resource that includes distributional and semantic information of medical morphemes. Using this information, we have built an ontology-driven Italian-English machine translation prototype, based on a set of Finite State Transducers, and we have carried out an experiment on *Orphanet* medical corpus to evaluate the feasibility of this approach.

Italiano. *In questo lavoro si introduce lo sviluppo di un sistema per la traduzione automatica della terminologia medica. Si propone un approcio morfologico, che utilizza gli elementi formativi neoclassici, i* confissi*. Si introduce* MedIta*, un'ontologia multilingua di morfemi del dominio medico, che mira ad offrire una risorsa validata secondo gli standard medici e che contiene informazioni semantiche e statistiche. La fattibilit della risorsa viene valutata tramite un prototipo di sistema di traduzione italiano-inglese basato su Trasduttori a Stati Finiti.L'applicazione viene poi testata su un campione estratto dal corpus medico* Orphanet.

1 Introduction

Automating Machine Translation (MT) of a technical language is a challenging task that requires an in-depth analysis both from a linguistic point of view and as regards the implementation of a complex system. This becomes even more complex in medical language. Indeed the translation of medical terminology must always be validated by a domain expert following official classification standards. For this reason currently there are no translation support tools specifically created for the medical domain. In this work we propose an MT system based on a set of Finite State Transducers that uses cross-language morpheme information provided by a lexical resource. The underlying idea is that in a technical language a morpho-semantic approach (Dujols et al., 1991) may be more effective than a probabilistic one in term-by-term translation tasks. Even though our approach could seem a bit "old fashioned", we must consider that proper nature of medical language, fully based by morphemes derived from neoclassical formative elements (Thornton, 2005). Neoclassical formative elements are morphological elements that come into being from Latin and Greek words, they combine with each other following compositional morphology rules. Due to the heterogeneous nature of these elements, they have received different definitions, we prefer to use the term *confixes*, a morpheme with full semantic value, which has been predominantly used in the literature (Sgroi, 2003; D'Achille, 2003; De Mauro, 2003). In this work we focused only on word formation related to the medical domain.

2 Related Work

In the following section we briefly present the most relevant studies or applications regarding the use of a morpho-semantic approach, and studies that exploited morphological rules in machine

translation tasks. Morpho-semantic approaches have already been applied to the medical domain in many languages. Works that deserve to be mentioned are those by (Lovis et al., 1998) that identified the ICD[1] (International Classification Diseases) codes in diagnoses written in different languages; (Hahn et al., 2001) that segmented the subwords in order to recognise and extract medical documents; and (Grabar and Zweigenbaum, 2000) that used machine learning methods on the morphological data of the SNOWMED[2] nomenaclature (French, Russian, English). As regards morphological approaches in machine translation tasks, we mention a lexical morphology based Italian-French MT tool (Cartoni, 2009); MT models for morphologically rich languages, like Russian and Arabic (Toutanova et al., 2008; Minkov et al., 2007), a German-English biomedical terms MT tool (Daumke et al., 2006) and an approach based on finite state technologies (Amtrup, 2003). Furthermore we notice an unsupervised morph-tokens analysis applied to MT tasks (Virpioja et al., 2007) and an approach that applies morphological analysis to statistical MT systems (Lee, 2004).

3 Proposed approach

The proposed approach can be divided in two main phases:

- the creation of a lexical resource: an ontology of morphemes belonging to the medical domain to be used as a knowledge base. This ontology represents medical morphemes and provide both semantic and statistical (e.g. distributional profiles) information about them.
- the implementation of a MT prototype that exploits information provided by this lexical resource to perform an effective medical term translation.

Currently tested languages are Italian and English, but one of the advantages of the morpho-semantic method is that linguistic analyses designed for a language can often be transferred to other languages that share the common basis of neoclassical formative elements (Deléger et al., 2007).

[1] http://www.cdc.gov/nchs/icd/icd10cm.htm

[2] http://www.ihtsdo.org/snomed-ct

3.1 Medical morphemes ontology (MedITA)

Our starting point is an ontology of medical morphemes (prefixes, suffixes and confixes), that includes various kinds of information for each morpheme, like distributional profiles extracted from medical corpora, medical classifications and definitions. This resource is made possible by the formative elements underlying medical terms: morphemes may detect and describe the semantic relations existing between those words that share portions of meaning. Relying on words sharing morphemes endowed with a particular meaning (e.g. *-acusia*, hearing disorders) it is not difficult to find sets of near-synonyms (Namer, 2005). Moreover, we can infer the medical subdomain to which the synonym set belongs (e.g."*otolaryngology*") and we can differentiate any item of the set by exploiting the meaning of the other morphemes involved in the words.

- **synset**: *iper-acusia*, *ipo-acusia*, *presbi-acusia*, *dipl-acusia*;
- **subdomain**: *-acusia* "otolaryngology";
- **description**: *ipo-* "lack", *iper-* "excess", *presbi-* "old age", *diplo-* "double".

On the basis of the morphemes meaning, we can also infer relations between words that are not morphologically related, but which are composed of morphemes that share at least one semantic feature and/or the medical subdomain (see Table 1). This is made possible using formative elements, that do not represent mere terminations, but possess their own semantic self-sufficiency (Iacobini, 2004).

Related to	Morpheme	Subdomain
Tumors	*cancero-*, *carcino-*,	oncology
Stomach	*stomac-*, *gastro-*	gastroenterology
Skin fungus	*fung-*, *miceto-*, *mico-*	dermatology

Table 1: Morphemes that share semantic features

To start building the ontology we used a top-down approach: first of all we have divided the medical specialties into 22 categories (e.g. "internal medicine", "cardiology", "traumatology", etc...), with the support of a domain expert. The lexical resource used as source is the electronic version of the GRADIT[3] (De Mauro, 1999). Using

[3] Electronic version of *Grande Dizionario Italiano dell'Uso*

the GRADIT it has been possible to extract every kind of morpheme related to the medical domain and group them on the basis of their subdomains. Each morpheme has been compared with the morphemes included in the Open Dictionary of English[4]. The respective English translation has been manually added to each element. The resulting set of medical morphemes have been formalized into a resource that specifies their category:

- Confixes (*cfx*): neoclassical formative elements with a full semantic value (i.e. *pupillo-*, *mammo-*, *-cefalia*);
- Prefixes (*pfx*): morphemes in the first part of the word, able to connote it with a specific meaning (i.e. *-ipo*, *-iper*);
- Suffixes (*sfx*): morphemes in the final part of the word, able to connote it with a specific meaning (i.e. *-oma*, *-ite*);

Subsequently a set of semantic information has been added to every morpheme. These semantic labels provide descriptions about the meaning they confer to the words composed with them and information about morpheme classification. Such semantic information regards the three following aspects:

- Meaning: describes the specific meaning of the morpheme;
- Medical Class: gives information regarding the medical subdomain to which the morpheme belongs;
- Translation: presents the corresponding morpheme in the English language.

3.2 MT System

We built a Morphology-based Machine Translation prototype that works in two steps. The system is composed of a set of Finite State Automata to find approximate morpheme matching and a set of Finite-State Transducers[5] able to translate the Italian term into the English one. In the first step a partial matching to recognize Italian medical terms from text was performed, after that each recognized morpheme that composes the word was tagged with semantic information. To maximize the morphological recognition with minimum effort a set of patterns able to recognize different sequences of morphemes are identified (e.g. : *cfx-cfx*; *cfxs-sfx*; etc.) These patterns are derived from distributional profiles of morphemes: the most frequent compositions of morphemes extracted from a sample of 1000 words from ICD-10 for Italian and UMLS[6] (Unified Medical Language System) for English. A new category named *cfxs* is needed to reduce systematic kinds of errors in specific cases. *cfxs* identifies all the confixes that can appear before a suffix, with its correspondent English morpheme deprived of the final part, to avoid repetition in case of suffixation (i.e. *cystoitis*, *cfx-sfx*, is not valid, but *cystitis*, *cfxs-sfx* is valid). After that, the Transducer takes as input the morphemes and produces the corresponding translations. In the end, using the same morpheme sequences, it tags every Italian Medical Term with the respective English translation.

4 Experiment and Evaluation

To evaluate the approach described above and to assess its feasibility, we built a test dataset: a corpus of terms extracted from the Italian version of Orphanet[7], a resource that provides an inventory of more than 6000 rare diseases and a classification of diseases elaborated using existing published expert classifications. Orphanet has been chosen because the vast majority of rare diseases are composed of several morphemes (e.g. *hemimegalencephaly*, *acrocephalopolydactyly*). For each disease, Orphanet offers a brief summary with connections with other medical terminologies (MeSH[8], UMLS, MedDRA[9]) or standard classifications (ICD-10). In this early stage in order to test the performance of our morpho-semantic translator we evaluated the Precision score on a sample of 100 rare diseases extracted from Orphanet corpus. The "gold standard" taken into account is the translation provided from ICD-10. Our results were compared to those obtained using other MT systems widely used in recent years as a case study:

- **Google Translate**[10], the wildly popular MT service provided by Google. It uses a propri-

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

[5] ASF and TSF are built using OpenFST Library (avaible at http://openfst.org/, in particular the python wrapper PyFst http://pyfst.github.io/

[6] http://www.nlm.nih.gov/research/umls/

[7] http://orpha.net/

[8] https://www.nlm.nih.gov/mesh/

[9] http://www.meddra.org/

[10] https://translate.google.com

etary statistical machine translation technology.

- **BabelNet**[11](Navigli and Ponzetto, 2010), a multilingual semantic network and ontology obtained as an integration of WordNet and Wikipedia.
- **HeTOP**[12](Grosjean et al., 2013), a controlled vocabulary that combines the best known biomedical terminology, vocabularies and classifications. It also integrates UMLS.

MT System	Precision
MedITA	91%
Google Translate	85%
BabelNet	73%
HeTOP	68%

Table 2: Precision comparison on Orphanet corpus

Although it must be considered that the system is based on an incomplete resource still in development and the test sample is quite small, this first analysis shows interesting results (see Table 2). In particular, a qualitative analysis of the results reveals some important aspects that deserve a deeper analysis. A brief summary and explanation of the most relevant aspects deriving from the Orphanet translation follows:

- On rare diseases the system has a precision higher than other systems, perhaps due to the intrinsic properties of the medical language, most evident in the case of rare diseases, as mentioned above. Notice that - in some cases - Google Translate and BabelNet provide a translation using a broader term (e.g. Google it: *"acromatopsia"* - en: *"colorblindess"*; it: *"iperargininemia"* - en: *"argininemia"*). Although in a broader context these translations could be considered as valid, in an extremely specific domain such as the medical one they are *de-facto* errors.
- In several cases the system proposes a translation that does not fit exactly with the standard: e.g. *polyendocrinopathia/polyendocrinopathy*. Many proposed translations can be considered *acceptable* because, although they are not yet formalized in the standard, they occur in other available resources, like technical papers, web pages, etc.
- The system never fails when other MT systems are wrong (see Table 3). This occurs with complex and extremely rare words; in these "extreme" cases we can argue that a morphological based translation could be better than a probabilistic one.

Another relevant aspect is that the system can work as spellchecker. This is a "side effect" of a morphological approach, despite that it may prove a useful function to improve precision, especially if it works on raw or uncontrolled data.

5 Conclusions

In this work we presented a morphology-based machine translation prototype specifically suited for medical terminology. The prototype uses ontologies of morphemes and Finite State Transducers. Even though the approach may seem a little out-of-date, the preliminary results showed that it can work as well as a probabilistic system in such a specific domain. It is worth mentioning that at this early stage we tested the prototype only on samples, since the evaluation is an extremely time-consuming task: every translated term must be manually compared with one or more medical standards. Medical standards are often not aligned, therefore an Orpha-number (disease id) does not necessarily match a disease listed in ICD-10. Moreover, these resources are not easily usable in an automated way, therefore the evaluation should entirely be done manually. Finally, even if at this preliminary stage there are many open issues, but the encouraging results suggest possible future developments: morpho-semantic approach, allows to easily extend the system to other languages; we can enrich the ontology to cover a bigger number of morphemes and we can take into account complex multiword expressions. A possible application of the system could be in the context of cross-border healthcare services in the European Union (Directive 2011/24/EU on patients' rights in cross-border healthcare)[13] and as a translation support tool for the international systems of coding diagnoses and disability (ICD and ICF[14]).

[11] http://babelnet.org/
[12] http://www.hetop.eu/hetop/
[13] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:088:0045:0065:EN:PDF
[14] http://www.who.int/classifications/icf/en/

Orphanet (it)	ICD-10 (en)	MedITA	Google	BabelNet	HeTOP
iperlisinemia	hyperlysinemia	✓	✗	✗	broader term
acrocefalopolisindattilia	acrocephalopolysyndactyly	✓	✓	✗	✗
polimicrogiria	Polymicrogyria	✓	✓	✓	✗
anisachiasi	anisakiasis	✓	✗	✗	✗
balantidiasi	balantidiasis	✓	✗	✗	✓
difillobotriasi	diphyllobothriasis	✓	✗	✗	✓
emimegalencefalia	hemimegalencephaly	✓	✗	✗	✓
poliembrioma	polyembryoma	✓	✗	✗	✗

Table 3: Translation comparison on rare diseases

References

Jan W Amtrup. 2003. Morphology in machine translation systems: Efficient integration of finite state transducers and feature structure descriptions. *Machine Translation*, 18(3):217–238.

Bruno Cartoni. 2009. Lexical morphology in machine translation: A feasibility study. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–138. Association for Computational Linguistics.

P. D'Achille. 2003. *L'italiano contemporaneo*. Il Mulino.

Philipp Daumke, Stefan Schulz, and Kornél Markó. 2006. Subword approach for acquiring and cross-linking multilingual specialized lexicons. *Programme Committee*, 1.

Tullio De Mauro. 1999. *Grande Dizionario Italiano dell'Uso*, volume 8. UTET.

Tullio De Mauro. 2003. *Nuove Parole Italiane dell'uso*, volume 7 of *GRADIT*. UTET.

Louise Deléger, Fiammetta Naner, Pierre Zweigenbaum, et al. 2007. Defining medical words: Transposing morphosemantic analysis from french to english. *Studies in Health Technology and Informatics*, pages 535–539.

Pierre Dujols, Pierre Aubas, Christian Baylon, and François Grémy. 1991. Morpho-semantic analysis and translation of medical compound terms. *Methods of Information in Medicine*, 30(1):30.

Natalia Grabar and Pierre Zweigenbaum. 2000. Automatic acquisition of domain-specific morphological resources from thesauri. In *Proceedings of RIAO*, pages 765–784. Citeseer.

Julien Grosjean, Tayeb Merabti, Lina F Soualmia, Catherine Letord, Jean Charlet, Peter N Robinson, Stéfan J Darmoni, et al. 2013. Integrating the human phenotype ontology into hetop terminology-ontology server. *Studies in health technology and informatics*, 192.

Udo Hahn, Martin Honeck, Michael Piotrowski, and Stefan Schulz. 2001. Subword segmentation–leveling out morphological variations for medical document retrieval. In *Proceedings of the AMIA Symposium*, page 229. American Medical Informatics Association.

Claudio Iacobini. 2004. Composizione con elementi neoclassici. In M. Grossmann & F. Rainer, editor, *La formazione delle parole in italiano*, pages 69–95. Niemeyer, T'ubingen.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 57–60. Association for Computational Linguistics.

Christian Lovis, Robert Baud, Anne-Marie Rassinoux, Pierre-André Michel, and Jean-Raoul Scherrer. 1998. Medical dictionaries for patient encoding systems: a methodology. *Artificial intelligence in medicine*, 14(1):201–214.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *ACL*, volume 7, pages 128–135.

Fiammetta Namer. 2005. Acquisizione automatica di semantica lessicale in francese: il sistema di trattamento computazionale della formazione delle parole dérif. In Anna Maria Thornton et Maria Grossmann, editor, *Atti del XXVII Congresso internazionale di studi Società di Linguistica Italiana: La Formazione delle parole*, pages 369–388.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

S. C. Sgroi. 2003. Per una ridenizione di "confisso": composti confissati, derivati confissati, parasintetici confissati vs etimi ibridi e incongrui. *Quaderni di semantica*, 24:81–153.

A. M. Thornton. 2005. *Morfologia*. Carocci.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *ACL*, pages 514–522.

Sami Virpioja, Jaakko J Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, 2007:491–498.

Computing, memory and writing: some reflections on an early experiment in digital literary studies

Giorgio Guzzetta[1], **Federico Nanni**[2]
[1] Italian Department, University College Cork
[2] Department of Philosophy and Communication Studies, University of Bologna
guzzettg@gmail.com, federico.nanni8@unibo.it

Abstract

English. In this paper we present the first steps of a research that aims at investigating the possible relationship between the emergence of a discontinuity in the study of poetic influence and some early experiments of humanities computing. The background idea is that the evolution of computing in the 1960s and 1970s might have triggered a transformation of the notion of influence and that today's interactions between the field of natural language processing and digital humanities are still sustaining it.

In order to do so, we studied a specific interdisciplinary project dedicated to the topic and we reproduced those experiments. Then we compared the results with different text mining techniques in order to understand how contemporary methods can deal with the rethinking of the notion of influence.

Italiano. *In questo articolo presentiamo i primi passi di una ricerca che intende investigare il possibile legame tra l'emergere di una discontinuità‡ nello studio dell'influenza poetica e alcuni tra i primi esperimenti di informatica umanistica. La nostra idea è che l'evoluzione dell'informatica tra gli anni Sessanta e Settanta abbia avviato tale trasformazione nel concetto di influenza e che le contemporanee interazioni tra i campi di natural language processing e digital humanities la stiano ancora sostenendo. Di conseguenza abbiamo deciso di studiare uno specifico progetto interdisciplinare dedicato all'argomento e abbiamo riprodotto gli esperimenti descritti nell'articolo. Quindi abbiamo confrontato i risultati ottenuti con quelli ricavati applicando contemporanei approcci di text mining per capire come tali metodi possano permetterci di comprendere più approfonditamente il ripensamento della nozione di influenza.*

1 Introduction

In recent years, when the institutional presence of digital humanities grew stronger, the need for a closer look at the history of humanities computing (the name under which this community of researchers was originally gathered) became more urgent, and some answers have been attempted (Nyhan et al., 2012; Sinclair and Rockwell, 2014). Rehearsing the history of humanities computing proved to be a challenging task, because of the hybrid and interdisciplinary nature of it and because of its entanglement up with a field like computing, which is epistemologically unstable and interdisciplinary as well (Castano et al., 2009).

This paper is a contribution to this attempt to develop an history of humanities computing, trying to combine together the histories of computing, of computational linguistics, and of literary studies. We present here the first steps of a research that aims at investigating the possibility that the emergence of a "discontinuity" in the notion of literary influence (and, consequently, of literary source), or at least the critical awareness of it, might be related to some early experiments of humanities computing. The background idea is that the evolution of computing in the 1960s and 1970s might have contributed to this transformation of the notion of influence, and that today's interactions between the field of natural language processing and digital humanities might have further developed it, possibly in new directions.
In order to do so, the paper is organised as

follows. First of all, we define the problem of influence from the point of view of literary studies and literary theory. With that background in mind, we study a specific interdisciplinary project dedicated to the topic (Raben, 1965) and we analysed Raben key-role on Goodman and Villani (1965) computational analysis. Moreover, to get a complete understanding of the methods applied in this work, we recreated the same approach with a script in Python.
Then, we decided to update Goodman and Villani's approach by adopting contemporary text mining methods. While conducting this specific task our purpose was not only to compare the different techniques and to highlight possible mistakes in Goodman and Villani's approach, but we aimed especially at understanding whether contemporary methods commonly adopted in natural language processing and digital humanities are able to answer questions and solve problematic issues emerged during the rethinking of the notion of influence in recent years.

2 The problem of literary influence

In the 1970s Harold Bloom "developed a theory of poetry by way of a description of poetic influence, or the story of the intra-poetic relationships" (Bloom, 1997), introducing the notion of "anxiety of influence" as a feature of creative writing. After him, critics used the term influence "to designate the affiliative relations between past and present literary texts and/or their authors" (Renza, 1990). That poets were influenced by previous masters of the craft is not a new idea, in fact it is as old as poetry itself. What is new is Bloom's aim to "de-idealize our accepted accounts of how one poets helps another" (Bloom, 1997). The traditional humanistic notion of infra-poetic relations "performed a conservative cultural function" because critical assessment of it was "focusing on the ways literary works necessarily comprise revision or updating of their textual antecedents", emphasising "homogeneity" and "continuity" in Western (canonical) literature:

Grounded in nineteenth-century philological notions of historical scholarship [...] this tradition-bound position regards literary influence as a benign, even reverential, endorsement of humanism: the ongoing project to transform the world into the image and likeness of human beings (Renza, 1990).

The past was not at all a burden in this humanistic tradition. Since the 1970s, though, thanks mainly to *The Anxiety of Influence* (Bloom, 1973), it became one. A few years earlier, in 1970 - but we have to consider that the first draft of Bloom's *Anxiety* was written in 1967 - Walter Jackson Bate, whom Renza considers Bloom's main critical precursor, introduced a discontinuity in this traditional view of influence, discussing the burden of the past in English poets:

[Bate] concedes that the mimetic view of influence pertains to the major portion of Western literary history. Only in the eighteenth century does the poets first suffer "the burden of the past"; only then does he experience a "loss of self-confidence" about what to write and how to write it "as he compares what he feels able to do with the rich heritage of past art and literature" (Renza, 1990).

3 Trasferring of memory and language: Raben's project

In 1964, during one of the first conferences announcing the coming of age of Digital Literary Studies, the *Literary Data Processing Conference*, sponsored by IBM, Joseph Raben, founder and main editor of *Computers and the Humanities* in 1966, described an interdisciplinary project in which computers played a significant role. Raben's project is interesting because it introduced some differences from a tradition that by then was consolidated (even though far from being a mainstream literary trend), involving the use of statistical methods to define and describe the style of an author. The two main elements of this tradition, that dominated the first experiments in literary humanities computing, are concordances on one side and authorship attribution based on stylometry on the other. Both were developed between late nineteenth/early twentieth century, and in both we can notice an essentialist idea of style, sort of a writer's trademark quantifiable from a linguistic point of view. With this experiment they were not trying to define the character of an author, a notion still essentialist and stiff; instead, they were

more interested in the agency of the writer that reuse the material at his disposal, transforming and manipulating it for his own purposes. The identity of the writer is not therefore an essence, but an active process of rewriting and remediation.

3.1 A specific case study

In his paper Raben described a specific case study based on the automatic detection of similarities between Milton's *Paradise Lost* and Shelley's *Prometheus Unbound*. The aim of Raben was "to illuminate the relationship of Shelley and Milton". To do this, he discussed Shelley's ideas on poetry and imitation, which were both considered "mimetic art":

It creates, but it creates by combination and representation. Poetical abstractions are beautiful and new, not because the portions of which they are composed had no previous existence in the mind of man or in nature, but because the whole produced by their combination has some intelleg-ible and beautiful analogy with those sources of emotion and thought, and with the contemporary condition of them: one great poet is a masterpiece of nature which another not only ought to study but must study. (Raben, 1965)

Combination of linguistic expressions previ-ously used by great poets was indeed part of the creative writing process. Shelley himself was doing that in his own creation, as Raben tried to demonstrate using computational techniques to confront his work with that of Milton:

Shelley is not merely echoing Milton's lan-guage: he is absorbing the sense of Milton's lines, transmuting it and re-expressing it in some of the same words in which it was originally invested. (Raben, 1965)

The ambition was that of entering in the (re)writing process as a creative endeavour, giv-ing a very detailed description of the way in which Shelley reused and reshaped images and vocabu-lary from Milton.

4 Detecting similarities between two texts

The method adopted in order to detect textual similarities between *Promethus Unbound* and *Paradise Lost* has been described by Goodman and Villani (1965), the two researchers who helped Raben in his work.

The authors described in detail the different steps of their analysis. First of all, they clarified the specific reason of their focus on detecting sentence-based similarities. If fact, Goodman and Villani remarked that they preferred to compare sentences instead of lines, as the former usually contain "a whole idea".

For every sentence each token was lemmatised ("reduced to its root") and stop-words were re-moved. To detect similarities between sentences, documents were then merged in a combined sentence-term matrix. By using it, results were ordered following the number of common words between two sentences from different books.

In order to re-create Goodman and Villani's experiment we used a set of tools available in Python NLTK library. We performed tokeniza-tion, stop-words removal and lemmatization[1]. Then, we detected the sentences that have more words in common and we ranked the results following this value. By doing this, we con-firmed Goodman and Villani's discovery, in fact we retrieved as one of the first results the two sentences presented in their paper (see Fig. 1).

All else had been subdued to me; alone
The soul of man, like unextinguished fire,
Yet burns towards heaven with fierce reproach, and doubt,
And lamentation, and reluctant prayer,
Hurling up insurrection, which might make
Our antique empire insecure, though built
On eldest faith, and hell's coeval, fear;
And though my curses through the pendulous air,
Like snow on herbless peaks, fall flake by flake,
And cling to it; though under my wrath's night
It climb the crags of life, step after step,
Which wound it, as ice wounds unsandalled feet,
It yet remains supreme o'er misery,
Aspiring, unrepressed, yet soon to fall;
Even now have I begotten a strange wonder,
That fatal child, the terror of the earth,
Who waits but till the destined hour arrive,
Bearing from Demogorgon's vacant throne
The dreadful might of ever-living limbs
Which clothed that awful spirit unbeheld,
To redescend, and trample out the spark.

The conquered, also, and enslaved by war,
Shall, with their freedom lost, all virtue lose,
And fear of God—from whom their piety feigned
In sharp contest of battle found no aid
Against invaders; therefore, cooled in zeal,
Thenceforth shall practise how to live secure,
Worldly, or dissolute, on what their lords
Shall leave them to enjoy; for the Earth shall bear
More than enough, that temperance may be tried.
So all shall turn degenerate, all depraved,
Justice and temperance, truth and faith, forgot;
One man except, the only son of light
In a dark age, against example good,
Against allurement, custom, and a world
Offended. Fearless of reproach and scorn,
Or violence, he of their wicked ways
Shall them admonish, and before them set
The paths of righteousness, how much more safe
And full of peace, denouncing wrauth to come
On their impenitence, and shall return
Of them derided, but of God observed
The one just man alive: by his command
Shall build a wondrous Ark, as thou beheld'st,
To save himself and household from amidst
A world devote to universal wrack.

Figure 1: Two sentences with a high similarity score following Goodman and Villani's approach. The one on the left is from Prometheus Unbound and the other from Paradise Lost.

[1]These algorithms have been trained on contemporary documents. For this reason we conducted a post-evaluation of the performances and we noticed that, for the purpose here described (tokenization, stop-words removal, lemmatization and later part of speech tagging) their performances could be in general considered solid. For what concerns stop-words removal we excluded four other extremely frequent tokens ("thou", "thy", "ye", "thro"). In the future we intend to im-prove this step of our work by developing in-domain tools.

4.1 A different approach, pt. 1

In order to get a better understanding of Goodman and Villani's approach, we compare their results with a standard natural language processing procedure for detecting similarities between two sentences. We re-ran their experiment considering each sentence as a tf-idf word vector (Sparck Jones, 1972). We removed extremely short sentences (with less than 4 words) and we use NLTK Part of Speech tagger to avoid homonyms. Then we computed the cosine similarity between each passage and we ranked the results according to that value.
In making these changes our aim was to normalise the word-frequency by the length of the sentence, in order to avoid that extremely long sentences would come out as first results (as in Goodman and Villani's work).
By using the inverse document frequency, sentences which have in common words that are rare in the books were ranked in higher positions.

Prometheus Unbound

Thus I am answered : strange!

Paradise Lost

Whom thus the angelick Virtue answered mild.

Figure 2: Two sentences with a high similarity score with the approach described in 4.1.

4.2 A different approach, pt. 2

We then performed a transformative improvement to this approach by looking at the mutual position of co-occurrences in the two texts. We ranked in higher positions couples that share two or more rare words which appear close in both sentences.

Prometheus Unbound

How canst thou hear? Who knowest not the language of the dead?

Paradise Lost

What callest thou solitude? Is not the Earth With various living creatures, and the air Replenished, and all these at thy command To come and play before thee? Knowest thou not Their language and their ways? They also know, And reason not contemptibly: With these Find pastime, and bear rule; thy realm is large.

Figure 3: Two sentences with a high similarity score with the approach described in 4.2.

By looking at the two sentences with a high similarity-values which we obtained recreating Goodman and Villani's approach, the major issue in their method is clearly evident. As they didn't normalise the word frequency by the length of the sentences, they ranked long sentences as top results. However it is clear that the similarity between these sentences is not very indicative of influence between authors.
If we instead consider two of the highest sentences detected by using our approaches we could notice that the common use of "rare words" and the close co-occurrence of specific couple of words could be useful to highlight poetic influence.
Starting from these encouraging results, in the near future our intention is to extend this study by considering other computational approaches, which, for example, have been developed in the area of text re-use (Clough et al., 2002), and by selecting a reliable metric for evaluating the different outputs of these ranking systems.

5 Results and conclusion

To conclude, for what concerns more specifically the possible boundary between humanities computing and the study of influence in literary studies, we believe that computational techniques helped to develop a keen sense of the issues involved, foregrounding the role of mechanical reading in de-idealising the problem. The use of machines in turn triggered Bloom's creation of his own "machine for criticism" (Bloom, 1976) built in order to understand how the anxiety of influence was dealt with by writers. If it's true that Bloom's theory "anxiously responds to [...] the subliminally perceived threat of textual anonymity promoted by the 'mechanical reproduction' of available texts" (Renza, 1990), that is to say something that with the digital will soon reach an entirely new, we can consider Raben's work as moved by a similar tension. Compared to Bloom, however, the advantage of the latter was that he was working from within, so to speak, the machine language. For this reason, he was able to begin transforming the traditional approach, that emphasised continuity within literary traditions, with a different one more focused on the way in which creative writing works, and on what the role of linguistic memory and of "creative" reading (or, as Prose (2006) said, reading as a writer).

References

Harold Bloom. 1973. The anxiety of influence. *Journal of PsychoAnalysis*, 76(1): 19-24.

Harold Bloom. 1976. *Poetry and Repression: Revisionism from Blake to Stevens*. New Haven, USA.

Harold Bloom. 1997. *The anxiety of influence: A theory of poetry*. Oxford University Press, Oxford, UK.

Silvana Castano, Alfio Ferrara and Stefano Montanelli. 2009. *Informazione, conoscenza e Web per le scienze umanistiche*. Pearson, Milano, IT.

Paul Clough, Robert Gaizauskas, Scott S.L. Piao and Yorick Wilks. 2002. Meter: Measuring text reuse. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Seymour Goodman and Raymond D. Villani. 1965. An algorithm for locating multiple word co-occurrences in two sets of texts. *Literary Data Processing Conference Proceedings*, 275-292.

Julianne Nyhan, Andrew Flinn, and Anne Welsh. 2012. A short Introduction to the Hidden Histories project and interviews. *DHQ: Digital Humanities Quarterly*, 6(3): 275-292.

Francine Prose. 2006. *Reading Like a Writer*. HarperCollins, New York, USA.

Joseph Raben. 1965. A computer-aided study of literary influence: Milton to Shelley. *Literary Data Processing Conference Proceedings*, 230-274.

Louis A. Renza. 1990. *Influence*, in Lenticchia and McLaughlin, eds., *Critical Terms for Literary Study.* University Press of Chicago, Chicago, USA.

Stefan Sinclair and Geoffrey Rockwell. 2014. Towards an archaeology of text analysis tools. Unpublished paper (DH2014, Lausanne, 7-12 July: http://dharchive.org/paper/DH2014/Paper-778.xml)

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, (28.1): 230-274.

Effectiveness of Domain Adaptation Approaches for Social Media PoS Tagging

Tobias Horsmann, Torsten Zesch
Language Technology Lab
Department of Computer Science and Applied Cognitive Science
University of Duisburg-Essen, Germany
{tobias.horsmann,torsten.zesch}@uni-due.de

Abstract

English. We compare a comprehensive list of domain adaptation approaches for PoS tagging of social media data. We find that the most effective approach is based on clustering of unlabeled data. We also show that combining different approaches does not further improve performance. Thus, PoS tagging of social media data remains a challenging problem.

Italiano. *Confrontiamo diversi approcci di adattamento al dominio per il PoS tagging di dati social media. Osserviamo che l'approccio più efficace si basa sul clustering di dati non annotati. Inoltre, mostriamo che la combinazione di diversi approcci non migliora ulteriormente le prestazioni. Di conseguenza, il PoS tagging di dati social media rimane un problema difficile.*

1 Introduction

Part-of-Speech (PoS) tagging of social media data is still challenging. Instead of tagging accuracies in the high nineties on newswire data, on social media we observe significantly lower numbers. This performance drop is mainly caused by the high number of out-of-vocabulary words in social media, as authors neglect orthographic rules (Eisenstein, 2013). However, special syntax in social media also plays a role, as e.g. pronouns at the beginning of sentence are often omitted like in "went to the gym" where the pronoun 'I' is implicated (Ritter et al., 2011). To make matters worse, existing corpora with PoS annotated social media data are rather small, which has led to a wide range of domain adaptation approaches being explored in the literature.

There are two main paradigms: First, adding more labeled training data by adding foreign or machine-generated data (Daumé III, 2007; Ritter et al., 2011). Second, incorporating external knowledge or guiding the machine learning algorithm to extract more knowledge from the existing data (Ritter et al., 2011; Owoputi et al., 2013). The first strategy affects from *which* data is learned, the second one *what* is learned.

Using more training data Usually there is only little PoS annotated data from the social media domain, so just using *re-training* on domain-specific data does not suffice for good performance. *Mixed re-training* adds additional annotated text from foreign domains to the training data. In case there is much more foreign data than social media data, *Oversampling* (Daumé III, 2007) can be used to adjust for the difference in size. Finally, *Voting* can be used to provide more social media training data by relying on multiple already existing taggers.

Using more knowledge Instead of adding more training data, we can also make better use of the existing data in order to lower the out-of-vocabulary rate. *PoS dictionaries* provide for instance information about the most frequent tag of a word. Another approach is *clustering* which group words according to their distributional similarity (Ritter et al., 2011).

In this paper, we evaluate the potential of each approach for solving the task.[1]

2 Baseline Tagger

We re-implement a state-of-the-art tagger in order to control all aspects of the process. It is based on CRFsuite[2] in version 0.12 as part of the text-classification framework DKProTC (Daxenberger

[1] Our experiments are available at http://tinyurl.com/neptn9e
[2] https://github.com/chokkan/crfsuite

et al., 2014). As training algorithm we use *Adaptive Regularization Of Weight* (AROW).

Our feature set follows previous work (Gimpel et al., 2011; Hovy et al., 2014). We use the word itself and the preceding and following word. We use boolean features for words containing capital letters, special characters, numbers, hyphens and periods, and for detecting words entirely composed of special characters or capital letters. We furthermore use the 1000 most frequent character bi- to four grams.

Our tagger achieves an accuracy of 96.4% on the usual WSJ train/test split which is close to the 96.5% by TNT tagger (Brants, 2000) and only slightly worse than the 97.2% of the Stanford tagger (Toutanova et al., 2003). When we evaluate our newswire tagger as is on the 15,000 token Twitter corpus by Ritter et al. (2011), accuracy drops to 76.1% which confirms their findings.

Having established these baselines, we now test the different domain adaption strategies. In order to reflect the domain difference, we will call the WSJ corpus NEWS and the Twitter corpus SOCIAL in the remainder of the paper.

3 Domain Adaptation Approaches

In this section, we explore existing domain adaptation approaches that can be divided into (i) using more training data or (ii) more knowledge.

3.1 More Training Data

We test three strategies (*re-training*, *oversampling*, and *voting*) using 10-fold cross validation on SOCIAL.

Re-training Simply re-training on SOCIAL improves accuracy from 76.1% to 81.9%, but still is far behind the 97% on newswire text. To estimate the potential of re-training, we show in Figure 1 the learning curve using increasing subsets of SOCIAL. The plot shape indicates that annotating additional in-domain data would be beneficial, but annotating more data is often so unattractive that domain adaption strategies are preferred anyway.

Another quite simple approach is training on NEWS and SOCIAL together, which we call *mixed re-training*. We evaluate this setting by cross validating only over SOCIAL and always adding the full NEWS corpus to the train set. This yields an accuracy of 82.7% compared to 81.9% on SOCIAL alone by adding two orders of magnitude more data (10^6 instead of 10^4 tokens).

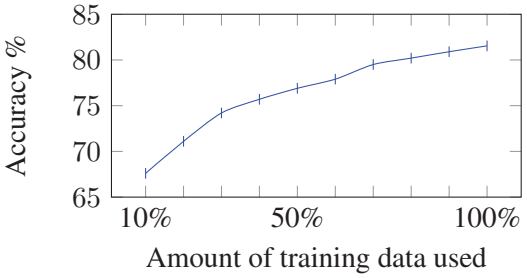


Figure 1: Re-training learning curve

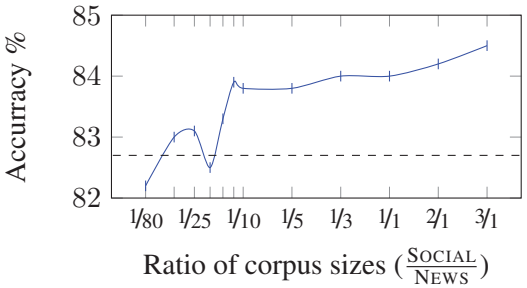


Figure 2: Oversampling results

Oversampling To overcome the size problem in *mixed-retraining*, oversampling the smaller corpus can be used (Daumé III, 2007; Neunerdt et al., 2014). The idea is to boost the importance of the small SOCIAL data by adding it multiple times (or adjusting the feature weights). We show the effect of varying oversampling rates i.e. the ratio of SOCIAL (size varied) to NEWS (size kept constant) in Figure 2. At an oversampling rate of 1:4, we achieve an accuracy of 84.5% which exceeds the *mixed-retraining* baseline of 82.7%.

Voting In this approach, a sample of unlabeled social media data is tagged using multiple existing PoS taggers. If they all assign the same label sequence (i.e. they all *voted* the same) the sentence is added to the training set as it is less likely that all taggers make the same mistakes. We use the PTB tagset taggers ClearNLP[3], OpenNLP[4] and Standford, setting the PoS tags for *Hashtags, Urls, Atmention* and *Retweet* manually in post-processing (Ritter et al., 2011). The results in Figure 3 show that it doesn't really matter how much voted data is added, we roughly see the same increase, with no real trend. We reach an accuracy of 83.5% at $6{\cdot}10^5$ additional tokens *voted* data. We show as comparison the curve if NEWS is added instead and find no disadvantages to the voting approach.

[3] http://www.clearnlp.com

[4] https://opennlp.apache.org

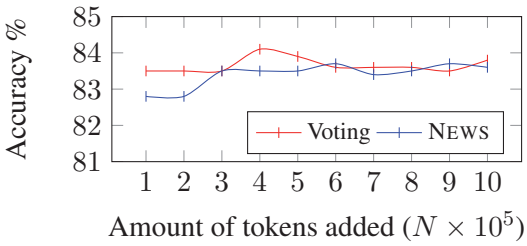


Figure 3: Voting vs. mixed-retraining

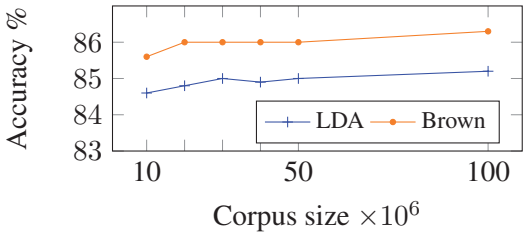


Figure 4: Clustering results

3.2 More Knowledge

In this section, we discuss the effect of adding more knowledge in the form of PoS dictionaries or word clusters.

PoS Dictionaries We use a dictionary that stores the PoS distribution for each word form that occurs in a corpus. The underlying corpus can either be manually annotated or machine-tagged (Gimpel et al., 2011; Rehbein, 2013).

We use two dictionaries in our experiments: ManualDict, created from the manually annotated Brown corpus (Nelson Francis and Kuçera, 1964), and MachineDict, created from 100 million tokens of the machine-tagged English WaCky corpus (Baroni et al., 2009). Surprisingly, both dictionaries equally improve the performance to 83.8%, the much bigger MachineDict providing no advantage. MachineDict covers about 60.3% of the tokens in SOCIAL while ManualDict only covers 54.0%. It seems that the higher quality of manual PoS annotations in ManualDict counters the higher coverage of MachineDict. The rather low coverage of both dictionaries is caused by cardinal numbers and social media phenomena such as Hashtags.

Clustering We experiment with two versions of clustering: LDA[5] (Blei et al., 2003; Chrupala, 2011) and hierarchical Brown clustering[6] (Brown et al., 1992). Following Owoputi et al. (2013) and Rehbein (2013), we create 1,000 Brown clusters with a minimal word frequency of 40, and 50 LDA clusters with a minimal word frequency of ten. We encode Brown cluster information following Owoputi et al. (2013).

Figure 4 shows that Brown clusters work much better than LDA, where the 100 million token Brown clusters reach the highest accuracy of 86.3%. Using the 800 million token Brown clusters provided by Owoputi et al. (2013) does not further improve results yielding an accuracy of 86.2%. We thus find that clustering is highly effective, but that very large corpus sizes might not translate into further increases.

[5] https://bitbucket.org/gchrupala/lda-wordclass/

[6] https://github.com/percyliang/brown-cluster

	Trained on	Acc. %
	Baseline	76.1
1	Re-training	81.9
2	Mixed re-training	82.7
3	Mixed re-training (Oversampling)	84.5
4	Re-training + Voting	83.5
5	PoS dictionary	83.8
6	Clustering	86.3
	Combo (4,5,6)	86.9

Table 1: Tagging accuracies per approaches

4 Combining Approaches

Combining approaches might further increase accuracy over the individual approaches summarized in Table 1. As the different strategies for adding more data are hard to combine, we select strategy #4 that provides good accuracy at much lower costs compared to oversampling.

PoS dictionaries and clustering seem to be effective and can easily be used together. Thus, our final combined model consists of re-training with the manually annotated SOCIAL data, 10,000 additional machine-tagged voting tokens, the MachineDict PoS dictionary, and the 100 million token Brown cluster. We achieve an accuracy of 86.9% accuracy, which is only a small improvement over clustering alone.

Comparison with State of the Art While our goal is not to exactly replicate previous work, it is quite informative to make the comparison. Ritter et al. (2011) reported 88.3% accuracy on the same dataset, but additionally added the NPS chat corpus for training, which is inline with our interpretation of Figure 1 that adding more hand-annotated

Adjectives		Interjections	
Token	Gold / Combo	Token	Gold / Combo
Happy	JJ	**Thanks**	UH / NNS
Berlated	JJ / NNP	and	CC
Birthday	NN	I	PRP
!	.	will	MD
When	WRB	in	IN
I	PRP	the	DT
Get	VBP	street	BB
Old	JJ / NNP	**loll**	UH / NN
,	,	.	.

Table 2: Adjective and interjection confusions

Word class	Combo fine	Combo coarse
ADJ	76.0	76.9
ADV	85.3	85.6
NN	80.9	91.6
V	81.9	91.4
All PoS	86.9	91.5

Table 3: Fine vs. coarse-grained accuracy

data is probably a good idea. Owoputi et al. (2013) reported 90%, but additionally use several name lists to detect proper nouns. We are going to explore the impact of this kind of tag specific optimization in section 5.

Error Examples Table 2 shows representative errors for the frequently occurring classes adjectives and interjections. The first adjective error shows a confusion of an out-of-vocabulary item with capital letter. The second error is also caused by the first letter in uppercase. Interjections are notoriously hard to tag, as they are mainly pragmatic markers.

5 Practical Issues

We now turn to some practical issues that influence the interpretation of the obtained results.

5.1 Coarse-grained Performance

Tagging social media is hard also because the lack of context and informal writing sometimes make fine-grained decisions about a certain PoS tag almost impossible. For example, in *He dance on the street* the word *dance* is a verb, but its intended tense is not easily determinable. We thus test whether the accuracy improvement mainly happens within a coarse tag class or between classes (e.g. only confusions between regular (NN) and proper nouns (NNP) are corrected).

Table 3 shows the re-calculated accuracy of the *Combo* approach, counting as correct not only exact matches, but also if the assigned PoS tag matches the coarse-grained PoS class. For nouns and verbs, we see that accuracy improves from the low 80's to the low 90's which means that many mistakes are intra-class here (e.g. NN vs NNP). Thus, tagging accuracy for coarse-grained word classes is already much higher than the numbers might show and tagging of adjectives and adverbs is the biggest remaining problem.

5.2 Influence of the System Architecture

While experimenting with CRFsuite, we noticed that the same set of train/test data yields different results on different system architectures (Windows 7, OS-X 10.10, and Ubuntu).[7]

Just by chance, changing platform might give you a performance increase that is in the same range as the best domain adaptation strategy discussed in this paper. This shows that failure to reproduce previous results can have unexpected causes far beyond the actual research question to be tested.

6 Conclusion

In this paper, we analyzed domain adaptation approaches for improving PoS tagging on social media text. We confirm that adding more manually annotated in-domain data is highly effective, but annotation costs might often prevent application of this strategy. Adding more out-domain training data or machine-tagged data is less effective than adding more external knowledge in our experiments. We find that clustering is the most effective individual approach. However, clustering based on very large corpora did not further increase accuracy. As combination of strategies did only yield minor improvements, clustering seems to dominate the other strategies.

References

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.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan.

[7] Cause is a shuffling operation of the train set that is initialized differently among operating system architectures.

2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Thorsten Brants. 2000. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA.

Peter F Brown, Peter V DeSouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18:467–479.

Gzegorz Chrupala. 2011. Efficient induction of probabilistic word classes with LDA. In *Proceedings of the Fifth International Joint Conference on Natural Language Processing : IJCNLP 2011*, page 363, Chiang Mai, Thailand.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. When pos data sets don't add up: Combatting sample bias. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

W. Nelson Francis and Henry Kuçera. 1964. Manual of Information to Accompany a Standard Corpus of Present-day Edited American English, for use with Digital Computers.

Melanie Neunerdt, Michael Reyer, and Rudolf Mathar. 2014. Efficient Training Data Enrichment and Unknown Token Handling for POS Tagging of Non-standardized Texts. In *12th Conference on Natural Language Processing (KONVENS)*, pages 186–192, Hildesheim, Germany.

Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ines Rehbein. 2013. Fine-Grained POS Tagging of German Tweets. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 162–175.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA.

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*, NAACL '03, pages 173–180, Stroudsburg, PA, USA.

Building a Corpus on a Debate on Political Reform in Twitter

Mirko Lai[1], Daniela Virone[2], Cristina Bosco[1], Viviana Patti[1]
[1]Dipartimento di Informatica, Università degli Studi di Torino
[2]Dipartimento di studi umanistici, Università degli Studi di Torino
{lai,bosco,patti}@di.unito.it, dvirone@unito.it

Abstract

English. The paper describes a project for the development of a French corpus for sentiment analysis focused on the texts generated by the participants to a debate about a political reform, i.e. the bill on homosexual wedding in France. Beyond the description of the data set, the paper shows the methodologies applied in the collection and annotation of data. The collection has been driven by the detection of the hashtag mainly used by the participants to the debate, while the annotation has been based on polarity but also extended with the target semantic areas involved in the debate.

Italiano. *L'articolo descrive un progetto per lo sviluppo di un corpus per la sentiment analysis composto di testi in francese prodotti dai partecipanti ad un dibattito su una riforma politica, i.e. la legge sul matrimonio gay in Francia. Oltre a descrivere il dataset, l'articolo mostra le metodologie applicate nella raccolta e nell'annotazione dei dati. La raccolta dei dati è stata guidata dalla presenza dell'hashtag maggiormente utilizzato dai partecipanti al dibattito, mentre l'annotazione è basata oltre che sulla polarità anche sulle aree semantiche toccate dai partecipanti nel dibattito.*

1 Introduction

The recent trends in sentiment analysis are towards hybrid approaches or to computational semantics oriented frameworks where linguistic and pragmatic knowledge are encompassed for describing a global notion of communication. This notion includes e.g. context, themes, dialogical dynamics in order to detect the affective content even if it is not directly expressed by words, like, for instance, when the user exploits figurative language (e.g. irony, metaphor or hyperbole) or, in general, when the communicated content does not correspond to words meaning but depends also on other communicative behaviors.

On this perspective, a particular interesting domain is related to the political debates and, in particular, in the specific form that such debates assumes in social media, which strongly differentiate them from other kinds of classical conversational contexts (Rajadesingan and Liu, 2014). In the last years social media, and in particular Twitter, have been used in electoral campaigns by different actors involved in the process: by campaign staffs in order to disseminate information, organize events; by the news media in order to inform and promote news content; and by voters to express and share political opinions. Therefore recently many studies focused on understanding the phenomenon, by studying the effect of this technology on the election outcomes (Skilters et al., 2011), its possible use to gauge the political sentiment (Tumasjan et al., 2011) and the users' stance on controversial topics (Rajadesingan and Liu, 2014), or by studying the networks of communication in order to investigate the political polarization issue (Conover et al., 2011).

This study contributes to this area by showing a methodology for the collection and annotation of a data set composed by texts from different media where a political debate has been developed. As a starting point of the project in this paper we will present a dataset of Twitter messages in French language about the reform "Le Mariage Pour Tous" (Marriage for everyone, i.e. marriage for all), discussed in France in 2012 and 2013. The collection of th dataset has been driven by a hashtag, i.e. #mariagepourtous, created to mark the messages about the debate on the reform, while the

selection of tags to be annotated has been based on the detection and analysis of the semantic areas involved in users posts. The detection of these areas is the result of a set of analysis we applied on the corpus described in more details in (Lai et al., 2015).

The paper is organized as follows. The next two sections respectively describe related works and the data set, showing the criteria and methodologies applied for the selection of data. Fourth section is in instead devoted to the annotation of collected data.

2 Related work

Several works rely on sentiment analysis techniques (Pang and Lee, 2008) to analyze politics (Tumasjan et al., 2011; Li et al., 2012; He et al., 2012), a domain where the problems related to the exploitation of figurative language devices described in (Maynard and Greenwood, 2014; Bosco et al., 2013; Reyes et al., 2012; Reyes et al., 2013; Gianti et al., 2012; Davidov et al., 2011) and in the Semeval15-11 shared task (Ghosh et al., 2015) have been detected as frequent. Moreover, some research focused on aspects concerning the political polarization in Twitter (Conover et al., 2011; Skilters et al., 2011), or on the detection of the stance of Twitter users from their tweets debating a controversial topic, such as abortion, gun reforms and so on (Rajadesingan and Liu, 2014)[1]. All such perspectives are very interesting also in the dataset we are describing in the current work.

Other works, instead, addressed the issues related to the arguments accompanying the political messages, like (Eensoo and Valette, 2014) where an analysis devoted to discover in the tweets the argumentation related to evaluative discourse is presented and applied to the case of the racism anti-Rom in the Web; it is shown that a discourse where a form of evaluation is expressed does not necessarily exploits semantic and linguistic markers traditionally linked to the evaluation, but it can be also based on dialogical and dialectical components. This is a strong motivation for the development of annotated corpora where this kind of knowledge can be reliably described. The idea to focus the analysis on the debate around a reform can lead to get some new insights on the communicative behavior in using subjective and evaluative language in politics.

Finally let us to notice that most of the works carried on so far in this area focus their analysis on English datasets only, while under this respect several languages, like French or Italian, are currently under-resourced, with some exception (Stranisci et al., 2015).

3 Collection and composition of the data set

This work is collocated in the context of an ongoing project about communication in different media and is focused on the debate about homosexual couple wedding in France. The project includes the collection of the following datasets from different media and sources:

- TW-MariagePourTous: texts from Twitter selected by filtering the tweets posted in the time-lapse 16th December 2010 - 20th July 2013 for French language and for the presence of the hashtag *#mariagepourtous*.
- NEWS-MariagePourTous: texts from French newspapers, i.e. LeMonde online and sources retrieved by using the Factiva search engine[2], published in the time-lapse 7th June 2011 - 4th February 2013 and filtered by the keyword *#mariagepourtous*.
- NEWSTITLE-MariagePourTous: titles only of the texts collected in the NEWS-MariagePourTous corpus.
- DEBAT-MariagePourTous: texts from parliamentary debates about the first discussion of the bill on homosexual wedding (meetings of the National Assembly and Senate of the French Parliament from 27th January 2013 to 12th February 2013) and the following meetings (from 4th to 12 April 2013 and from 15th to 23th April 2013) where the bill has been approved[3].

The largest corpus is NEWS-MariagePourTous, which includes around 24,000 articles, while

[1] A new task on *Detecting Stance in Tweets* has been proposed in Semeval-2016 (Task 6) as part of the Sentiment Analysis Track: `http://alt.qcri.org/semeval2016/task6/`

[2] See `http://new.dowjones.com/products/factiva/`.

[3] See `http://www.assemblee-nationale.fr/14/debats/` for the transcription of debates of the National Assembly, and those `http://www.senat.fr/seances/comptes-rendus.html` for the debates in Senate made available by the French Government.

Figure 1: A cloud-style representation of words distribution in the TW-MPT dataset.

NEWSTITLE-MariagePourTous is the smallest.

The current study focus on the MariagePourTous dataset (henceforth TW-MPT), which includes 254,366 messages. 88,157 of them have been retweetted by one or more users during the time of the corpus collection[4]. Each tweet is associated with the metadata related to the posting time and the user that posted it, information that can be exploited in the analysis of data.

The collection of this corpus is based on the detection of the hashtag *#mariagepourtous*. Hashtags are single words or expressions (with words not separated by spaces) preceded by the symbol '#', well known in Twitter and exploited by users to create communities of people interested in the same topic (Cunha et al., 2011), by making it easier for them to find and share information related to it (think, for instance, of the hashtags/slogans created during election campaigns). When a user exploits an existing hashtag, he/she wants to be recognized as belonging to the group using it, to be accepted within the dialogical and social context growing around the topic (Chiusaroli, 2012), but not necessarily in order to assume the same opinion about the content of the hashtag. For instance, *#mariagepourtous* has been used by people expressing both positive and negative opinions about homosexual wedding in France.

[4]We didn't include in the annotated corpus the retweetted messages but we have this information available for further processing and statistics (Lai et al., 2015).

By selecting a hashtag as our main filtering criterion, we easily collected several arguments and different opinions expressed by the persons interested in the web debate about the topic. Furthermore, we could observe the "life" of the hashtag during its first propagation among Twitter users, and then diffusion within the community (see (Lai et al., 2015)).

4 Data analysis and data-driven annotation

As previously reported, the limited amount of resources available for French sentiment analysis makes the development of a sentiment annotation for the TW-MPT corpus an especially valuable effort. Nevertheless, our main goal was to test a methodology for the definition of a data-driven annotation scheme, which can be applied also in other cases, and, in particular, in socio-political debates for making explicit the features of this kind of conversational context. Therefore, our annotation scheme extends the standard annotation for marking the polarity of opinions and sentiments, usually applied in corpora annotated for sentiment analysis, by including both tags for marking figurative language devices and a set of semantically oriented labels. The analyses based on linguistic and non linguistic features described in (Lai et al., 2015), which we applied for detecting the dynamics of communicative behavior of

users in exploiting subjective and evaluative language, meaningfully helped us in designing this annotation scheme.

For what concerns polarity, we applied in this project the same approach applied in (Gianti et al., 2012; Bosco et al., 2013; Bosco et al., 2014; Bosco et al., 2015) for the annotation of Italian corpora for sentiment analysis, which includes the tags of table 1.

The annotation of figurative devices is based on three labels: HUM POS for marking the presence of irony featured by positive polarity, HUM NEG for negative irony, and a yes/no feature for METAPHOR. Also other figurative devices (e.g. hyperbole) can be of interest for sentiment analysis, but the extension of the schema in this direction will be object of future work.

label	polarity
POS	positive
NEG	negative
NONE	neutral
MIXED	both positive and negative
UN	unintelligible content
RP	repetition of a post

Table 1: Polarity tags annotated in the TW-MPT corpus.

For what concerns, instead, the set of semantically oriented tags, we defined them according to an analysis of the dataset. In fact, observing the corpus and the other collected data, we hypothesized that the debate developed around some particular topic, and the experiments performed validated our hypothesis. We classified the most frequently occurring words by the application of the cloud extraction techniques described in (Lai et al., 2015) to the full TW-MPT corpus tag. The result is that represented in Figure 1, showing that user tweets focused on few quite sharply distinguishable semantic areas encompassing several other relevant discussed themes: *family* (we labeled as FAMILLE), *legal aspects* (we labeled as LOI), *public manifestations* (we labeled as MANIF), *socio-political debate* (we labeled as DEBAT).

The annotation scheme has been applied on a first portion of 2,872 tweets of the TW-MPT corpus, i.e. all the posts where the hashtag occurs immediately after or before the verb "etre" (*to be*), namely the messages where the hashtag is in some way evaluated or defined by users. After the discussion and definition of a set of the guidelines to be shared, the annotation has been done by two independent skilled annotators, and the disagreement has been calculated and analyzed. Before the final release of the corpus, which will be available soon, a third annotation will be applied in order to improve the reliability of data, but some preliminary hints can be derived from the analysis of the currently available data.

The disagreement on polarity appears in 861 of the 2,872 annotated tweets, but it is mainly focused on cases where irony is involved. In 184 tweets only one annotator detected irony when the other doesn't, but both detected the same polarity. For instance, annotator-1 used POS and annotator-2 used HUM-POS, or viceversa or the same with the labels HUM-NEG and NEG. This confirms the hypothesis of a variable perception of irony among humans (González-Ibáñez et al., 2011). Only a limited amount of cases (177) have been found where the annotators disagreed annotating opposed polarities (i.e. POS and NEG). In a few remaining cases the disagreement depends on the annotation of a neutral polarity versus a defined polarity (173) or mixed polarity with respect to a sharp one (86). For what concerns the annotation of the semantic areas, it is featured by a very high agreement (the annotators selected the same label in 1958 cases). However, further investigations are needed in order to find areas where they mainly disagree. Finally, for what concerns the detection of metaphors, the related annotation is still in progress, as it has been applied to the corpus in a second stage.

5 Conclusions and future work

The paper presents a data-driven methodology for collecting and annotating corpora for sentiment analysis, which has been applied to a French corpus of a Twitter debate about a political reform. The collection is driven by a hashtag expoited by users expressing opinions of a controversial topic. The annotation is based on a set classical polarity labels, extended with tags for figurative language devices (i.e. irony) and for a few semantic areas detected in posts, intended as *aspects* of the reform on which users express their opinions.

The investigation of further aspects and information sources that can be found in data, e.g. emojis, links and images, is matter of future work.

Acknowledgements

The authors thank all the persons who supported the work, and in particular Federica Ramires that meaningfully contributed to the annotation and analysis of the corpus as part of her Bachelor's degree thesis.

References

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.

Cristina Bosco, Leonardo Allisio, Valeria Mussa, Viviana Patti, Giancarlo Ruffo, Manuela Sanguinetti, and Emilio Sulis. 2014. Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in Felicittà. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Opena Data, ESSSLOD 2014*, pages 56–63, Reykjavik, Iceland. ELRA.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2015. Developing corpora for sentiment analysis: The case of irony and senti-tut (extended abstract). In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4158–4162. AAAI Press / International Joint Conferences on Artificial Intelligence.

Francesca Chiusaroli. 2012. Scritture brevi oggi. tra convenzione e sistema. In Francesca Chiusaroli and Fabio Massimo Zanzotto, editors, *Scritture brevi di oggi*, pages 4–44. Università Orientale di Napoli.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Political polarization on twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos Andre Goncalves, and Fabricio Benevenuto. 2011. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 58–65, Portland, Oregon. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2011. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the CONLL'11*, pages 107–116, Portland, Oregon (USA).

Egle Eensoo and Mathieu Valette. 2014. Approche textuelle pour le traitement automatique du discours evaluatif. *Langue francaise*, 184(4):109–124.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and John Barnden. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proc. Int. Workshop on Semantic Evaluation (SemEval-2015), Co-located with NAACL and *SEM*.

Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli, and Luigi Di Caro. 2012. Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2013)*, pages 1–7, Istanbul, Turkey. ELRA.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586. Association for Computational Linguistics.

Yulan He, Hassan Saif, Zhongyu Wei, and Kam-Fai Wong. 2012. Quantising opinions for political tweets analysis. In *Proceedings of the LREC'12*, pages 3901–3906, Istanbul, Turkey.

Mirko Lai, Daniela Virone, Cristina Bosco, and Viviana Patti. 2015. Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. In *Proceedings of IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015)*. IEEE. In press.

Hong Li, Xiwen Cheng, Kristina Adson, Tal Kirshboim, and Feiyu Xu. 2012. Annotating opinions in German political news. In *Proceedings of the LREC'12*, pages 1183–1188, Istanbul, Turkey.

Diana Maynard and Mark 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'14)*, Reykjavik, Iceland, may. ELRA.

Bo Pang and Lillian Lee. 2008. *Opinion Mining and Sentiment Analysis (Foundations and Trends(R) in Information Retrieval)*. Now Publishers Inc.

Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in twitter debates. In William G. Kennedy, Nitin Agarwal, and Shanchieh Jay Yang, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 8393 of *Lecture Notes in Computer Science*, pages 153–160. Springer International Publishing.

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.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Jurgis Skilters, Monika Kreile, Uldis Bojars, Inta Brikse, Janis Pencis, and Laura Uzule. 2011. The pragmatics of political messages in twitter communication. In Raul Garcia-Castro, Dieter Fensel, and Grigoris Antoniou, editors, *ESWC Workshops*, volume 7117 of *Lecture Notes in Computer Science*, pages 100–111. Springer.

Marco Stranisci, Cristina Bosco, Patti Viviana, and Delia Irazú Hernández Farias. 2015. Analyzing and annotating for sentiment analysis the socio-political debate on "La Buona Scuola". In *Proceedings of the 2th Italian Conference on Computational Linguistics (CLiC-IT 2015)*, Trento, Italy. In Press.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2011. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the ICWSM-11*, pages 178–185, Barcelona, Spain.

The OPATCH corpus platform – facing heterogeneous groups of texts and users

Verena Lyding[1], Michel Généreux[1], Katalin Szabò[2], Johannes Andresen[2]
[1]EURAC research, viale Druso 1, 39100 Bolzano, Italy
[2]Dr. Friedrich Teßmann library, Via A.-Diaz 8, 39100 Bolzano, Italy
firstname.lastname@eurac.edu,firstname.lastname@tessmann.it

Abstract

English. This paper presents the design and development of the OPATCH[1] corpus platform for the processing and delivery of heterogeneous text collections for different usage scenarios. Requirements and technical solutions for creating a multipurpose corpus infrastructure are detailed by describing its development.

Italian. *L'articolo presenta il design e lo sviluppo della piattaforma OPATCH[1] per l'elaborazione e la distribuzione di testi eterogenei in differenti contesti d'uso. La procedura evidenzia le esigenze e le soluzioni tecnologiche nella creazione di una ampia infrastruttura per i corpora.*

1 Introduction

Nowadays, electronic text collections have become an important information source in different usage contexts, including linguistic research and research in the humanities. With regard to application contexts and user groups, requirements related to tools for analyzing the text collections can differ. This mainly concerns the level of search interfaces, while the underlying data processing and annotation procedures typically build on standard NLP technologies.

In this paper, we are presenting the OPATCH project[1], which aims at creating a multipurpose corpus platform, by combining a uniform system back-end with varied front ends for different usage scenarios. Its flexible use is illustrated through two sample portals: one for content research and one for a linguistic search scenario.

2 Related work

Recent works related to the use of NLP technologies for enhancing humanities research environments include ALCIDE (Moretti et al., 2014), a platform for the automatic textual analysis of historical text documents, and GutenTag (Brooke et al., 2015), a tool which serves as a reader, subcorpus builder and tagging tool for literary texts.

3 Overview of the OPATCH platform

The OPATCH platform aims at providing a comprehensive infrastructure for the processing and delivery of digital text documents. The platform is designed to serve multiple purposes with regard to both the textual resources and the usage contexts it can accommodate. The platform consists in a system back-end, a unique component for text processing, and an extendable number of front-end portals, the user interfaces.

The *system back-end* combines a variety of standard text processing and annotation tools into ready-to-use tool chains. These tool chains are designed in such a way that intermediate text and corpus formats are in accordance with established standards for corpus data exchange, and output data is made compliant with standard formats of corpus and text search environments. Furthermore, the platform back-end strictly relies on open source tools.

The *system front-end* consists of an extendable series of portals for specific usage scenarios. Within the project scope, its applicability for different use cases is demonstrated by two portals:

(1) a *historical newspaper portal* for research and investigations on cultural content
(2) a *linguistic corpus portal* for language research purposes

[1] OPATCH - *Open Platform for Access to and Analysis of Textual Documents of Cultural Heritage*, financed by the 'Provincia Autonoma di Bolzano-Alto Adige, Diritto allo studio, università e ricerca scientifica, Legge provinciale 13 dic. 2006, n. 14'; project duration: Jan 2014 – Mar 2016; http://commul.eurac.edu/opatch/website/index.html

4 Data

The two use cases deliver two different text collections: The content search on local history is based on a collection of historical newspapers and the linguistic search is based on a collection of documents of current South Tyrolean German.

4.1 Historical newspaper data

The newspaper corpus contains 100.000 pages of German newspapers from (South) Tyrol for the years 1910 to 1920. They are part of the historical newspaper archive from the Alpine region held at the Dr. Friedrich Teßmann library and were selected with regard to maximizing OCR (Optical Character Recognition) quality and to including full (not partial) news issues. Table 1 shows the division by newspaper title and years.

Newspaper	Years	Pages
Bozner Nachrichten	1910-20	24786
Der Tiroler	1910-20	19933
Meraner Zeitung	1910-20	19286
Bote für Tirol	1910-19	9497
Volksblatt	1910-20	9275
Lienzer Zeitung	1910-15, 19	8479
Tiroler Volksbote	1910-19	6746
Bozner Zeitung	1911, 13, 15	1100
Pustertaler Bote	1917-20	898
Total		**100000**

Table 1. Subdivision of newspaper corpus

4.2 'Korpus Südtirol' core corpus

The other text collection consists in the core part of the 'Korpus Südtirol' (KST) initiative (Abel et al., 2009) enhanced with further newspapers. The core corpus consists of balanced texts of four genres: fiction, informative, functional (e.g. user manuals) and journalistic texts. It has a size of 3.5 Mio tokens and spans the entire 20th century.

5 Platform design

The platform design has been informed by user requirements and specified with reference to format standards and available open source tools.

5.1 User requirements

Both portals deliver documents of South Tyrolean cultural heritage and aim at fostering research on local history and language. The newspaper portal serves humanities related research with a focus on *textual content*, while the linguistic portal targets studies on *linguistic characteristics* of the South Tyrolean texts. Accordingly, the user studies addressed different target groups: historians, town/family chroniclers, teachers and students for the newspaper portal, and linguists and language planners/testers for the linguistic portal.

For the newspaper portal, two user studies were performed. In study (1), interviews with 13 library users yielded the following requirements:

- Research topics: local history, family history, world war, media history, literary study
- Objectives: research work, thesis preparation (students), preparation of teaching material
- Modes of access: by date, by topic, full text search with focus on names and events
- Use of results: saving results, references and query history, export and printing, notes
- Additional features: highlighting of search terms, overview of data base, user space

Study (2) evaluated an early interface prototype via an online questionnaire compiled by 55 respondents. It yielded the following results:

- Navigation: clear interface structure (80%)
- Modes of access:[2] text based search (80%), by title (35%), by date (25%)
- Required search facilities: [3] multiword searches, Boolean, Regular Expressions, searches, combination of text-based and filters, search by page number, fuzzy search
- Results display: standard view, pdf and download are most used (>60% often/ always; < 18% rarely/never); animated and tiles view are hardly used (< 10% often/always; > 65% rarely/never)
- Additional features:[3] persistent links to results, more (Italian) content, query storage, download of entire articles, ordering of results, adaptation to mobile devices, API

For the linguistic search, OPATCH relied on user studies from previous corpus projects (cf. Wisniewski et al. (to appear), Lyding et al. (2013)). Accordingly, primary requirements are:

- Powerful query facilities
- Search on linguistic annotations / metadata
- Focus on frequencies
- Visualization of frequencies in concordances

5.2 Format and annotation requirements

The design of the OPATCH platform is oriented on established standards for the description of language resources. With reference to the European infrastructure initiative CLARIN [4] ,

[2] respondents who "often/always" use this search
[3] as listed in free text field by several respondents
[4] http://clarin.eu

OPATCH aims at compatibility with CMDI (Component MetaData Infrastructure) for the exchange of metadata, and at providing an FCS (Federated Content Search) endpoint for the final linguistic corpus portal. Furthermore, different standard formats for encoding texts throughout processing are employed: METS/ALTO[5] format for OCRed newspaper issues files, ALTO format for single pages of text with linguistic annotations, Lucene/SOLR indexes for newspaper portal back-end, IMS Open Corpus Workbench (openCWB)[6] vertical format for linguistic portal back-end, and custom text format for the *Double Tree* (Culy/Lyding 2010) visualization front-end.

Regarding linguistic annotations and mark-up, all text documents are required to be tokenized and split into sentences, and to be annotated with metadata, lemma, part-of-speech and NE (Named Entity) information.

5.3 Portal specifications

The configuration of each portal has been specified in relation to its general purpose and in response to results of the user studies.

The **newspaper portal** aims at serving research on cultural topics and thus targets the retrieval of textual *content*. The design foresees:
(1) different search modes
(2) full access to the source data

Search options are designed to combine the access via metadata filtering (e.g. by year, title, etc.) with full text search and search by linguistic annotations (e.g. NE). This way, the portal offers text browsing and targeted searches.

The presentation of search results is designed to allow for a comprehensive view on the data, by providing the digitized text as well as the original image files. Furthermore, the possibility to highlight search terms, and download or print search results and related documents is foreseen.

The **linguistic corpus portal** aims at serving research on structural language characteristics. Accordingly, it aims at providing:
(1) powerful query facilities
(2) access to contextualized text and statistics

The query facilities are designed to support searches on text and annotation layers, including Regular Expression searches and the use of Boolean operators.

The presentation of search results, next to a standard KWIC display, foresees a *Double Tree* view, which highlights frequent word combinations and allows for the interactive exploration of results with regard to sequential and frequency characteristics of the data.

5.4 Back-end specification

The specification of the system back-end is based on the functional demands that have been derived from the user studies for both portals and the technical requirements for text processing and annotation related to them. In order to serve the two portals, the OPATCH system has to accommodate a series of tools into a flexible tool chain. The processing measures specified for the OPATCH system are presented in Table 2, in order of their application. The Table also indicates the tools used and the applicability of measures to the data of each portal.

Processing measure	**Tool**	**News**	**KST**
OCR recognition	ABBYY's fine reader	yes	partly
OCR post-processing	Custom model	yes	-
layout recognition	Formal description	-	-
metadata collection	manually	yes	yes
tokenization	Treetagger	yes	yes
lemmatization	Treetagger	yes	yes
part-of-speech tagging	Treetagger	yes	yes
NE recognition	Stanford NER and lists	yes	yes
transformation to format of retrieval tools	Lucene/SOLR and open CWB	yes	yes

Table 2. Text processing and annotation

6 Processing and annotation chain

The following subsections describe in detail the processing procedures listed in Table 2 and discuss particular challenges related to the two types of text collections and portals

6.1 OCR and post-processing

Processing of the newspaper data started from OCRed text files (METS/ALTO format), which had been produced using ABBYY's Fine Reader.[7] Due to the printing in 'Fraktur' font and a partly deteriorated paper quality, the data

[5] flexible XML schema for describing complex digital objects, maintained by the Library of Congress
[6] http://sourceforge.net/projects/cwb/

[7] OCR recognition was done within the *Europeana Newspaper Project (Europeana)*, see: http://www.europeana-newspapers.eu/

showed a very low quality. An evaluation of 10 pages of the OCRed collections gives an average bag-of-word (BoW) index success rate of 67.5%. BoW evaluations apply well to texts with complex layout structures (newspapers), cf. Pletschacher et al. (2014), while more refined evaluations that go beyond Levenshtein or edit distance may be better suited for more uniform layout such as books, cf. Reynaert (2014).

The post-correction of the OCRed texts was approached by applying a multi-step transformation model of edit operations on single or multiple letters, trained on manually corrected data. In an experiment, we could show significant reductions in error rates for words no further than two edit-operations from their true value. The task of correcting OCRed texts of newspapers is made more difficult by complex layouts, dislocated or merged words and incomplete dictionaries (Généreux et al., 2014).

At project start, KST texts have been available in digital format. OCR post-correction has been no issue, as texts were either genuine electronic text or high quality prints in modern fonts. [8]

6.2 Layout recognition

The feasibility of automatic layout recognition for historical newspapers has been investigated, related to experimentations of the project partner library Teßmann.[9] A manual analysis of section headings and layouts of ten newspapers showed:

- Text appears in three columns (rarely two)
- Vertical and horizontal separation of articles by lines or little star signs
- No headlines, but titles in 2-3 font sizes
- Very compact printing, little free space
- Advertisements with varied layout/fonts
- Content-related subdivisions are recurrent

Within OPATCH, the automatic layout division has been limited to separation of pages and distinction of header data and core text.

6.3 Metadata collection

For the newspaper corpus, metadata was collected for entire issues and includes 'title/publisher', 'publication date', 'no. of printed issues', 'no. of pages' and 'font type'. The metadata was recorded during the OCR step or added after.

For the KST corpus, great effort has been dedicated to the systematic collection of detailed metadata sets (Anstein et al., 2011). In particular, literary, functional and informative texts are associated with detailed descriptions of the author, publisher or text content, which for newspaper data would need to be assigned on article level.

The shared set of metadata among both collections covers: title, publisher, publication date.

6.4 Linguistic annotations: tokenization, lemmatization, part-of-speech and NER

For tokenization, lemmatization and part-of-speech tagging the IMS Treetagger (Schmid, 1994) trained for German has been used. Regarding Named Entity Recognition (NER), for both, the newspaper collection and the KST corpus, two approaches were combined: the Stanford NER tool re-trained for South Tyrolean German and the exact matching of texts with detailed lists of South Tyrolean names (place names[10], person names[11], addresses and organization names[12]).

The corrected OCRed output complies with the latest ALTO-XSD specification (v2.1, Feb. 20, 2014), which enforces a consistent enumeration of all entities, including multi-word entities.

6.5 Transformation for retrieval engines

The newspaper portal and the linguistic portal are based on different retrieval engines, in order to respond to the relative requirements. The newspaper portal relies on Lucene/SOLR which allows for the efficient retrieval of plain text and facetted searches based on metadata.

The linguistic portal relies on the openCWB which provides support for linguistic annotations on token level and a powerful query processor which allows for Regex and Boolean searches. Transformations towards the required input data formats have been handled by custom scripts.

7 Conclusion

This article reported on the design and development of the OPATCH corpus platform. Based on two usage scenarios for different target groups, relevant considerations concerning requirements towards a comprehensive corpus infrastructure

[8] today, texts in standard fonts yield OCR accuracies of 90% (Kettunen et al., 2014; Kettunen, 2015)

[9] experimentations carried out within *Europeana*

[10] from database for South Tyrolean place names, http://www.uibk.ac.at/germanistik/fachbereiche/germanistische_linguistik/forschung_flurnamen.html

[11] list of South Tyrolean names, cf. Strickner (2011)

[12] taken from historical address books, 1911-1922

have been illustrated and the technical solutions chosen in OPATCH have been presented.

References

Andrea Abel, Stefanie Anstein, and Stefanos Petrakis. 2009. Die Initiative Korpus Südtirol. In *Linguistik Online*, vol. 38, no. 2, 2009.

Stefanie Anstein, Margit Oberhammer, and Stefanos Petrakis. 2011. Korpus Südtirol - Aufbau und Abfrage. In A. Abel & R. Zanin (eds.), *Korpora in Lehre und Forschung*. Bozen - Bolzano: University Press, 15-28.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: an NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, held at NAACL HLT 2015, 42-47.

Chris Culy and Verena Lyding. 2010. Double Tree: An Advanced KWIC Visualization for Expert Users. In *Information Visualization, Proceedings of IV 2010, 14th International Conference Information Visualization*, 98-103.

Michel Généreux, Egon Stemle, Lionel Nicolas, and Verena Lyding. 2014. Correcting OCR Errors for German in Fraktur Font. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-It 2014)*, edited by R. Basili, A. Lenci & B. Magnini, Pisa, Italy.

Kimmo Kettunen, Timo Honkela, Krister Lindén, Pekka Kauppinen, Tuula Pääkkönen, and Jukka Kervinen. 2014. Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. In *IFLA World Library and Information Congress Proceedings: 80th IFLA General Conference and Assembly*.

Kimma Kettunen. 2015. Keep, Change or Delete? Setting up a Low Resource OCR Post-correction Framework for a Digitized Old Finnish Newspaper Collection, presented at the *11th Italian Research Conference on Digital Libraries - IRCDL 2015,* Bozen-Bolzano, Italy, 29-30 January, 2015, http://ircdl2015.unibz.it/papers/paper-01.pdf

Verena Lyding, Claudia Borghetti, Henrik Dittmann, Lionel Nicolas, and Egon Stemle. 2013. Open Corpus Interface for Italian Language Learning. In *Proceedings of ICT4LL 2013, 6th edition of the ICT for Language Learning Conference*. libreriauniversitaria.it

Giovanni Moretti, Sara Tonelli, Stefano Menini, and Rachele Sprugnoli. 2014. ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-It 2014)*, edited by R. Basili, A. Lenci & B. Magnini, Pisa, Italy.

Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. 2014. *Performance Evaluation Report of Europeana Newspapers*, *A Gateway to European Newspapers Online*, D3.5, http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D3.5_Performance_Evaluation_Report_1.0.pdf

Martin Reynaert. 2014. On OCR ground truths and OCR post-correction gold standards, tools and formats. In *Proceedings of Digital Access to Textual Cultural Heritage, Datech 2014*. New York: ACM, 159-166.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Sieglinde Strickner. 2011. *Nachnamen in Südtirol 2010*. Autonome Provinz Bozen-Südtirol, Landesinstitut für Statistik – ASTAT. Bozen 2011.

Katrin Wisniewski, Andrea Abel, and Verena Lyding. 2015. The MERLIN platform: exploring CEFR-related learner texts. Software demo presented at the *Third Learner Corpus Research Conference*, Nijmegen, 11-13 Sept. 2015, In *LCR 2015 Book of Abstracts*, 172-174.

Generare messaggi persuasivi per una dieta salutare

Alessandro Mazzei
Università degli Studi di Torino
Corso Svizzera 185, 10149 Torino
mazzei@di.unito.it

Abstract

English. In this paper we consider the possibility to automatically generate persuasive messages in order to follow a healthy diet. We describe a simple architecture for message generation based on *templates*. Moreover, we describe the influence of some theories about persuasion on the message design.

Italiano. *In questo lavoro si considera la possibilità di generare automaticamente dei messaggi persuasivi affinché degli utenti seguano una dieta salutare. Dopo aver descritto una semplice architettura per la generazione dei messaggi basata su template, si considera la relazione tra il design dei messaggi e alcune teorie della persuasione.*

1 Introduzione

MADiMAN (Multimedia Application for DIet MANagement) è un progetto che studia la possibilità di applicare l'intelligenza artificiale nel contesto della dieta alimentare. L'idea progettuale è realizzare un *dietista virtuale* che aiuti le persone a seguire una dieta salutare. Sfruttando l'ubiquità dei dispositivi mobili si vuole costruire un sistema di intelligenza artificiale che permetta (1) di recuperare, analizzare, conservare i valori nutritivi di una specifica ricetta, (2) di controllarne la compatibilità con la dieta che si sta seguendo e (3) di persuadere l'utente a fare la scelta migliore rispetto a questa dieta.

Nell'ipotetico scenario applicativo, l'interazione tra uomo e cibo è mediata da un sistema artificiale che, sulla base di vari fattori, incoraggia o scoraggia l'utente a mangiare uno specifico piatto. I fattori che il sistema deve considerare sono: la dieta che si intende seguire, il cibo che è stato

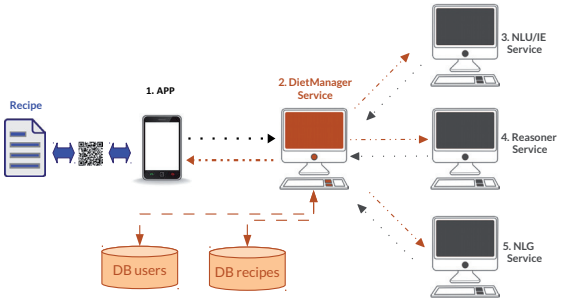


Figura 1: L'architettura di MADiMAN.

mangiato nei giorni precedenti, i valori nutrizionali dello specifico piatto che si vuole scegliere. L'architettura applicativa che il progetto vuole realizzare è un sistema di *web services* (Fig. 1) che interagisca con l'utente mediante una APP (Fig. 1-1.), analizzi il contenuto di una specifica ricetta mediante un modulo di *information retrieval* (Fig. 1-3.), ragioni mediante un modulo di ragionamento automatico (Fig. 1-4.) e, sulla base del ragionamento, generi un messaggio *persuasivo* per convincere l'utente a fare la scelta migliore usando un modulo di generazione automatica del linguaggio naturale (*NLG*, Fig. 1-5.).

Il trattamento automatico del linguaggio entra in gioco nella fase di analisi della ricetta (Mazzei, 2014), così come nella generazione del messaggio persuasivo. In particolare, il sistema di generazione dei messaggi deve usare come input l'output del ragionatore. Allo stato attuale dello sviluppo del progetto, il ragionatore è un sistema basato sulla teoria dei *Simple Temporal Problems*, che produce una costante (*I1, I2, C1, C2, C3*. *I*: incompatibile, *C*: compatibile) insieme a una semplice spiegazione del risultato (*IPO vs. IPER*) (Anselma et al., 2014). Ad esempio, alla fine del ra-

gionamento, un piatto può risultare incompatibile con una dieta perché ha dei valori delle proteine troppo bassi (*I1+IPO*) o troppo alti (*I1+IPER*); oppure un piatto può risultare compatibile seppur iper-proteico (*C2+IPER*) ma, nel caso venga scelto, dovrà essere bilanciato scegliendo nel futuro piatti ipo-proteici.

Il presente lavoro è strutturato come segue: nella Sezione 2 si descriverà il modulo di generazione del linguaggio, nella Sezione 3 si prenderanno in rassegna alcune teorie della persuasione che hanno ispirato il design del modulo di generazione, mentre nella Sezione 4 si concluderà il lavoro con alcune considerazioni sullo stato del progetto e sui futuri sviluppi.

2 Una semplice architettura di NLG

Un'architettura che si è affermata come standard per la generazione del linguaggio naturale prevede tre moduli distinti: il Document Planning , il micro-planning e la Surface Realization (Reiter and Dale, 2000).

Nel document planning si decide cosa dire (i contenuti informativi) e come strutturare il discorso (la struttura retorica). Nel micro-planning, il focus riguarda la progettazione di una serie di caratteristiche che riguardano il linguaggio e i contenuti linguistici delle frasi che saranno generate (ad esempio, usare una frase attiva o una passiva). Nella surface realization le frasi vengono infine prodotte considerando gli specifici vincoli morfo-sintattici della specifica lingua.

Poiché l'ingresso del generatore è costituito dall'output del ragionatore, ovvero da una costante (*I1, I2, C1, C2, C3*) e simbolico per indicare la direzione della deviazione (*IPO vs. IPER*), l'input contiene già la selezione delle informazioni riguardanti cosa dire e quindi nell'architettura di MADiMAN la pianificazione del documento viene fatta a tutti gli effetti dal ragionatore (Reiter, 2007).

La scelta più semplice, da noi adottata, per poter implementare il micro-planning e la surface generation è quella di usare un sistema basato su template. Esiste un'accesa discussione nel campo dell'NLG su cosa esattamente sia la generazione basata su template o meglio su cosa non lo sia (Van Deemter et al., 2005). Nel contesto di MADiMAN, immaginiamo di avere un certo numero di frasi prototipali che possono essere usate contestualmente al risultato del generatore. Nella Tabella 1 elenchiamo alcuni dei possibili output del generatore: le parti sottolineate sono le parti variabili del messaggio che variano di volta in volta a seconda degli output del ragionatore.

Il sistema MADiMAN usa cinque template per poter comunicare i cinque casi principali prodotti in output dal ragionatore: nella tabella 1 sono riportati l'output del ragionatore (colonna **Class**), la direzione della deviazione (colonna **Dev**) e il messaggio generato (colonna **Template**). Lo schema generale seguito per costruire un messaggio è che questo deve contenere (i) una risposta diretta (es. *Ora non puoi mangiare questo piatto*), con eventualmente (ii) una spiegazione (es. *perché è poco proteico*) e, eventualmente anche (iii) un suggerimento (es. *Ma se domenica mangi un bel piatto di fagioli allora lunedì potrai mangiarlo.*). Nella prossima sezione si spiegheranno le motivazioni che ci hanno portato a questi specifici templates.

Per semplicità di esposizione, non descriviamo qui il l'algoritmo usto in generazione per combinare il risultato sui tre distinti macro-nutrienti, cioè proteine, lipidi e carboidrati. In breve, i messaggi sui tre macro-nutrienti devono essere *aggregati* in un solo messaggio, rispettando una serie di vincoli che riguardano la coordinazione delle singole frasi (Reiter and Dale, 2000).

3 Teorie della persusione per la generazione del linguaggio

Nell'ottica di generare messaggi persuasivi, abbiamo considerato tre approcci alla teoria della persuasione presentati negli ultimi anni (Reiter et al., 2003; Fogg, 2002; Guerini et al., 2007).

Il primo approccio è stato ideato nel progetto di un sistema di generazione automatica del linguaggio chiamato STOP, per generare delle lettere che inducano il lettore a smettere di fumare (Reiter et al., 2003). In STOP, il sistema persuasivo è basato essenzialmente sul riconoscimento di tipi-utente (*tailoring*). L'idea di base è di far compilare un questionario ad ogni utente e, sulla base delle risposte, individuare un profilo utente specifico e una serie di informazioni chiave per poi generare, grazie a queste informazioni, delle lettere sulla base di template. Questo tipo di approccio diretto quanto semplice alla persuasione non ha dato i risultati sperati. La sperimentazione basata anche sull'uso di un gruppo di controllo ha evidenziato che l'efficacia della personalizzazione era trascurabile. Secondo gli autori l'inefficacia potrebbe essere ricondotta all'uso di un canale di comunica-

Class	Dev	Template
I.1	IPO	Questo piatto non va affatto bene, contiene davvero pochissime proteine!
I.2	IPO	Ora non puoi mangiare questo piatto perché è poco proteico. Ma se domenica mangi un bel piatto di fagioli allora lunedì potrai mangiarlo.
C.1	IPO	Va bene mangiare le patatine ma nei prossimi giorni dovrai mangiare più proteine.
C.2	IPO	Questo piatto va bene, è solo un po' scarso di proteine. Nei prossimi giorni anche fagioli però! :)
C.3	-	Ottima scelta! Questo piatto è perfetto per la tua dieta :)

Tabella 1: I 5 templates per i messaggi persuasivi (colonna **Template**): la sottolineatura denota le parti variabili nel template. La colonna **Class** contiene la classificazione prodotta dal ragionatore, mentre la colonna **Dev** contiene la direzione della deviazione: *IPO* (*IPER*) indica che il piatto è scarso (ricco) nel valore di uno specifico macro-nutriente.

zione non adatto, ovvero l'invio all'utente di una singola email al giorno.

Allo stato attuale del progetto, la possibilità di creare messaggi personalizzati non è stata presa in analisi in MADiMAN, ma come evidenziato da esperienze simili (es. *myFoodPhone*, vedi sotto), la personalizzazione del feedback rende generalmente un'applicazione più efficiente. A tal proposito, un sistema di tailoring più vicino alle tematiche di MADiMAN viene descritto in (Kaptein et al., 2012): l'idea è quella di spedire messaggi SMS per ridurre il consumo di snack degli utenti. In questo specifico progetto, i messaggi contenuti negli SMS rispettano alcuni degli schemi di persuasione definiti nella teoria generale sulla persuasione di Cialdini (Cialdini, 2009). Le sei strategie descritte sono: *reciprocity*, "people feel obligated to return a favor"; *scarcity*, "people will value scarce products"; *authority*, "people value the opinion of experts", *consistency*, "people do as they said they would"; *consensus*, "people do as other people do"; *liking*, "we say yes to people we like". Rispetto a questa catalogazione, possiamo notare che i messaggi definiti nel sistema di generazione di MADiMAN appartengono essenzialmente alle categorie *authorithy, consistency* e *liking* (vedi messaggi della Tabella 1).

Il secondo approccio alla persuasione non riguarda direttamente la linguistica computazionale, ma è più legato alla psicologia e al design industriale (Fogg, 2002). Fogg è uno psicologo dell'Università di Stanford, dove dirige il laboratorio di CAPTOLOGY (*computers as persuasive technologies*). La CAPTOLOGY è lo studio di come il computer possa essere usato per persuadere un utente a seguire un certo comportamento. È l'approccio di Fogg quello seguito nell'ideazione delle frasi prototipo nel servizio di generazione di MADiMAN. Il punto di partenza della teoria di Fogg è che il computer viene percepito dagli utenti in tre forme coesistenti: Tool-Media-SocialActor, e ognuna di queste tre forme può esercitare una qualche forma di persuasione. Come tool il computer può potenziare le capacità di un utente: nel caso di MADiMAN i calcoli sul contenuto nutritivo dei piatti potenzia le capacità di potere giudicare correttamente la compatibilità di un piatto con una dieta. Come media il computer "fornisce esperienza": nel caso di MADiMAN la memoria umana viene potenziata dal sistema di ragionamento, che indirettamente gli ricorda cosa ha mangiato negli ultimi giorni. Come socialActor il computer crea una relazione empatica con l'utente richiamandolo alle "regole sociali". Nel caso di MADiMAN, i messaggi guidano l'utente verso la scelta di un'alimentazione bilanciata invitandolo a seguire la dieta che egli stesso ha deciso. Parlando di persuasione positiva, ovvero di sistemi software che migliorano in maniera indubbiamente positiva lo stile di vita delle persone, Fogg fa riferimento ad una applicazione, chiamata MyFoodPhone, che ha diverse analogie con MADiMAN: "An example of a positive technology is a mobile application called MyFoodPhone. While mobile persuasive devices have not been studied rigorously, they have several unique properties that may improve their abilities to persuade. First, they are personal devices: people carry their mobile phones everywhere, customize them, and store personal information in them. Second, intrinsic to them being mobile, these devices have the potential to intervene at the right moment, a concept called kairos" (Fogg, 2003). I punti cruciali che MyFoodPhone ha in comune con MADiMAN sono l'ubiquità dei dispositivi mobili e la possibilità di intervenire nel momento giusto (*kairo*). Ancora lo stesso Fogg enuncia delle regole per progettare dei sistemi che siano efficacemente persuasivi (Fogg, 2009), ed alcune di queste regole sono applicabili in MADiMAN. Ad esempio, la regola "Learn what is preventing the target behaviour" chiede di classificare le cause del comportamento scorretto degli

utenti in una delle tre categorie: "lack of motivation, lack of ability, lack of a well-timed trigger to perform the behaviour". Nel contesto MADiMAN tutte e tre le tipologie di cause entrano in gioco: un utente segue una dieta scorretta perché non è abbastanza motivato, perché non sa che il piatto che sta per mangiare è in contrasto con la dieta che sta seguendo, perché non ha lo stimolo giusto nel momento della scelta del piatto. Il sistema di ragionamento e generazione del messaggio lavora proprio su questi due ultimi elementi: il ragionatore potenzia le capacità dell'utente mettendolo in grado di avere le informazioni salienti al momento giusto, il sistema di generazione crea uno stimolo motivazionale nel preciso momento in cui è davvero necessario, ovvero quando bisogna decidere cosa mangiare.

Un approccio alla persuasione, distante da quello di Fogg ma più legato alle tematiche e tecnologie dell'intelligenza artificiale, è quello che si basa sul concetto di computer come agente intelligente (Hovy, 1988; De Rosis and Grasso, 2000; Guerini et al., 2007; Guerini et al., 2011). Il sistema si comporta a tutti gli effetti come un'entità autonoma, spesso modellata attraverso la specifica BDI (*Beliefs, Desires, Intentions*), il cui scopo principale è persuadere l'utente a comportarsi in una specifica maniera. È evidente come un approccio di questo tipo è più vicino alla ricerca sulla persuasione in un contesto di agenti artificiali e umani che interagiscono, piuttosto che alla sua applicazione pratica immediata. Comunque, questi modelli ad agenti permettono una maggiore flessibilità nelle scelte implementative del sistema di generazione del linguaggio. In contrasto, la scelta da noi fatta in MADiMAN prevede un sistema che unifica il micro-planning e la realizzazione in un unico modulo basato su template. Comunque, l'analisi di un sistema flessibile basato su agenti ci permette alcune riflessioni anche sulle scelte fatte in MADiMAN. Hovy definisce una serie di regole euristiche che legano il livello argomentale, definito nel processo di sentence planning. Ad esempio: "Adverbial stress words can only be used to enhance or mitigate expressions that carry some affect already" (Hovy, 1988). Nei messaggi definiti in MADiMAN questa regola è stata applicata in un certo numero di occasioni, come ad esempio "Nei prossimi giorni un po' meno proteine quindi! :)". De Rosis e Grasso definiscono delle regole euristiche sulla struttura argomentale per enfatizzare o mitigare lessicalmente il messaggio. L'uso di alcuni avverbi, come *little bit*, *very*, *really*, sono usate contestualmente ad alcune specifiche strutture argomentali (De Rosis and Grasso, 2000). L'uso di queste parole nei messaggi definiti in MADiMAN seguono spesso queste costruzioni. Guerini et al. definiscono un'architettura per la persuasione molto dettagliata, in cui la pianificazione dell'agente parte dalla strategia persuasiva da adottare e definisce la struttura retorica che il messaggio adotterà in fase di generazione (Guerini et al., 2007). Rispetto alla tassonomia di strategie proposte, possiamo notare come MADiMAN adotti unicamente la strategia *action_inducement/goal_balance/ positive_consequence*, ovvero una strategia che induca un'azione (scegliere un piatto), usando i goal dell'utente (una dieta bilanciata), usando i benefici della scelta del piatto giusto.

Le possibilità di persuasione dei canali multimediali sono ancora in una fase di sperimentazione, ma alcuni risultati sono già direttamente applicabili nel contesto di MADiMAN. Come oramai attestato da alcuni studi l'uso delle emoticons nei testi scritti può aumentare l'efficacia comunicativa del messaggio. Ad esempio (Derks et al., 2008) dimostra che l'uso delle emoticons dà un tono di tipo amicale al messaggio e può aumentarne il valore positivo. Nel contesto di MADiMAN abbiamo considerato questo studio nell'inserire le emoticons nei messaggi relativi alle situazioni in cui l'utente scegliendo il piatto analizzato fa proprio la scelta giusta.

4 Conclusioni e sviluppi futuri

In questo lavoro abbiamo descritto le principali caratteristiche di un sistema di generazione di messaggi con intenti persuasivi nel contesto della dieta alimentare.

Attualmente, per poter verificare quantitativamente la bontà dell'approccio proposto, prevediamo di seguire due modalità sperimentali distinte. Inizialmente, stiamo realizzando una simulazione che tenga conto dei vari fattori che influenzano il successo del nostro sistema. Da un lato è necessario modellare la propensione dell'utente a essere persuaso, dall'altro è necessario considerare dei valori numerici sensati per modellare la dieta e i piatti. Se la simulazione darà risultati promettenti, intendiamo successivamente valutare il sistema in un trial medico realistico. Seguendo il modello

valutativo proposto da Reiter per il sistema STOP (Reiter et al., 2003), intendiamo testare il sistema mediante gruppi di controllo in un contesto medico specifico, cioè quello delle cliniche per trattare l'obesità essenziale.

Ringraziamenti

Questo lavoro è stato supportato dal progetto MADiMAN, parzialmente finanziato dalla Regione Piemonte, Innovation Hub for ICT, POR FESR 2007/2013 - Asse I - Attività I.1.3. `http://di.unito.it/madiman`

References

Luca Anselma, Alessandro Mazzei, Luca Piovesan, and Franco De Michieli. 2014. Adopting STP for diet management. In *Proc. of IEEE International Conference on Healthcare Informatics*, page 371, September.

Robert B. Cialdini. 2009. *Influence : science and practice*. Pearson Education, Boston.

Fiorella De Rosis and Floriana Grasso. 2000. Affective natural language generation. *Affective interactions*, pages 204–218.

Daantje Derks, Arjan E. R. Bos, and Jasper von Grumbkow. 2008. Emoticons in computer-mediated communication: Social motives and social context. *Cyberpsy., Behavior, and Soc. Networking*, 11(1):99–101.

B.J. Fogg. 2002. *Persuasive Technology. Using computers to change what we think and do*. Morgan Kaufmann Publishers, Elsevier, San Francisco.

B. J. Fogg. 2003. Motivating, Influencing, and Persuading Users. In Julie A. Jacko and Andrew Sears, editors, *The Human-computer Interaction Handbook*, chapter Motivating, Influencing, and Persuading Users, pages 358–370. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

B.J. Fogg. 2009. The new rules of persuasion. *RSA Digital Journal*, page 1:4, Summer. Online.

Marco Guerini, Oliviero Stock, and Massimo Zancanaro. 2007. A taxonomy of strategies for multimodal persuasive message generation. *Applied Artificial Intelligence*, 21(2):99–136.

Marco Guerini, Oliviero Stock, Massimo Zancanaro, Daniel J. O'Keefe, Irene Mazzotta, Fiorella Rosis†, Isabella Poggi, Meiyii Y. Lim, and Ruth Aylett. 2011. Approaches to Verbal Persuasion in Intelligent User Interfaces. In Roddy Cowie, Catherine Pelachaud, and Paolo Petta, editors, *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pages 559–584. Springer.

Eduard H. Hovy. 1988. *Generating Natural Language Under Pragmatic Constraints*. Lawrence Erlbaum, Hillsdale, NJ.

Maurits Kaptein, Boris De Ruyter, Panos Markopoulos, and Emile Aarts. 2012. Adaptive persuasive systems: A study of tailored persuasive text messages to reduce snacking. *ACM Trans. Interact. Intell. Syst.*, 2(2):10:1–10:25, June.

Alessandro Mazzei. 2014. On the lexical coverage of some resources on italian cooking recipes. In *Proc. of CLiC-it 2014, First Italian Conference on Computational Linguistics*, pages 254–259, December.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Ehud Reiter, Roma Robertson, and Liesl M. Osman. 2003. Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence*, 144:41–58.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pages 97–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kees Van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Real versus template-based natural language generation: A false opposition? *Comput. Linguist.*, 31(1):15–24, March.

FacTA: Evaluation of Event Factuality and Temporal Anchoring

Anne-Lyse Minard[1], Manuela Speranza[1], Rachele Sprugnoli[1-2], Tommaso Caselli[3]
[1]Fondazione Bruno Kessler, Trento
[2]Università di Trento
[3]VU Amsterdam
{minard,manspera,sprugnoli}@fbk.eu
t.caselli@vu.nl

Abstract

English. In this paper we describe FacTA, a new task connecting the evaluation of factuality profiling and temporal anchoring, two strictly related aspects in event processing. The proposed task aims at providing a complete evaluation framework for factuality profiling, at taking the first steps in the direction of narrative container evaluation for Italian, and at making available benchmark data for high-level semantic tasks.

Italiano. *Questo articolo descrive FacTA, un nuovo esercizio di valutazione su fattualità ed ancoraggio temporale, due aspetti dell'analisi degli eventi strettamente connessi tra loro. Il compito proposto mira a fornire una cornice completa di valutazione per la fattualità, a muovere i primi passi nella direzione della valutazione dei contenitori narrativi per l'italiano e a rendere disponibili dati di riferimento per compiti semantici di alto livello.*

1 Introduction

Reasoning about events plays a fundamental role in text understanding; it involves different aspects, such as event identification and classification, temporal anchoring of events, temporal ordering, and event factuality profiling. In view of the next EVALITA edition (Attardi et al., 2015),[1] we propose FacTA (*Factuality and Temporal Anchoring*), the first task comprising the evaluation of both factuality profiling and temporal anchoring, two strictly interrelated aspects of event interpretation.

Event factuality is defined in the literature as the level of committed belief expressed by relevant sources towards the factual status of events mentioned in texts (Saurí and Pustejovsky, 2012). The notion of factuality is closely connected to other notions thoroughly explored by previous research conducted in the NLP field, such as subjectivity, belief, hedging and modality; see, among others, (Wiebe et al., 2004; Prabhakaran et al., 2010; Medlock and Briscoe, 2007; Saurı et al., 2006). More specifically, the factuality status of events is related to their degree of certainty (from absolutely certain to uncertain) and to their polarity (affirmed *vs.* negated). These two aspects are taken into consideration in the factuality annotation frameworks proposed by Saurí and Pustejovsky (2012) and van Son et al. (2014), which inspired the definition of factuality profiling in FacTA.

Temporal anchoring consists of associating all temporally grounded events to time anchors, i.e. temporal expressions, through a set of temporal links. The TimeML annotation framework (Pustejovsky et al., 2005) addresses this issue through the specifications for temporal relation (TLINK) annotation, which also implies the ordering of events and temporal expressions with respect to one another. Far from being a trivial task (see systems performance in English (UzZaman et al., 2013) and in Italian (Mirza and Minard, 2014)), TLINK annotation requires the comprehension of complex temporal structures; moreover, the number of possible TLINKs grows together with the number of annotated events and temporal expressions. Pustejovsky and Stubbs (2011) introduced the notion of *narrative container* with the aim of reducing the number of TLINKs to be identified in a text while improving informativeness and accuracy.

A narrative container is a temporal expression or an event explicitly mentioned in the text into which other events temporally fall (Styler IV et al., 2014). The use of narrative containers proved to be useful to accurately place events on timelines in the domain of clinical narratives (Miller et al., 2013). Temporal anchoring in FacTA moves in the direction of this notion of narrative container by fo-

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

cusing on specific types of temporal relations that link an event to the temporal expression to which it is anchored. However, anchoring events in time is strictly dependent of their factuality profiling. For instance, counterfactual events will never have a temporal anchor or be part of a temporal relation (i.e. they never occurred); this may not hold for speculated events, whose association with a temporal anchor or participation in a temporal relation is important to monitor future event outcomes.

2 Related Evaluation Tasks

Factuality profiling and temporal anchoring of events are crucial for many NLP applications (Wiebe et al., 2005; Karttunen and Zaenen, 2005; Caselli et al., 2015) and therefore have been the focus, either direct or indirect, of several evaluation exercises, especially for English.

The ACE Event Detection and Recognition tasks of 2005 and 2007 (LDC, 2005) took into consideration factuality-related information by requiring systems to assign the value of the *modality* attribute to extracted events so as to distinguish between asserted and non-asserted (e.g. hypothetical, desired, and promised) events. Following the ACE evaluation, a new task has recently been defined in the context of the TAC KBP 2015 Event Track.[2] The Event Nugget Detection task aims at assessing the performance of systems in identifying events and their *realis* value, which can be ACTUAL, GENERIC or OTHER (Mitamura et al., 2015). Other tasks focused on the evaluation of speculated and negated events in different domains such as biomedical data and literary texts (Nédellec et al., 2013; Morante and Blanco, 2012).

The evaluation of event modality was part of the Clinical TempEval task at SemEval 2015 (Bethard et al., 2015),[3] which also proposed for the first time the evaluation of narrative container relations between events and/or temporal expressions.

Temporal anchoring has been evaluated in the more general context of temporal relation annotation in the 2007, 2011 and 2013 TempEval evaluation exercises (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013) as well as in the EVENTI task (Caselli et al., 2014) on Italian at EVALITA 2014. The TimeLine task at SemEval 2015 (Minard et al., 2015) was the first evaluation exercise focusing on cross-document event ordering; in view of the creation of timelines, it requires temporal anchoring and ordering of certain and non-negated events.

With respect to the aforementioned tasks, FacTA aims at providing a complete evaluation framework for factuality profiling, at taking the first steps in the direction of narrative container evaluation for Italian, and at making new datasets available to the research community.

3 Task Description

The FacTA task consists of two subtasks: factuality profiling and temporal anchoring of given gold event mentions. Participants may decide to take part to both or only one of the proposed subtasks.

3.1 Subtask 1: Factuality Profiling

Tonelli et al. (2014) propose an annotation schema of factuality for English based on the annotation framework by van Son et al. (2014).[4] This schema was then adapted to Italian by Minard et al. (2014). Following this, we represent factuality by means of a combination of three attributes associated with event mentions: certainty, time, and polarity. For each given gold event mention, participant systems are required to assign values for three factuality attributes.

The *certainty* attribute relates to how sure the main source is about the mentioned event[5] and admits the following four values: `certain`, `possible`, `probable`, and `underspecified`.

The *time* attribute specifies the time when an event is reported to have taken place or to be going to take place. Its values are `non_future` (for present and past events), `future` (for events that will take place), and `underspecified`.

The *polarity* attribute captures if an event is affirmed or negated and, consequently, it can be either `positive` or `negative`; when there is not enough information available to detect the polarity of an event mention, it is `underspecified`.

[2] `http://www.nist.gov/tac/2015/KBP/Event/index.html`

[3] Systems were required to distinguish actual, hedged, hypothetical and generic events.

[4] van Son et al.'s annotation framework, inspired by FactBank (Saurí and Pustejovsky, 2009), enriches it with the distinction between future and non-future events.

[5] The main source is either the utterer (in direct speech, indirect speech or reported speech) or the author of the news (in all other cases). In this framework, where factuality depends strictly on the source, factuality annotation is also referred to as attribution annotation.

Factuality value. The combination of the attributes described above determines the value of an event: `factual`, `counterfactual` or `non_factual`. More specifically, the overall factuality value is factual if its values are certain, non_future, and positive (e.g. 'rassegnato' in [1]), while it is counterfactual (i.e. the event is reported as not having taken place) if its values are certain, non_future, and negative (e.g. 'nominato' in [2]). In any other combination, the event is non factual, either because it is non certain, or future (e.g. 'nomineranno' in [1]).

(1) *Smith ha rassegnato ieri le dimissioni; nomineranno il suo successore entro un mese.* ("Smith resigned yesterday; they will appoint his replacement within a month.")

(2) *Non ha nominato un amministratore delegato.* ("He did not appoint a CEO.")

No factuality annotation. Language is used to describe events that do not correlate with a real situation in the world (e.g. 'parlare' in [3]). For these event mentions participant systems are required to leave the value of all three attributes empty.

(3) *Guardate, penso che sia prematuro parlare del nuovo preside* ("Well, I think it is too early to talk about the new dean")

3.2 Subtask 2: Temporal Anchoring

Given a set of gold events, participant systems are required to detect those events for which it is possible to identify a time anchor. Our definition of time anchor includes two different types of elements: the temporal expressions occurring in the text, as well as the Document Creation Time (DCT), which is part of the metadata associated with each document. The subtask thus includes temporal expression (or TIMEX3) detection and normalization,[6] as well as identification of temporal relations (or TLINKs) between events and temporal expressions.

TIMEX3 detection and normalization. Based on the annotation guidelines produced within the NewsReader project (Tonelli et al., 2014), which in turn are based on the ISO-TimeML guidelines (ISO TimeML Working Group, 2008), this consists of:

- TIMEX3 detection: identification and classification of temporal expressions of type date and time (durations and sets of times, on the other hand, are excluded from the task).
- TIMEX3 normalization: identification of the *value* attribute for each temporal expression.

For instance, in [1], *ieri* is a TIMEX3 of type date with value *2015-07-28* considering *2015-07-29* as DCT.

TLINK identification. This consists of detecting TLINKs of types IS_INCLUDED and SIMULTANEOUS holding between an event and a TIMEX3 (i.e. the anchor of the event), as defined in (Tonelli et al., 2014). The event (the source of the TLINK) and the TIMEX3 (the target) can either appear in the same sentence or in different sentences. For instance, in [1], *rassegnato* is anchored to *ieri* (*rassegnato*, IS_INCLUDED, *ieri*).

4 Dataset Description

4.1 Subtask 1: Factuality Profiling

As a training dataset, participants can use Fact-Ita Bank (Minard et al., 2014), which consists of 170 documents selected from the Ita-TimeBank (Caselli et al., 2011), which was first released for the EVENTI task at EVALITA 2014.[7] Fact-Ita Bank contains annotations for 10,205 event mentions and is already distributed with a CC-BY-NC license.[8]

System evaluation will be performed on the "first five sentences" section of WItaC, the NewsReader Wikinews Italian Corpus (Speranza and Minard, 2015).[9] It consists of 15,676 tokens and has already been annotated with event factuality (as this annotation has been projected from English, it will need some minor revision).

4.2 Subtask 2: Temporal Anchoring

For temporal expression detection and normalization, participant systems can be trained on the dataset used for the EVENTI Task at Evalita 2014 (Caselli et al., 2014). It also contains TLINKs between events and TIMEX3s in the same sentence but not in different sentences. To make it usable as a training corpus for temporal anchoring, we would have to add the TLINKs between events and

[6] Here, and in the remainder of the paper, we are not distinguishing between the two types of elements and we refer to them simply as temporal expressions or TIMEX3s.

[7] https://sites.google.com/site/eventievalita2014/home

[8] http://hlt-nlp.fbk.eu/technologies/fact-ita-bank

[9] The reason for selecting the first sentences was to maximise the number of articles in the corpus, while at the same time including the most salient information.

TIMEX3s in different sentences and the TLINKs between events and the DCT, which would require a big effort. Thus, we are instead planning to add the needed relations to only a subset of the corpus, namely the same 170 documents that compose Fact-ItaBank.

As test data we will use the "first five sentences" section of WItaC (Speranza and Minard, 2015), which is already annotated with TIMEX3s and with TLINKs between events and TIMEX3s in the same sentences;[10] the test set thus needs to be completed through the addition of TLINKs between events and TIMEX3s in different sentences.

5 Evaluation

Each subtask will be evaluated independently. No global score will be computed as the task aims to isolate the two phenomena.

5.1 Subtask 1: Factuality Profiling

Participant systems will be evaluated in terms of precision, recall and their harmonic mean (i.e. F1 score). We will perform the evaluation of:

- values of the factuality attributes (polarity, certainty and time);
- detection of events to which factuality values should not be assigned (i.e. "no factuality annotation" events);
- assignment of the overall factuality value (combination of the three attributes), including also the non-assignment of factuality attributes.

The official ranking of the systems will be based on the evaluation of the overall factuality value.

5.2 Subtask 2: Temporal Anchoring

For the temporal anchoring subtask, we will evaluate the number of event-TIMEX3 relations correctly identified in terms of precision, recall and F1 score. Two relations in the reference and the system prediction match if their sources and their targets match. Two sources (i.e. events) are considered as equivalent if they have the same extent, whereas two targets (i.e. TIMEX3s) match if their values are the same. Participant systems will be ranked according to the F1 score.

We will not apply the metric for evaluating temporal awareness based on temporal closure graphs proposed by UzZaman and Allen (2011), which is unnecessarily complex as we have reduced the relations to only IS_INCLUDED and SIMULTANEOUS.

6 Discussion and Conclusions

The FacTA task connects two related aspects of events: factuality and temporal anchoring. The availability of this information for Italian will both promote research in these areas and fill a gap with respect to other languages, such as English, for a variety of semantic tasks.

Factuality profiling is a challenging task aimed at identifying the speaker/writers degree of commitment to the events being referred to in a text. Having access to this type of information plays a crucial role for distinguishing relevant and non-relevant information for more complex tasks such as textual entailment, question answering, and temporal processing.

On the other hand, anchoring events in time requires to interpret temporal information which is not often explicitly provided in texts. The identification of the correct temporal anchor facilitates the organization of events in groups of narrative containers which could be further used to improve the identification and classification of in-document and cross-document temporal relations.

The new annotation layers will be added on top of an existing dataset, the EVENTI corpus, thus allowing to re-use existing resources and to promote the development of multi-layered annotated corpora; moreover a new linguistic resource, WItaC, will be provided. The availability of these data is to be considered strategic as it will help the study the interactions of different language phenomena and enhance the development of more robust systems for automatic access to the content of texts. The use of well structured annotation guidelines grounded both on official and *de facto* standards is a stimulus for the development of multilingual approaches and promote discussions and reflections in the NLP community at large.

Considering the success of evaluation campaigns such as Clinical TempEval at SemEval 2015 and given the presence of an active community focused on extra-propositional aspects of meanings (e.g. attribution[11]), making available new annotated data in the framework of an evaluation campaign for a language other than English can have a large impact in the NLP community.

[10] This also includes TLINKs between events and the DCT.

[11] Ex-Prom Workshop at NAACL 2015 `http://www.cse.unt.edu/exprom2015/`

Acknowledgements

This work has been partially supported by the EU NewsReader Project (FP7-ICT-2011-8 grant 316404) and the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3).

References

Giuseppe Attardi, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli. 2015. State of the Art Language Technologies for Italian: The EVALITA 2014 Perspective. *Intelligenza Artificiale*, 9(1):43–61.

Steven Bethard, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics*.

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 *Linguistic Annotation Workshop*, pages 143–151.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Tommaso Caselli, Antske Fokkens, Roser Morante, and Piek Vossen. 2015. SPINOZA VU: An NLP Pipeline for Cross Document TimeLines. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

ISO TimeML Working Group. 2008. ISO TC37 draft international standard DIS 24617-1, August 14. `http://semantic-annotation.uvt.nl/ISO-TimeML-08-13-2008-vankiyong.pdf`.

Lauri Karttunen and Annie Zaenen. 2005. Veridicity. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, extracting and reasoning about time and events*, Dagstuhl, Germany.

LDC. 2005. ACE (Automatic Content Extraction) English Annotation Guidelines for Events. In *Technical Report*.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *ACL*, volume 2007, pages 992–999. Citeseer.

Timothy A Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana K Savova. 2013. Discovering narrative containers in clinical text. *ACL 2013*, page 18.

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 2014, First Italian Conference on Computational Linguistic*.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, Rubén Urizar, and Fondazione Bruno Kessler. 2015. SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics.

Paramita Mirza and Anne-Lyse Minard. 2014. FBK-HLT-time: a complete Italian Temporal Processing system for EVENTI-EVALITA 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Teruko Mitamura, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event Nugget Annotation: Processes and Issues. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 66–76.

Roser Morante and Eduardo Blanco. 2012. * SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 265–274. Association for Computational Linguistics.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1014–1022. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

James Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. *The specification language TimeML*, pages 545–557.

Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Roser Saurı, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proceedings of 19th International FLAIRS Conference*.

Manuela Speranza and Anne-Lyse Minard. 2015. Cross-language projection of multilayer semantic annotation in the NewsReader Wikinews Italian Corpus (WItaC). In *Proceedings of CLiC-it 2015, Second Italian Conference on Computational Linguistic*.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. 2014. NewsReader Guidelines for Annotation at Document Level. Technical Report NWR2014-2-2, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf`.

Naushad UzZaman and James Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of SemEval 2013*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Chantal van Son, Marieke van Erp, Antske Fokkens, and Piek Vossen. 2014. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: TempEval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

TED-MWE: a bilingual parallel corpus with MWE annotation

Towards a methodology for annotating MWEs in parallel multilingual corpora

Johanna Monti[1]*, Federico Sangati[2], Mihael Arcan[3]
[1]Sassari University, Sassari, Italy
[2]Fondazione Bruno Kessler, Trento, Italy
[3]National University of Ireland, Galway, Ireland
jmonti@uniss.it,sangati@fbk.eu,mihael.arcan@insight-centre.org

Abstract

English. The translation of Multiword expressions (MWE) by Machine Translation (MT) represents a big challenge, and although MT has considerably improved in recent years, MWE mistranslations still occur very frequently. There is the need to develop large data sets, mainly parallel corpora, annotated with MWEs, since they are useful both for SMT training purposes and MWE translation quality evaluation. This paper describes a methodology to annotate a parallel spoken corpus with MWEs. The dataset used for this experiment is an English-Italian corpus extracted from the TED spoken corpus and complemented by an SMT output.

Italiano. *La traduzione delle polirematiche da parte dei sistemi di Traduzione Automatica (TA) rappresenta un sfida irrisolta e benché i sistemi abbiano compiuto notevoli progressi, traduzioni errate di polirematiche occorrono ancora molto di frequente. È necessario sviluppare ampie collezioni di dati principalmente corpora paralleli annotati con polirematiche che siano utili sia per l'addestramento della TA di tipo statistico sia per la valutazione della qualità della traduzione delle polirematiche. Questo contributo descrive una metodologia per annotare un corpus parallelo del parlato con le polirematiche e il corpus stesso. La collezione di dati usata per questo esperimento è un corpus inglese-italiano estratto dal TED, corpus del parlato, integrato dalla traduzione di un sistema statistico di TA.*

*Johanna Monti is author of sections 2 and 3.2, Federico Sangati is author of sections 4 and 5, Mihael Arcan is author of sections 3.1 and 4.1. Introduction and conclusions are in common.

1 Introduction

Multiword expressions (MWEs) represent one of the major challenges for all Natural Language Processing (NLP) applications and in particular for Machine Translation (MT) (Sag et al., 2002). The notion of MWE includes a wide and frequent set of different lexical phenomena with their specific properties, such as idioms, compound words, domain specific terms, collocations, Named Entities or acronyms. Their morpho-syntactic, semantic and pragmatic idiomaticity (Baldwin and Kim, 2010) together with translational asymmetries (Monti and Todirascu, 2015), i.e. the differences between an MWE in the source language and its translation, prevent technologies from using systematic criteria for properly handling MWEs. For this reason their automatic identification, extraction and translation are very difficult tasks.

Recent PARSEME surveys[1] have highlighted that there is lack of MWE-annotated resources, and in particular parallel corpora. Moreover, the few available ones are usually limited to the study of specific MWE types and specific language pairs. The focus of our research work is therefore to provide a methodology for annotating a parallel corpus with all MWEs (with no restrictions to a specific type) which can be used both for training and testing SMT systems. We have refined this methodology while developing the English-Italian MWE-TED corpus, which contains 1.5K sentences and 31K EN tokens.It is a subset of the TED spoken corpus annotated with all the MWEs detected during the annotation process. This contribution presents the corpus[2] together with the annotation guidelines in section 3, the annotation process in section 4 and the MWE annotation statistics in section 5.

[1]Translating Multiword Expressions - PARSEME WG3 State of the Art Report - forthcoming

[2]http://tiny.cc/TED_MWE

2 Related work

As mentioned in the previous section, the research work in this field is mainly focused on the annotation of specific MWE types, such as (i) the SzegedParalell English-Hungarian parallel corpus (Vincze, 2012) which contains 1370 occurrences of light verb constructions (LVCs), (ii) 4FX, a quadrilingual parallel corpus annotated manually for LVCs (Rácz et al., 2014) containing 673 LVCs in English, 806 in German, 938 in Spanish and 1059 in Hungarian.

Unlike the above methodologies, our aim is to provide a more general approach to MWE annotation in a parallel and multilingual corpus. In this respect, Schneider et al. (2014) present an interesting *comprehensive annotation approach*, in which all different types of MWEs are annotated in a 55K-word corpus of English web text.

Annotating MWEs in parallel texts involves several problems due to the translational asymmetries between languages and presence of discontinuity, but it is considered very important to compensate for the lack of training and benchmark resources for MT.

There are few corpora specifically built to evaluate MT translation quality with reference to MWE translation, such as (i) Ramisch et al. (2013) where an English-French corpus annotated with Phrasal Verbs (PVs) is used to assess the quality of PV translation by a phrase-based system (PBS) and a hierarchical system (HS) or (ii) Schottmüller and Nivre (2014), who describe a German-English corpus containing Verb-particle constructions (VPCs), used to compare the results obtained from Google Translate and Bing Translate, and finally Barreiro et al. (2013), who use parallel corpora (English to Italian, French, Portuguese, German and Spanish) containing 100 English Support Verb Constructions (SVC) and their translations in the target languages done by OpenLogos and the Google Translate.

3 TED-MWE

3.1 The TED Corpus

We have used the WIT^3 web inventory (Cettolo et al., 2012) which offers access to a collection of transcribed and translated talks. The core of WIT^3 is the TED Talks corpus, that basically redistributes the original content published by the TED Conference website. The WIT^3 corpus repurposes the original TED content in a way which is more convenient for MT researchers. For our experiments we used the WIT^3 data released for the IWSLT 2014 Evaluation Campaign, which contains the training data of 190K parallel sentences, needed to build an SMT system. We base our annotations and analysis on the test set, which we will refer to as the MWE-TED corpus.

3.2 MWE Annotation Guidelines

The judgement of whether an expression should qualify as an MWE relies on the annotation guidelines, which are based on the PARSEME MWE template and the testing of MWE properties.

The PARSEME MWE Template provides information and examples for all different MWE syntactic structures (nominal verbal, adjectival, prepositional, clausal MWEs), the fixedness/flexibility of MWE parts, the different levels of idiomaticity (lexical, syntactic, semantic, pragmatic, statistical idiomaticity) and finally the rhetoric relations within an MWE. In addition to the template, annotators were provided with a set of tests (Monti, 2012) to be used to assess if a certain group of words can be considered as a MWE:

Non-substitutability : one element of the MWE cannot be replaced without a change of meaning or without obtaining a non-sense (*in deep water* → *in hot water*; *gas chamber* → **gas room*);

Non-expandability : insertion of additional elements is not possible (*get a head start* → **get a quick head start*);

Non-reducibility : the elements in the MWE cannot be reduced and pronominalisation of one of the constituents is also not possible (*take advantage* → **what did you take? advantage*; **Did you take it?*;

Non-literal translatability : the meaning cannot be translated literally. The difficulty of a literal translation across cultural and linguistic boundaries is mainly a property of MWEs with limited or no variation of distribution, such as idioms (e.g., *it's raining cats and dogs* → it. **sta piovendo cani e gatti*), but also of many collocations (e.g., *heavy rain* → it. **pioggia pesante*), fixed expressions (e.g., *by and large* → it. **da e largo*), proverbs (e.g., *there's no such thing as a free lunch* → it. **non esiste una cosa come un pranzo gratuito*), phrasal verbs (e.g., *bring somebody down* → it. **Portare qualcuno giù*);

Invariability : Invariability can affect both the morphological and the syntactic level. Inflectional variations of the constituents of the MWEs are not always possible. Invariability affects both the head elements and its modifiers (*fish out of water* → **fishes out of water*; *dead on arrival* → **dead on arrivals*; *in high places* → **in high place*); syntactical variations inside an MWU may also not be acceptable (*credit card* → **card of credit*);

Non-displaceability : displacement and a different order of constituents are not possible (*wild card* → **is wild this card?*) - (*back and forth* → **forth and back*);

Institutionalisation of use : certain word units, even those that are semantically and distributionally "free", are used in a conventional manner. The Italian expression *in tempo reale* (a loan translation of the English expression *in real time*) is an example of this feature since its antonym **in tempo irreale* (**in unreal time*) seems to be unmotivated and not used at all.

In order to consider a certain word unit as an MWE it is sufficient that it shows at least one of the above-mentioned properties. Nevertheless, during the annotation process, the property which turned out to characterise the majority of MWEs is the non-literal translatability.

4 Annotation Process

The annotation was organised in three distinct phases: individual annotation, inter-annotation check, validation.

Individual annotation. During the first phase, thirteen annotators with linguistic background in Italian and English were asked to annotate the 1,529 sentences in the MWE-TED corpus. The sentences were organised in a spreadsheet (see figure 1) containing the following information: (i) the English source text, (ii) the Italian *manual* translations (from the parallel corpus) and finally (iii) the Italian *SMT* output (see section 4.1). The annotators were asked to identify all the MWEs in the source text together with their translations in approximately 300 random sentences each and to evaluate the automatic translation correctness[3]. If the *manual* or the *SMT* generated translations were wrong, the annotators were asked to specify the correct translations.

The annotation took into account all MWE types detected in the source text with no restrictions to a particular type of MWE and in particular, both contiguous and discontinuous MWE types were recorded in the dataset. The MWEs identified during the annotation process were recorded as sequences of tokens with no further information about their internal syntactic structure or semantic features.

Inter-annotation check. In the second phase, each annotator was confronted with the anonymized annotations by the other annotators on his/her annotation subset, in order to decide about his/her choices, i.e. to confirm or change the annotations for each source text/manual/SMT set (see table 1).

Sentence: 369	
Source: people sort of think i went away between " titanic " and " avatar " and was buffing my nails someplace , sitting at the beach .	
Your MWE(s)	[sort of, buffing my nails, someplace]
Ann.10 MWE(s)	[sort of, buffing my nails]
Sentence: 432	
Source: now that 's back from high school algebra , but let 's take a look .	
Your MWE(s)	[back from]
Ann.6 MWE(s)	[take a look]
Sentence: 539	
Source: that 's a key element of making that report card .	
Your MWE(s)	[report card]
Ann.12 MWE(s)	[key element, report card]

Table 1: Annotation phase 2: inter-annotation check.

Validation. Finally, in the last phase, we have randomly selected about half of the annotated sentences (801) and asked the annotators to integrate and resolve the possible annotation conflicts (see figure 2).

4.1 Statistical Machine Translation

In order to gather automatic translations of the source text, we used the Moses toolkit (Koehn et al., 2007), where the word alignments were built with GIZA++ (Och and Ney, 2003). The IRSTLM toolkit (Federico et al., 2008) was used to build the 5-gram language model. The parameters within the SMT system are optimized on the development data set using MERT (Bertoldi et al., 2009). The system performed in line with the state-of-the-art results on the test set.

[3]The annotation work was organised in such a way that each sentence was annotated by at least two annotators

SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	MWE SOURCE TEXT	MWE MANUAL TEXT	MWE MANUAL CHECK (Y/N)	MWE AUTO TEXT	MWE AUTO CHECK (Y/N)
369	people sort of think i went away between " titanic " and " avatar " and was buffing my nails someplace , sitting at the beach .	la gente pensa quasi che me ne sia andato tra " titanic " e " avatar " e che mi stessi girando i pollici seduto su qualche spiaggia .	persone come pensare partii tra " titanic " e " avatar " e fu buffing mie unghie da qualche parte , seduto in spiaggia .	buffing my nails	girando i pollici	Y	buffing mie unghie	N

Figure 1: Annotation phase 1: individual annotation.

SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	ANN #	MWE SOURCE TEXT	MWE MANUAL TEXT	MWE MANUAL CHECK (Y/N)	MWE AUTO TEXT	MWE AUTO CHECK (Y/N)
26	" don , " i said , " just to get the facts straight , you guys are famous for farming so far out to sea , you don 't pollute . "	" don " , gli ho detto " tanto per capire bene , voi siete famosi per fare allevamento così lontano , in mare aperto , che non inquinate . "	" non " , ho detto , " per ottenere i fatti dritto , siete famosa per coltivare così lontano in mare , non inquinante . "						
				3	to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				9	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				13	get...stright	capire bene	Y	per ottenere...dritto	N
				FINAL	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N

Figure 2: Annotation phase 3: validation

English	Italian
pointed at	indicò
no longer	non ... più
don 't get me wrong	non fraintendetemi
got bitten by	sono stato affetto dal
a lot of	un sacco di
in the dead of winter	nella tristezza dell' inverno

Table 2: Sample of annotated MWE EN-IT pairs.

5 MWE Annotation Statistics

After the first two phases of the annotation process, out of 1,529 annotated sentences, 541 (35.9%) showed a good inter-annotation agreement, i.e. at least two annotators completely agreed on the annotations. In total we have collected 2,484 English MWEs types out of which 2,391 (96%) are contiguous and 93 (4%) are discontinuous. At least two annotators agreed for the 27% (671) of the MWEs and in 45% of them (1,115) at least two annotators showed an overlapping (at least one word in common).

This general low agreement scores confirm the difficulty of the annotation task. In order to resolve the numerous annotation conflicts, we ran a third annotation phase in which 801 of the previous sentences were validated. This resulted in a total of 799 English MWE types (931 tokens), of which 729 (91%) are contiguous and the 9% (70) are discontinuous. Most MWEs have length 2 (515) and 3 (261), but there are MWEs up to length 8. In 52% of the cases (471) the annotators have evaluated the automatic translation to be incorrect. Table 2 reports a small sample of annotated English MWEs together with their Italian translations.

6 Conclusions

We have described the TED-MWE corpus, an English-Italian parallel spoken corpus annotated with MWEs, together with the methodology and the guidelines adopted during the annotation process. Ongoing and future work includes refinement of the annotation tools and guidelines, the extension of the methodology to further languages in order to develop a multilingual MWE-TED corpus. The main aim is to provide useful data both for SMT training purposes and MT quality evaluation.

Acknowledgments

We greatly acknowledge the PARSEME IC1207 COST Action for supporting this work. We are particularly grateful to Manuela Cherchi, Erika Ibba, Anna De Santis, Giuseppe Casu, Jessica Ladu, Ilaria Del Rio, Elisa Virdis, Gino Castangia for their annotation work.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 1, pages 267–292. CRC Press, Boca Raton, USA, second edition edition.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When multiwords go bad in machine translation. *MT Summit workshop Proceedings on Multi-word Units in Machine Translation and Transla tion Technology*, page 10.

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. 2009. Improved minimum error rate training in moses. *Prague Bull. Math. Linguistics*, 91:7–16.

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.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, pages 1618–1621.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics, Prague, Czech Republic.

Johanna Monti. 2012. *Multi-word unit processing in Machine Translation - Developing and using language resources for Multi-word unit processing in Machine Translation*. Ph.D. thesis, University of Salerno.

Johanna Monti and Amalia Todirascu. 2015. Multiword Units Translation Evaluation: another pain in the neck? In *Proceedings of Multi-word Units in Machine Translation and Translation Technology (MUMTTT15)*. Malaga.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Anita Rácz, István Nagy T., and Veronika Vincze. 2014. 4fx: Light verb constructions in a multilingual parallel corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Carlos Ramisch, Laurent Besacier, and Alexander Kobzar. 2013. How hard is it to automatically translate phrasal verbs from English to French? In *MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technology*. Nice, France.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461. European Language Resources Association (ELRA), Reykjavik, Iceland.

Nina Schottmüller and Joakim Nivre. 2014. Issues in translating verb-particle constructions from german to english. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 124–131. Association for Computational Linguistics, Gothenburg, Sweden.

Veronika Vincze. 2012. Light verb constructions in the szegedparalellfx english–hungarian parallel corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Digging in the Dirt: Extracting Keyphrases from Texts with KD

Giovanni Moretti[1], Rachele Sprugnoli[1-2], Sara Tonelli[1]
[1]Fondazione Bruno Kessler, Trento
[2] Università di Trento
{moretti,sprugnoli,satonelli}@fbk.eu

Abstract

English. In this paper we present a keyphrase extraction system called Keyphrase Digger (KD). The tool uses both statistical measures and linguistic information to detect a weighted list of n-grams representing the most important concepts of a text. KD is the reimplementation of an existing tool, which has been extended with new features, a high level of customizability, a shorter processing time and an extensive evaluation on different text genres in English and Italian (i.e. scientific articles and historical texts).

Italiano. *In questo articolo presentiamo un sistema di estrazione di espressioni-chiave chiamato Keyphrase Digger. Lo strumento usa sia misure statistiche che informazioni linguistiche per individuare una lista pesata di n-grammi corrispondenti ai concetti più importanti all'interno di un testo. KD è una reimplementazione di un sistema già esistente, che è stato esteso con nuove funzionalità, un alto livello di personalizzazione, maggiore velocità e una valutazione accurata su differenti generi testuali in inglese e italiano (i.e. articoli scientifici e testi storici).*

1 Introduction

This paper presents *Keyphrase Digger* (henceforth KD), a new implementation of the KX system for keyphrase extraction. Both KX (Pianta and Tonelli, 2010; Tonelli et al., 2012) and KD combine statistical measures with linguistic information to identify and extract weighted keyphrases from English and Italian texts. KX took part to the SemEval 2010 task on "Automatic Keyphrase Extraction from Scientific Articles" (Kim et al., 2010) achieving the 7th best result out of 20 in the final ranking. KX is part of TextPro (Pianta et al., 2008), a suite of NLP tools developed by Fondazione Bruno Kessler [1]. The aim of KX re-implementation was to improve system performance in terms of F-measure, processing speed and customizability, so to make its integration possible in web-based applications. Besides, its adaptation to different types of texts has become possible also for not expert users, and its application to large document collections has been significantly improved.

Keyphrases are n-grams of different length, both single and multi-token expressions, which capture the main concepts of a given document (Turney, 2000). Their extraction is very useful when integrated in complex NLP tasks such as text categorization (Hulth and Megyesi, 2006), opinion mining (Berend, 2011) and summarization (D'Avanzo and Magnini, 2005)[2]. Moreover keyphrases, especially if displayed using an effective visualization, can help summarize and navigate document collections so to allow their so-called 'distant reading' (Moretti, 2013). This need to easily grasp the concise content of a text through keyphrases is particularly relevant given the increasing availability of digital document collections in many domains. Nevertheless, outside the computational linguistics community, for example among humanities scholars, the extraction of keywords is often assimilated with the extraction of the most frequent (single) words in a text, see for instance the success of tools such as Textal[3] and Voyant[4] in the Digital Humanities community. In some cases, stopwords are still included among the top-ranked keywords, leading to a key-concept

[1] http://hlt-services2.fbk.eu/textpro/

[2] For a comprehensive survey of the state of the art in automatic keyphrases extraction see Hasan and Ng (2014).

[3] http://www.textal.org/how

[4] http://voyant-tools.org/

list which is little informative and just reflects the Zipfian distribution of words in language. KD, instead, is designed to be easily customized by scholars from different communities, while sticking to the definition of keyphrases in use in the computational linguistics community.

The remainder of the paper is structured as follows. In Section 2 we describe KD architecture and features while in Section 3 the system evaluation is reported. We present the application of KX to the historical domain and the web-based interface available online in Section 4. Future works and conclusions are drawn in Section 5.

2 System Overview

KD is a rule-based system that combines statistical and linguistic knowledge given by PoS patterns. The system takes in input a text pre-processed with a tokenizer, a lemmatizer and a PoS tagger, and derives an ranked and weighted list of single words and multi-token expressions, which represent the most important concepts mentioned in a document. Differently from KX, whose pre-processing step was performed with TextPro for both Italian and English texts, English documents can now be pre-processed with Stanford CoreNLP (Manning et al., 2014), TreeTagger (Schmid, 1994) or TextPro, making KD more flexible than its ancestor.

Furthermore, KD is based on a parallel architecture implemented in Java. This constitutes a major efficiency improvement with respect to KX, which is implemented in Perl and has a sequential architecture. In particular, KX extracts first all possible n-grams from a text and applies selection rules only in a second phase, slowing down the extraction of appropriate candidates, especially in case of long documents. Instead, in KD the following five steps are performed:

1. A file is first split in *n* slices. The number of slices can be decided by the user or is automatically defined by the tool according to the number of machine's CPUs. Each part is processed by an isolated and parallel thread that extracts the set of n-grams corresponding to the language-dependent PoS patterns defined in a configuration file. This file contains the chains of meaningful PoS tags to be extracted, e.g. *noun+adjective* and *noun+preposition+noun* for Italian. Such sequences can be manually edited, deleted or enriched by users, if necessary. The direct access to this configuration files (and also to the other configuration files) is realized by using the MapDB Java library[5] that grants good performances at the read/write serialization time. Differently from KX, in this step the user can choose whether to run KD on inflected word forms or on lemmas, so to cluster extracted key-concepts (e.g. cluster *lingue straniere* and *lingua straniera* under the same key-concept).

2. A function merges the n-grams extracted from different threads in a common list. N-grams with a frequency lower than a threshold defined by the user are removed. In addition, frequencies are recalculated so that if a short key-concept is nested in a longer one (Frantzi et al., 2000) (e.g. *solidarietà economica* and *solidarietà economica internazionale*), the frequency of the former is deducted from the frequency of the latter.

3. The system checks whether, in the preliminary list of extracted concepts, some of them can be treated as synonyms. If yes, the corresponding entries are merged. Synonym resolution is performed on the basis of a list defined by the user, containing n-grams that the tool must consider equivalent, e.g. *liberismo* and *liberalismo economico*.

4. A first relevance score is computed for each concept in the list, taking into consideration different parameters that can be activated or disactivated by the user in a configuration file: frequency and inverse document frequency of n-grams, length of n-grams (so to prefer single words or multi-token expressions), position of first occurrence in the text, presence of specific suffixes (for example to give higher score to abstract concepts ending with *-ismo* and *-itudine*), boost of specific PoS patterns considered important in a given domain. This latter parameter is not present in KX. Another new feature is given by the integration of Apache Lucene library[6]: its scoring system allows to compute efficiently tf/idf at document level.

[5] http://www.mapdb.org
[6] https://lucene.apache.org/core/

5. If the user wants to give preference to specific (i.e. longer) key-concepts, a final re-ranking step can be included. In this way, key-concepts that are specific but have a low frequency are given more relevance than key-concepts that are generic and thus have a higher frequency. This re-ranking step was already present in KX, but the boosting effect had in our opinion an excessive impact on the final keyphrase list, possibly leading to the deletion of top-ranked unigrams. In KD the impact of the re-ranking has been limited to an adjustment of some weights.

3 Evaluation

The evaluation of KD covers different aspects of the system. First, we replicated the SemEval 2010 evaluation using the same data and scorer provided in the keyword extraction task. In this way we checked system performance in terms of F-measure, precision and recall on English texts and on a specific domain, namely scientific papers. As for Italian, we assessed the quality of keyphrases extracted from a corpus of historical documents against a set of key-concepts previously defined by an expert. In addition, we calculated the speed of KD to process a corpus of Italian texts.

In task 5 of SemEval 2010 evaluation campaign, systems were required to automatically assign keyphrases to a corpus of scientific articles and were assessed by using an exact match evaluation metric over stems. This means that micro-averaged precision, recall and F-score were calculated considering the top 5, 10 and 15 candidates found by participating systems that perfectly match the set of manually assigned gold standard keyphrases (in other words, no partial match was allowed). Given that criteria for keyphrase identification depend on the domain, KD parameters were configured to deal with scientific papers. We used the 144 training files and the corresponding answer keys to identify recurrent relevant PoS patterns not present in the default pattern list and determine which ones need to be boosted. On one side, we needed to give importance to long multi-token expressions (e.g. *unified utility maximization model*), which are typical of the scientific domain, on the other we needed to recognize and boost non-expanded acronyms (e.g. *cscw*) that play a central role in this type of articles. For this reason, a specific rule has been added to automatically identify and give a higher weight to unigrams corresponding to acronyms. Furthermore, we noted that the majority of keyphrases provided as gold answers were bigrams and trigrams (74% of the total in the training), so we boosted their corresponding patterns. Overall, we found that the best system configuration on the training data was the following: *min frequency of occurrence* = 2; *max length of keyphrases* = 6; *IDF* = yes; *position of first occurrence at the beginning of the file* = yes; *use of Lucene scoring* = yes; *re-ranking algorithms* = no. Such configuration scored an F-measure of 27.5% on the training set (KX scored 25.6 on the same files). Table 3 shows the results obtained with the same configuration on the test set. Results over the 5, 10 and 15 top-ranked keyphrases are reported: the F-score for the top 15 candidates, i.e. 26.5%, corresponds to the second best results in SemEval 2010 competition with an improvement of almost 2 points with respect of KX performance (i.e. 23.9%). Note that the first-ranked system relied on a supervised approach, making KD the best performing rule-based system evaluated on this data set.

	Precision	Recall	F-score
Top 5	35.4%	12.7%	18.0%
Top 10	31.3%	21.4%	25.4%
Top 15	26.2%	26.8%	**26.5%**

Table 1: Precision, Recall and F-score of KD evaluated on the test set provided in task 5 of SemEval 2010

As for Italian, we asked a history scholar to manually identify a set of key-concepts considered relevant to characterize the corpus of Alcide De Gasperi's writings, dating back to the first half of the XX century (De Gasperi, 2006)[7]. This task was performed independently from the development of KD, so no specific instructions related to the keyphrase extraction task were given (e.g. the scholar could select also keyphrases which were not present in the documents). A set of about 60 keyphrases was defined for each of the five relevant periods of De Gasperi's political career, which we used as a gold standard to evaluate the system performance. Over these five periods, which correspond to five corpora, KD achieved a macro-average precision of 23.8% calculated in

[7] Alcide De Gasperi was the first Prime Minister of the Italian Republic and one of the founders of the European Union

an 'exact match' setting. Since some of the key-concepts identified by the expert do not appear in the text, it was impossible for KD to extract them. For instance, *Alleanza Atlantica* is an expression never used by De Gasperi who, instead, used the expression *Patto Atlantico*, correctly extracted by KD. We compared KD results with the ones obtained using the Distiller-CORE library developed by the University of Udine (De Nart et al., 2015) and available at `https://github.com/ailab-uniud/distiller-CORE` [8]. Distiller-CORE extracted 20 keyphrases from each of De Gasperi's subcorpora, achieving a macro-average precision of 15%. Considering only the first 20 keyphrases extracted by KD againts the full list of expert's keywords, our tool achieved a precision of 42%.

Since KX speed was a main issue when processing large document collections, we also ran a comparison between KX and KD processing time, running both systems on the same corpus of 101,000 Italian tokens and on the same machine[9]. As for parameters, we used the most comparable setting: two re-rank algorithms, frequency of occurrence set at 1, max length of 4 tokens for extracted keyphrases. It took KD 7 seconds to return the list of keyphrases, whereas KX needed 3.4 minutes to complete the task[10]. The improved system speed makes KD particularly suitable for integration in web applications, where texts can be processed on the fly. Some examples are reported in the following section.

4 Applications

KD has been integrated in the last version of ALCIDE[11] (*Analysis of Language and Content In a Digital Environment*), an online platform for Historical Content Analysis (Moretti et al., 2014). In ALCIDE the output of KD is displayed by means of two visualizations: a bar chart and a tag cloud. This analysis and the corresponding visualizations are available both at the corpus and at the single document level. Moreover, the user can search for a specific key-concept, retrieve the documents where it appears and display its distribution along a timeline. Within ALCIDE, KD has been applied to a corpus of F.T. Marinetti's writings (Daly, 2013), with the goal of exploring Futurism works with NLP tools. Figure 1 shows the 20 most frequent key-concepts extracted from all manifestos written by Marinetti between 1909 and 1921. Such key-concepts can be mainly divided into two categories: the ones related to the political program of Marinetti characterized by the exaltation of war and of his homeland (i.e. *guerra*, *Italia*, *patriottismo*, *eroismo*) and the ones associated with his artistic program, with particular emphasis on futurism style in poetry (*parole in libertà*) and theatre (*teatro della sorpresa*).

KD is also available as a web application at the link `http://celct.fbk.eu:8080/KD_KeyDigger/`, through which users can copy&paste sample documents and run the keyphrase extraction process. Four pre-defined parameter settings are available: one for scientific papers, one for historical texts, one for news articles and one for all the other types of texts. Besides, also single parameters can be further specified (e.g. maximum keyphrase length). Keyphrases can be visualized as bar chart and word cloud, and be exported in tab-separated format.

5 Conclusions

This paper presents KD, a keyphrase extraction system that re-implements the basic algorithm of KX but adds new features, a high level of customizability and an improved processing speed. KD currently works on English and Italian and can take in input texts pre-processed with different available PoS taggers and lemmatizers for these two languages. Nevertheless, the system could be easily adapted to manage more languages and additional PoS taggers by modifying few configuration parameters.

KD will be soon integrated in the next TextPro release[12] and it will be also released as a standalone module. Meanwhile, we made it available online as part of an easy-to-use web application, so that it can be easily accessed also by users without a technical background. This work targets in particular humanities scholars, who often do not know how to access state-of-the-art tools for keyphrase extraction.

[8] We also tried to use AlchemyAPI (`http://www.alchemyapi.com/api`) and Sensium (`https://www.sensium.io`) API endpoints but they do not allow processing long documents.

[9] CPU: 2.3GHz Intel Core i7, RAM: 8Gb 1600 mhz ddr3, Hard Disk: SSD serial SATA 3

[10] 7,000 ms versus 206,546 ms

[11] `https://youtu.be/PhkuOfIod1A`

[12] Check the TextPro website (`http://textpro.fbk.eu/` for updates

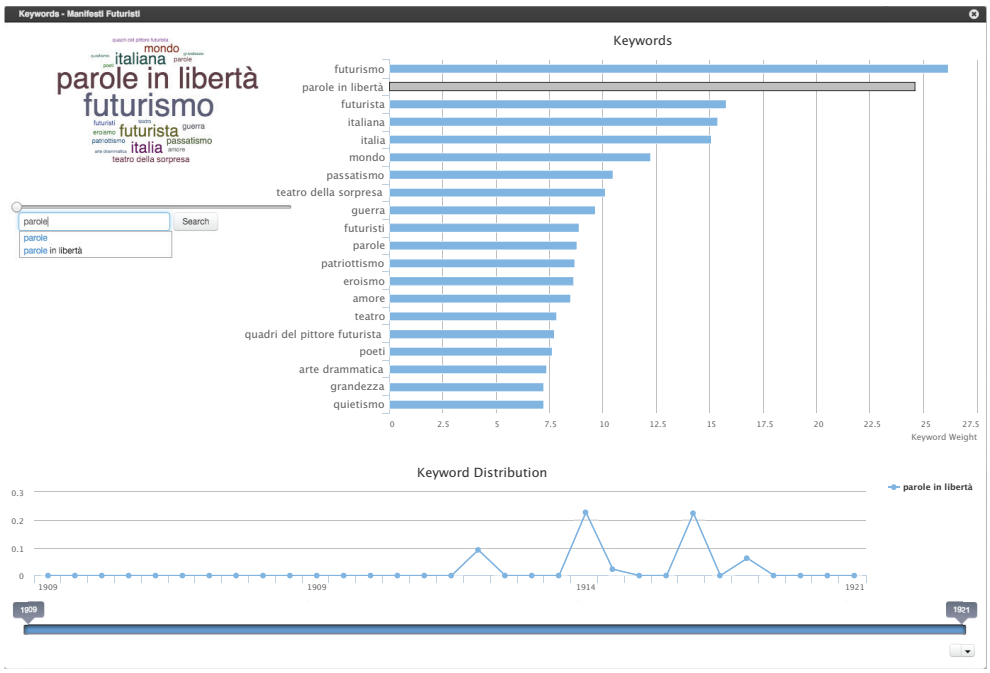


Figure 1: Visualization of key-concepts extracted with KD from Marinetti's manifestos in the ALCIDE platform

References

Gábor Berend. 2011. Opinion Expression Mining by Exploiting Keyphrase Extraction. In *Proceedings of IJCNLP*, pages 1162–1170.

Selena Daly. 2013. "The Futurist mountains": Filippo Tommaso Marinetti's experiences of mountain combat in the First World War. *Modern Italy*, 18(4):323–338.

Ernesto D'Avanzo and Bernado Magnini. 2005. A keyphrase-based approach to summarization: the lake system at DUC-2005. In *Proceedings of DUC*.

A. De Gasperi. 2006. Scritti e discorsi politici. In E. Tonezzer, M. Bigaran, and M. Guiotto, editors, *Scritti e discorsi politici*, volume 1. Il Mulino.

Dario De Nart, Dante Degl'Innocenti, and Carlo Tasso. 2015. Introducing distiller: a lightweight framework for knowledge extraction and filtering. In *Proceedings of the UMAP Workshops*.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Anette Hulth and Beáta B Megyesi. 2006. A study on automatically extracted keywords in text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 537–544.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26.

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*, pages 55–60.

Giovanni Moretti, Sara Tonelli, Stefano Menini, and Rachele Sprugnoli. 2014. ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment. In *Atti della prima Conferenza Italiana di Linguistica Computazionale*.

Franco Moretti. 2013. *Distant Reading*. Verso, London.

Emanuele Pianta and Sara Tonelli. 2010. KX: A flexible system for keyphrase extraction. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 170–173.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the Language Resources and Evaluation Conference*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49.

Sara Tonelli, Elena Cabrio, and Emanuele Pianta. 2012. Key-concept extraction from french articles with KX. *Actes de latelier de clôture du huitième défi fouille de texte (DEFT)*, pages 19–28.

Peter Turney. 2000. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(4):303–336.

Automatic extraction of Word Combinations from corpora: evaluating methods and benchmarks

Malvina Nissim[1], Sara Castagnoli[2], Francesca Masini[2], Gianluca E. Lebani[3], Lucia Passaro[3], Alessandro Lenci[3]
[1]CLCG, University of Groningen, [2]University of Bologna, [3]University of Pisa
m.nissim@rug.nl, {s.castagnoli,francesca.masini}@unibo.it,
{gianluca.lebani,lucia.passaro}@for.unipi.it,
alessandro.lenci@unipi.it

Abstract

English. We report on three experiments aimed at comparing two popular methods for the automatic extraction of Word Combinations from corpora, with a view to evaluate: i) their efficacy in acquiring data to be included in a combinatory resource for Italian; ii) the impact of different types of benchmarks on the evaluation itself.

Italiano. *Presentiamo i risultati di tre esperimenti che mirano a confrontare due metodi di estrazione automatica di combinazioni di parole da corpora, con lo scopo di: (i) valutare l'efficacia dei due metodi per acquisire dati da includere in una risorsa combinatoria per l'italiano, e (ii) analizzare e confrontare i metodi di valutazione stessi.*

1 Introduction

We use the term **Word Combinations** (WoCs) to encompass both Multiword Expressions, namely WoCs characterised by different degrees of fixedness and idiomaticity, such as idioms, phrasal lexemes, collocations, preferred combinations (Calzolari et al., 2002; Sag et al., 2002; Gries, 2008), and the distributional properties of a word at a more abstract level (argument structure, subcategorization frames, selectional preferences).

Currently, apart from purely statistical approaches, the most common methods for the extraction of WoCs involve searching a corpus via sets of patterns and then ranking the extracted candidates according to various association measures (AMs) in order to distinguish meaningful combinations from sequences of words that do not form any kind of relevant unit (Villavicencio et al., 2007; Ramisch et al., 2010). Generally, the search is performed for either shallow morphosyntactic (POS) patterns (**P-based approach**) or syntactic dependency relations (**S-based approach**) (Lenci et al., 2014; Lenci et al., 2015).

While P-based approaches have shown to yield satisfactory results for relatively fixed, short and adjacent WoCs, it has been suggested that syntactic dependencies might be more helpful to capture discontinuous and syntactically flexible WoCs (Seretan, 2011). The two methods intuitively seem to be highly complementary rather than competing with one another, and attempts are currently being proposed to put them together (Lenci et al., 2014; Lenci et al., 2015; Squillante, 2015). In previous work (Castagnoli et al., forthcoming), we compared the performance of the two methods against two benchmarks (a dictionary and expert judgments), showing that the two methods are indeed complementary and that automatic extraction from corpora adds a high number of WoCs that are not recorded in manually compiled dictionaries.

As an extension of that work, in this paper we shift the focus of investigation by addressing the following research questions: What is the effect of different benchmarks when evaluating an extraction method? What do our results tell us about the bechmarks themselves? And, as a byproduct, can experts / laypeople be exploited to populate a lexicographic combinatory resource for Italian?

2 Benchmarks

The performance of WoC extraction can be evaluated in various ways. A straightforward way is assessing extracted combinations against an existing dictionary of WoCs (Evaluation 1). Such resources, however, are often compiled manually on the basis of the lexicographers' intuition only. The dictionary can be seen as a one-expert judgement, in a top-down (lexicographic) fashion. Moreover, this type of evaluation assumes the dictionary as an absolute gold standard, without considering that any dictionary is just a partial representation of the

lexicon and that corpus-based extraction might be able to identify further possible WoCs.

Another way to assess the validity of extracted combinations is via human evaluation. One problem with this approach lies in the competence of the judges: experts are difficult to recruit, but it isn't completely clear whether people unfamiliar with linguistic notions are able to grasp the concept of WoCs, and to judge the validity of the extracted strings. Knowing whether this is a task that can be performed by laypeople is not only theoretically interesting, but also practically useful. To this end, we set up two distinct human-based experiments: one involving experts (Evaluation 2), and one involving laypeople (Evaluation 3). Table 1 summarises the characteristics of the three strategies, whose results are discussed and compared in the next sections, in terms of the kind and number of contributors, the procedure (bottom-up means that the evaluation is done directly on the corpus-extracted WoCs rather than against a pre-compiled list (top-down)), the assessment performed or required, and the data evaluated.

3 Experimental evaluation

3.1 Data and WoC extraction

We selected a sample of 25 Italian target lemmas (TLs) – 10 nouns, 10 verbs and 5 adjectives – and we extracted P-based and S-based combinatory information from *la Repubblica* corpus (Baroni et al., 2004)[1]. TLs were selected by combining frequency information derived from *la Repubblica* and inclusion in DiCI (Lo Cascio, 2013), a manually compiled dictionary of Italian WoCs, which is also used for (part of the) evaluation.

As regards the P-based method, we extracted all occurrences of each TL in a set of 122 predefined POS-patterns deemed representative of Italian WoCs, using the **EXTra** tool (Passaro and Lenci, forthcoming). EXTra retrieves all occurrences of the specified patterns as linear and contiguous sequences (no optional slots) and ranks them according to various association measures, among which we chose Log Likelihood (LL). The search considers lemmas, not wordforms. Only sequences with frequency over 5 were considered.

As regards the S-based method, we extracted the distributional profile of each TL using the **LexIt** tool (Lenci et al., 2012). The LexIt distributional profiles contain the syntactic slots (subject, complements, modifiers, etc.) and the combinations of slots (frames) with which words co-occur, abstracted away from their surface morphosyntactic patterns and actual word order. The statistical salience of each element in the distributional profile is estimated with LL. For each TL we extracted all its occurrences in different syntactic frames together with the lexical fillers (lemmas) of the relevant syntactic slots. Only candidate WoCs with frequency over 5 have been considered.

3.2 Evaluation against a dictionary

The gold standard we used for this part of the evaluation, fully presented in (Castagnoli et al., forthcoming), is the DiCI dictionary (Lo Cascio, 2013).

Recall is calculated as the percentage of extracted candidates out of the combinations found in the gold standard. Generally, EXTra performs better than LexIt for nominal and adjectival TLs, whereas LexIt has a higher recall for virtually all verbal TLs.[2] **R-precision**, which measures precision at the rank position corresponding to the number of combinations found in DiCI, is almost always higher for LexIt than for Extra, irrespective of POS. Total **overlap** is calculated as the percentage of cases in which EXTra/LexIt retrieve (or not) the same gold standard combinations. For instance, the entry for *giovane* 'young' in DiCI contains 50 combinations. Out of these, 20 are retrieved by both EXTra and LexIt, 27 are retrieved by neither, and only LexIt extracts 3 further WoCs. This means that the two systems perform similarly for 94% of cases found in the benchmark data. Total overlap runs between 59.07% and 94% (average 76.05%).

3.3 Human-based evaluation with experts

We recruited a number of linguists, mainly with a background in translation and/or corpus work. They were asked to assess the validity of candidates by assigning one of 3 possible values: Y (Yes, a valid WoC), N (No, not a valid WoC), U (Uncertain / may be part of a valid WoC). We obtained judgments for 2,000 candidates (50% EXTra, 50% LexIt, taking the top 100 results for 10 TLs from each system). We used two annotators per

[1]The version we used was POS-tagged with the tool described in (Dell'Orletta, 2009) and dependency-parsed with DeSR (Attardi and Dell'Orletta, 2009).

[2]This result may in part be due to the POS-patterns used, which were limited to a maximum of 4 slots, thus preventing EXTra from capturing longer verbal expressions. However, this can be seen as an inherent limitation of the P-based approach, given that the complexity/variability of patterns increases immensely as soon as we consider longer strings.

Table 1: Overview of evaluation strategies.

	Evaluation 1 (DiCI)	Evaluation 2 (experts)	Evaluation 3 (laypeople)
contributors	expert (1)	expert (> 1)	naive (> 1)
procedure	top-down	bottom-up	bottom-up
assessment	inclusion	validity (categorical)	typicality + idiomaticity (scalar)
candidates	all extracted (ca.105,000)	top extracted per TL (2,000)	random from Eval 2 (630)

candidate, and considered valid WoCs those that received either YY or YU values.

A total of 855 entries (EXTra: 408, LexIt: 447) were judged as valid. Out of these, 534 (62.5%) are not recorded in DiCI (EXTra: 273, LexIt: 261). If we intersect the two sets, we find that only 80 of these additional WoCs are in common, which means that we have 454 actual *new* valid WoCs, retrieved thanks to the corpus-based methodology.

3.4 Human-based evaluation with laypeople

Judgements from laypeople were obtained by setting up a crowdsourcing task on the Crowdflower platform (`http://www.crowdflower.com`). Compared to the previous experiment, annotators were asked to judge two aspects of the candidate combinations: how *typical* they are, i.e. how important it is that they are included in a multiword dictionary; and how *idiomatic* they are, i.e. how much their overall meaning is not directly inferrable from their parts (non-compositionality). Both judgements were asked on a scale from 1 to 5 rather than via the discrete values used by the experts (Y/N/U). The Appendix shows a snapshot of the instructions and the task the annotators were presented with. Note that candidates were presented in the form they were extracted from the corpora, i.e. lemmatized (e.g. *vero guerra* instead of *vera guerra* 'true war'). Further, LexIt examples may contain free slots (e.g. *pagare * multa* 'pay * fine').

This second human-based experiment was primarily expected to shed light on whether experts' and laypeople's judgements differ in the assessment of WoCs. Moreover, the additional question about idiomaticity was aimed at detecting potential differences in the degree of idiomaticity of the WoCs the two methods extract.

3.4.1 Participation and results

Potential annotators could train on some "gold" combinations, which were also used to assess the quality of the contributors. Such gold combinations were not part of the original dataset and are not further included in the analysis. Contributors who misclassified more than 60% of the test questions were not allowed to proceed with the rest of the combinations, so that out of 81 potential contributors we were left with 53 reliable ones, and only 36 actively working on the task (with contributions ranging from 300 to 20 annotated combinations).

As a result, this second human-based experiment is based on 630 combinations (a random subsample of the original 2,000 dataset of the expert-based evaluation) for which we managed to collect three independent judgements. The distribution between combinations extracted by Extra (322) and by Lexit (308) is approximately preserved.

In Figure 1 and 2 we report the results of the evaluation for the "typicality" and "idiomaticity" assessments (x in the chart labels), respectively, splitting the overall range into five subranges.

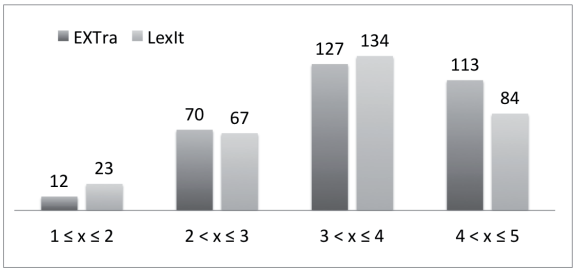


Figure 1: Results of the crowdsourcing evaluation for how *typical* combinations are (average of three annotations, global range 1–5).

If we deem *valid* any combination with average score > 3 (the two rightmost columns in the Figures), we can observe that laypeople judged as valid combinations the majority of candidates in both sets and more precisely: approx. 75% of candidates extracted by EXTra (240/322) and approx. 71% of candidates extracted by LexIt (218/308). The two methods perform similarly also regarding the capability of extracting combinations with stronger or weaker idiomaticity: approx. 38% of (those judged as) typical combinations obtained via

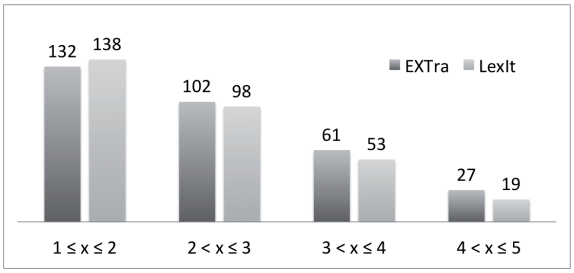


Figure 2: Results of the crowdsourcing evaluation for how *idiomatic* combinations are (average of three annotations, global range 1–5).

Table 2: Comparison of "valid" combinations according to laypeople and expert judges.

	valid for both	laypeople only	total
EXTra	124	116	240
LexIt	119	99	218

EXTra were also judged idiomatic (88), and approx. 33% of (those judged as) typical combinations obtained via LexIt were also judged idiomatic (72). Overall, EXTra appears to have a slightly better performance in both cases (although the difference is not statistically significant), and this is different from what we observed in the expert-based evaluation. The reason for this may lie in the fact that the LexIt candidates correspond to more abstract and schematic WoCs, which could eventually be harder to map onto specific instances by the evaluators.

3.4.2 Experts vs laypeople

Do experts and laypeople share the same notion of what a *typical* combination is? Given that in the crowdsourcing experiment we used a subset of the expert set, we checked how many of those combinations that were assessed as valid (> 3) by laypeople had also been evaluated as valid by the experts (YY or YU, see Section 3.3). Table 2 shows the results of such comparison. If we treat the experts' judgements as gold, we can interpret the values in the table as precision, resulting in 0.517 for EXTra and 0.546 for LexIt. Both figures are rather low, and suggest that the notion of "typicality" of a combination - or possibly the notion of a combination at all - isn't at all straightforward.

A qualitative analysis of the disagreements between laypeople and experts leads to some interesting insights. Combinations annotated as valid only by the former include: a) cases were the candidate differs from a proper WoC only for a small detail: e.g. *dichiarare una guerra* 'declare a war' (proper WoC: *dichiarare guerra* 'declare war', without indefinite article), *tenere il ostaggio* 'take the hostage' (proper WoC: *tenere in ostaggio* 'take s.one hostage'), showing little attention to details; b) cases of uncertain collocations: e.g. *libretto rosso* 'red booklet', *famiglia italiano* 'Italian family', *prendere - carta* 'take - paper'; c) blatantly incomplete/nonsensical combinations: e.g. *di guerra di* 'of war of', *di molto famiglia* 'of many family'; d) a few WoCs that were not recognised as valid by experts: e.g. *dare la mano* 'shake one's hand', *prendere corpo* 'to take shape', *guerra punica* 'punic war'.

4 Discussion and conclusion

As for extraction methods per se, we observed that recall against a manually compiled WoC dictionary is good for both EXTra and LexIt, and, especially, that the two systems are complementary. In the human evaluation performed by experts, 40% of WoCs automatically extracted with EXTra and LexIt are deemed valid, and more than half of these are not attested in DiCi. We can thus say that data from corpora proves to be very fruitful, especially if we use the two methods complementarily.

As for benchmarks, we observed that the dictionary we have evaluated is not an exhaustive resource, and should be complemented with corpus-extracted WoCs. We also observed that expert- and laypeople-based evaluations differ, which raises a number of interesting, albeit puzzling questions. Overall, it seems that the notion of WoC, as well as of idiomaticity, is quite a complex one to grasp for non-linguists: the collection of judgments took quite a long time to be completed (much more than we expected) and evaluators explicitly regarded the task and the instructions as particularly complex.

The results of our experiments thus leave us with a sort of methodological conundrum, as both a dictionary-based gold standard and a human-based evaluation have limitations. Using experts not only makes the evaluation expensive, but also little ecological, as it is standard practice in psycholinguistics and computational linguistics to resort to laypeople judgments. The fact that evaluating WoCs isn't easy for laypeole may cast some shadows on the concept of WoC itself. This suggests that improving extraction methods must go hand in hand with the theoretical effort of making the very notion of WoC more precise, in order to make it an experimentally solid and testable notion.

Acknowledgments

This research was carried out within the **CombiNet** project (PRIN 2010-2011 *Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary*, n. 20105B3HE8), funded by the Italian Ministry of Education, University and Research (MIUR). `http://combinet.humnet.unipi.it`.

References

Giuseppe Attardi and Felice Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of NAACL 2009*, pages 261–264.

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. In *Proceedings of LREC 2004*, pages 1771–1774.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of LREC 2002*, pages 1934–1940.

Sara Castagnoli, Gianluca E. Lebani, Alessandro Lenci, Francesca Masini, Malvina Nissim, and Lucia C. Passaro. forthcoming. Pos-patterns or syntax? comparing methods for extracting word combinations. In *Proceedings of EUROPHRAS 2015 (provisional)*.

Felice Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA 2009*.

Stefan Th. Gries. 2008. Phraseology and linguistic theory: a brief survey. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: an interdisciplinary perspective*, pages 3–25. John Benjamins, Amsterdam & Philadelphia.

Alessandro Lenci, Gabriella Lapesa, and Giulia Bonansinga. 2012. LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of LREC 2012*, pages 3712–3718.

Alessandro Lenci, E. Gianluca Lebani, Sara Castagnoli, Francesca Masini, and Malvina Nissim. 2014. SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations. In *Proceedings of CLiC-it 2014*, pages 234–238, Pisa, Italy.

Alessandro Lenci, E. Gianluca Lebani, S.G. Marco Senaldi, Sara Castagnoli, Francesca Masini, and Malvina Nissim. 2015. Mapping the Constructicon with SYMPAThy: Italian Word Combinations between fixedness and productivity. In *Proceedings of the NetWordS Final Conference*, pages 144–149, Pisa, Italy.

Vincenzo Lo Cascio. 2013. *Dizionario Combinatorio Italiano (DiCI)*. John Benjamins, Amsterdam/Philadelphia.

Lucia C. Passaro and Alessandro Lenci. forthcoming. Extracting terms with EXTra. In *Proceedings of EUROPHRAS 2015 (provisional)*.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a framework for multiword expression identification. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

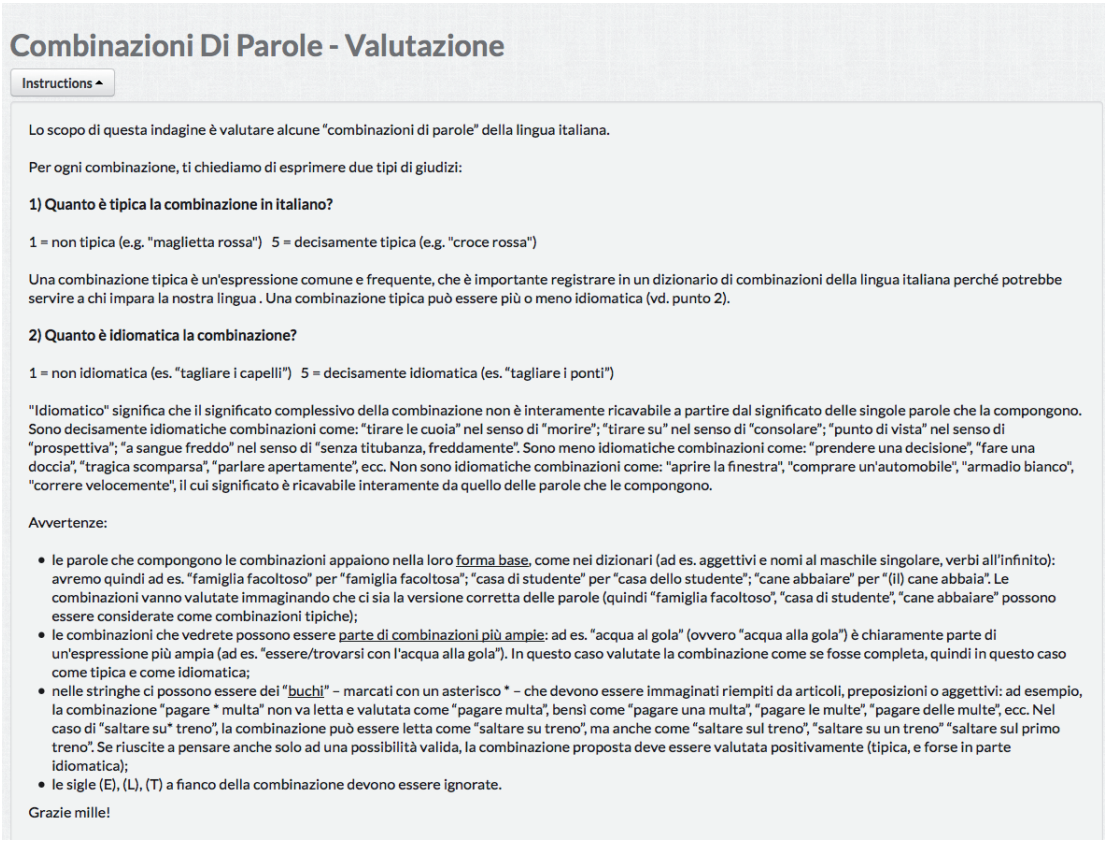
Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing 2002*, pages 1–15.

Violeta Seretan. 2011. *Syntax-based collocation extraction*. Springer, Dordrecht.

Luigi Squillante. 2015. *Polirematiche e collocazioni dell'italiano. Uno studio linguistico e computazionale.* Ph.D. thesis, Università di Roma "La Sapienza".

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of EMNLP-CoNLL 2007*, pages 1034–1043.

Appendix: Crowdflower Job



Combinazioni Di Parole - Valutazione

Instructions

Lo scopo di questa indagine è valutare alcune "combinazioni di parole" della lingua italiana.

Per ogni combinazione, ti chiediamo di esprimere due tipi di giudizi:

1) Quanto è tipica la combinazione in italiano?

1 = non tipica (e.g. "maglietta rossa") 5 = decisamente tipica (e.g. "croce rossa")

Una combinazione tipica è un'espressione comune e frequente, che è importante registrare in un dizionario di combinazioni della lingua italiana perché potrebbe servire a chi impara la nostra lingua . Una combinazione tipica può essere più o meno idiomatica (vd. punto 2).

2) Quanto è idiomatica la combinazione?

1 = non idiomatica (es. "tagliare i capelli") 5 = decisamente idiomatica (es. "tagliare i ponti")

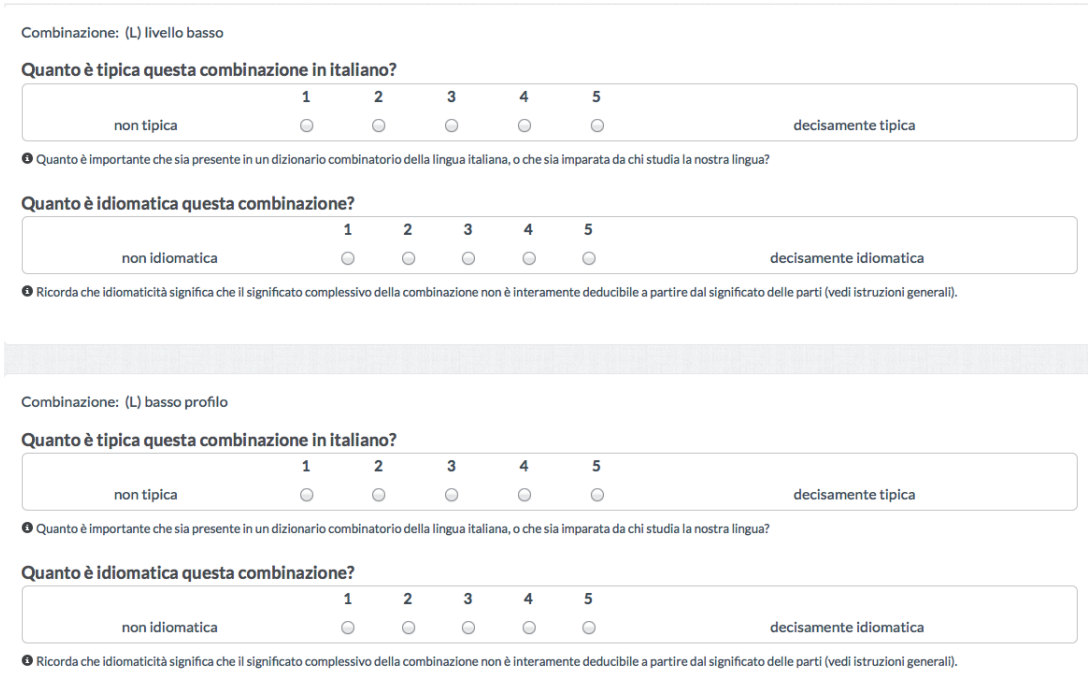
"Idiomatico" significa che il significato complessivo della combinazione non è interamente ricavabile a partire dal significato delle singole parole che la compongono. Sono decisamente idiomatiche combinazioni come: "tirare le cuoia" nel senso di "morire"; "tirare su" nel senso di "consolare"; "punto di vista" nel senso di "prospettiva"; "a sangue freddo" nel senso di "senza titubanza, freddamente". Sono meno idiomatiche combinazioni come: "prendere una decisione", "fare una doccia", "tragica scomparsa", "parlare apertamente", ecc. Non sono idiomatiche combinazioni come: "aprire la finestra", "comprare un'automobile", "armadio bianco", "correre velocemente", il cui significato è ricavabile interamente da quello delle parole che le compongono.

Avvertenze:

- le parole che compongono le combinazioni appaiono nella loro forma base, come nei dizionari (ad es. aggettivi e nomi al maschile singolare, verbi all'infinito): avremo quindi ad es. "famiglia facoltoso" per "famiglia facoltosa"; "casa di studente" per "casa dello studente"; "cane abbaiare" per "(il) cane abbaia". Le combinazioni vanno valutate immaginando che ci sia la versione corretta delle parole (quindi "famiglia facoltoso", "casa di studente", "cane abbaiare" possono essere considerate come combinazioni tipiche);
- le combinazioni che vedrete possono essere parte di combinazioni più ampie: ad es. "acqua al gola" (ovvero "acqua alla gola") è chiaramente parte di un'espressione più ampia (ad es. "essere/trovarsi con l'acqua alla gola"). In questo caso valutate la combinazione come se fosse completa, quindi in questo caso come tipica e come idiomatica;
- nelle stringhe ci possono essere dei "buchi" – marcati con un asterisco * – che devono essere immaginati riempiti da articoli, preposizioni o aggettivi: ad esempio, la combinazione "pagare * multa" non va letta e valutata come "pagare multa", bensì come "pagare una multa", "pagare le multe", "pagare delle multe", ecc. Nel caso di "saltare su* treno", la combinazione può essere letta come "saltare su treno", ma anche come "saltare sul treno", "saltare su un treno" "saltare sul primo treno". Se riuscite a pensare anche solo ad una possibilità valida, la combinazione proposta deve essere valutata positivamente (tipica, e forse in parte idiomatica);
- le sigle (E), (L), (T) a fianco della combinazione devono essere ignorate.

Grazie mille!

Screenshot of the Crowdflower job: instructions.



Combinazione: (L) livello basso

Quanto è tipica questa combinazione in italiano?

	1	2	3	4	5	
non tipica	○	○	○	○	○	decisamente tipica

Quanto è importante che sia presente in un dizionario combinatorio della lingua italiana, o che sia imparata da chi studia la nostra lingua?

Quanto è idiomatica questa combinazione?

	1	2	3	4	5	
non idiomatica	○	○	○	○	○	decisamente idiomatica

Ricorda che idiomaticità significa che il significato complessivo della combinazione non è interamente deducibile a partire dal significato delle parti (vedi istruzioni generali).

Combinazione: (L) basso profilo

Quanto è tipica questa combinazione in italiano?

	1	2	3	4	5	
non tipica	○	○	○	○	○	decisamente tipica

Quanto è importante che sia presente in un dizionario combinatorio della lingua italiana, o che sia imparata da chi studia la nostra lingua?

Quanto è idiomatica questa combinazione?

	1	2	3	4	5	
non idiomatica	○	○	○	○	○	decisamente idiomatica

Ricorda che idiomaticità significa che il significato complessivo della combinazione non è interamente deducibile a partire dal significato delle parti (vedi istruzioni generali).

Screenshot of the Crowdflower job: examples involving the TL *basso* 'low/short'.

Improved Written Arabic Word Parsing through Orthographic, Syntactic and Semantic constraints

Nahli Ouafae **Simone Marchi**

Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche
Via G. Moruzzi, 1, 56124 Pisa - Italy
{firstname.lastname}@ilc.cnr.it

Abstract

English. The Arabic script omits diacritics, which are essential to fully specify inflected word forms. The extensive homography caused by diacritic omission considerably increases the number of alternative parses of any morphological analyzer that makes no use of contextual information. Many such parses are spurious and can be filtered out if *diacriticization*, i.e. the process of interpolating diacritics in written forms, takes advantage of a number of orthographic, morpho-syntactic and semantic constraints that operate in Arabic at the word level. We show that this strategy reduces parsing time and makes morphological analysis of written texts considerably more accurate.

Italiano. Le convenzioni ortografiche della lingua araba consentono l'omissione dei diacritici, introducendo così numerosi casi di omografia tra forme flesse e la conseguente proliferazione di analisi morfologiche contestualmente spurie. Un analizzatore morfologico che utilizzi i vincoli ortografici, morfo-sintattici e semantici che operano a livello lessicale, può tuttavia ridurre drasticamente il livello di ambiguità morfologica del testo scritto, producendo analisi più efficienti e accurate.

1 Introduction

Arabic is a morphologically rich language, where a lot of information on morpho-syntactic and semantic relationships among words in context is directly expressed at the word level[1]. Some prepositions, conjunctions and other particles are morphologically realized as proclitics, while all pronouns are enclitics. Orthographic, morphological and syntactic characteristics of Arabic contribute to increasing the level of ambiguity of written word forms, which is made even more complex by the unsystematic use of diacritical markers in the Arabic script[2]. In this paper we suggest that spelling rules, morpho-syntactic and semantic constraints should be jointly evaluated as early as possible in parsing an Arabic text. In particular, the analysis of spelled-out forms requires simultaneous use of morpho-syntactic and semantic information to define constraints on NLP, and "interpolate" missing vowels/diacritics (diacriticization) in Arabic written texts.

2 Morphological structure of Arabic words

2.1 Maximal and minimal words

In Arabic, written tokens correspond to either a "minimal word form" (see infra) delimited by white spaces, or a morphologically more complex token resulting from a concatenation of a minimal word form with clitics (called "maximal word form"). In (1), we offer the example of a maximal word form, consisting of the inflected form of the verb *kataba* 'write' surrounded by clitics[3].

Example 1 ***wa=ta-ktub-u=hu***
and=2MS-write$_{IPFV}$-PRS.IND=it
'and you write it'

The morphological structure of (1)[4] can be schematized as follows:
proclitics=prefix-stem-suffixes=enclitics.

By removing clitics, the remaining word form (***ta-ktub-u***) is a minimally autonomous inflected form, whose structure consists of **prefix-stem-suffixes**. Due to these levels of morphological embedding, word tokenization in Arabic must be followed by a sub-tokenization phase demarcating the boundaries between proclitics, the minimal word and enclitics.

[1] Tsarfaty et all (2013).

[2] Farghaly A., and Shaalan K. (2009).

[3] Interlinear glosses follow the standard set of parsing conventions and grammatical abbreviations explained in: "**The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses**" February 2008. Hyphen marks segmentable morphemes and an equal sign marks clitic boundaries, both in transliterations and in the interlinear gloss.

[4] Dichy J. (1997).

2.2 Ambiguity in tokenization

In Arabic written texts, vowels, gemination and other signs are written as diacritics added above or below consonant letters. Their marking, however, is not systematic. For instance, the word *kataba* 'he wrote' can be written in any of the following variants: *ktb*, *katb*, *katab*, *ktaba*, *katba,* etc. Furthermore, *ktb* is shared by all members of its derivational family. This means that, by vocalizing the skeleton differently, one can obtain word forms of other lexical units than the base verb: *kutub* (books), *katb* (writing), *kattaba* (dictate; make write). As a result of these powerful morphological relations, omission of diacritics in written texts causes extensive homography in Arabic. Text reading and understanding is an active process of text interpretation, based on context, grammatical knowledge and vocabulary. For example, clitics can be in grammatical combination with only some minimal forms. Hence, one can use the presence of clitics in maximal forms to cut on the level of ambiguity of their embedded minimal forms.

Section 2.3 illustrates how addition of proclitics can help morpho-syntactic disambiguation. Section 2.4 shows how semantic features of the minimum word can help constrain the number of enclitics that can be added to it.

2.3 Morpho-syntactic characteristics

Arabic clitics are important because impose morpho-syntactic restrictions on the words they are attached to. Particularly when the particle is proclitic, morphological restrictions can be of help for the morpho-syntactic analysis of a spelled-out form. Consider the example 2, where the form *ktb* is preceded by the determiner and the preposition ***li***. In this case, the form *llktb* has a single reading because, in Arabic, all prepositions require genitive case:

Example 2 ***li=l=kutub-i***
to=DET=books-GEN$_{.DEF}$
'to the books'

Hence, to decrease the level of orthographic ambiguity, it is important to have a full list of clitics and the morphotactic constraints defining their compatibility with minimal words.

2.4 Verb semantics and agreement

Another peculiarity of Arabic is a complex system of N-V agreement rules. For example, when the subject refers to a rational entity (e.g. a person), its anaphoric clitic in the verb agrees with it in both number (SG, DU and PL) and gender (M and F). However, when the subject refers to an irrational entity, e.g. a non-human entity, its clitic marker in the verb is always in third person, and agrees with the noun in both number and gender only if the noun is singular or dual. If the noun is plural, the anaphoric clitic is 3SGF only. Consider the example 3 below. The verb *wahaja* requires an inanimate subject[5]. Thus, it can only select pronoun clitics in 3 SG/DU. Even if the subject is plural (3.b and 3.d), the verb is inflected in 3FSG. Furthermore, it cannot be inflected in the first and second person.

Example 3

a- **النَّارُ وَهَجَتْ**
'an=nār-u wahaj-at
DET=fire-NOM burn$_{.PST}$-3SGF
'The fire burns' (cf. DET='al)

b- **النِّيرَانُ وَهَجَتْ**
'an=nīrān-u wahaj-at
DET=fires-NOM burn$_{.PST}$-3SGF
'The fires burn' (cf. DET='al)

c- **العِطْرُ وَهَجَ**
'al='iṭr-u wahaj-a
DET=perfume-NOM spread$_{.PST}$-3SGM
'The perfume spreads' (cf. DET='al)

d- **العُطُورُ وَهَجَتْ**
'al='uṭūr-u wahaj-at
DET= perfume -NOM burn$_{.PST}$-3SGF
'The perfumes spread' (cf. DET='al)

To sum up, verbs are characterized by a conceptual structure that governs the selection and morpho-syntactic mapping of its arguments. The semantic properties of lexical units enforce constraints that can help predict their morpho-syntactic realization. Number and category of syntactic arguments are licensed by lexical restrictions imposed by the verb semantic class. These "selectional restrictions" on arguments are an essential part of the verb meaning and govern its morpho-syntactic behaviour[6]. Thanks to these restrictions, it becomes possible to successfully tackle possible ambiguities in the morpho-syntactic realization of the argument structure of a verb.

3 Word processing issues

We consider here the impact of the above-mentioned constraints on word processing in Arabic. Several software systems are available for the morphosyntactic analysis of Arabic texts.

[5] For example 'fire', which is feminine in Arabic and 'perfume', which is masculine.

[6] Jackendoff R. (2002), page 133 - 169

Buckwalter's Morphological Analyzer 1.0 (hereafter referred to as "AraMorph") is certainly one of the most popular such systems. Released in 2002, it is also offered as a Java port version, written by Pierrick Brihaye[7]. AraMorph's components are essentially two: the rule engine for morphological analysis and a repository of linguistic resources mainly composed of three lexicons: i) the dictStems lexicon, which contains 38.600 lemmas; ii) the dictPrefixes lexicon, which consists of sequences of proclitics and inflectional prefixes; iii) the dictSuffixes lexicon, which consists of sequences of inflectional suffixes and enclitics. These lexica are accompanied by three compatibility tables used for checking combinations of A (proclitics+prefixes), B (stems) and C (suffixes+enclitics). AraMorph analyzes transliterated Arabic text, and implements an algorithm for morphological analysis and for Part-of-Speech (POS) tagging that includes tokenization, word segmentation, dictionary look-up and compatibility checks. It finally produces an analytic report. In what follows, we consider some of the problems AraMorph encounters in tackling the extensive homography of Arabic written texts.[8] We then move on to our proposed solutions.

3.1 Problems and solutions

Case 1

In processing the written form *yaktub*, Aramorph produces the different parses listed in Table 1.[9]

	Analyses	Lemma
1	***ya-ktub***	*kataba* 'write'
2	**yu-ktab*	
3	**yu-ktib*	*'aktaba* 'dictate'
4	**yu-ktab*	

Table 1 – Aramorph's analyses for "*yaktub"*

Note that the AraMorph engine simply ignores the vowels present in the original spelling, and proposes a number of alternative parses, some of which are simply incompatible with the input form *yaktub*. This is the result of AraMorph's normalization strategy of written texts. To tackle lack of consistency in the Arabic spelling of diacritics, AraMorph gets rid of all diacritics marked in the original text, and parsed undiacritcized forms only. Buckwalter justifies this approach by claiming that writing without diacritics "is a common feature" of Arabic scripts. However, the approach generates spurious output analyses, based on a drastically underspecified spelling.[10] We suggest that diacritics marked in the original text should never be dispensed with, but rather used to filter out the set of candidate parses provided by AraMorph. For this reason, we designed a component assessing the compatibility of the vowel structure of AraMorph multiple parses with the original spelling in the text, to discard all candidates that are not compatible with the original spelling. Another noticeable aspect of Table 1 is that all parses simply ignore omission of the word final vowel in *yaktub*, a vowel used in the Arabic verb system to convey features of time and mood, as shown in example 4 below. This is due to AraMorph's suffix dictionary (dictSuffixes) lacking this information.

Example 4 ***ya-ktub-u***
IPFV.3-read-IND
ya-ktub-a
IPFV.3-read -SBJV
ya-ktub-Ø
IPFV.3-read -JUSS

To improve resulting parses, we augmented AraMorph's prefix and suffix dictionaries with missing information. Furthermore, it was necessary to update compatibility tables.

Case 2

Table 2 shows the analyses output by Aramorph upon processing the spelled-out form *whajt*.

solutions	Analyses	Lemma
1	**wa=hij-tu*	*hāja* 'be agitated'
2	**wa=hij-ta*	
3	**wa=hij-ti*	
4	**wa=hajj-ato*	*hajja* 'burn'
5	***wa=hajj-ato***	*hajjā* 'spell'
6	***wa=haj-ato***	*hajā* 'satirize'
7	**wahaj-tu*	*wahaja* 'burn; spread'
8	**wahaj-ta*	
9	**wahaj-ti*	
10	***wahaj-ato***	

Table 2 – Aramorph's analyses by "*whajt"*

Note that in this case, word segmentation differs depending on the output lemma. In solutions 1-6, each spelled-out form is an inflected form of the verbs *hāja/hajja/hajjā/hajā*, preceded by the clitic conjunction "wa=" (and). Solutions 7-10 are inflected forms of the verb *wahaja*. As in Case 1 parses 1, 2 and 3 may be filtered out if we take into account diacritics in the original spelling.

[7] AraMorph is downloadable from the LDC site at: http://www.nongnu.org/aramorph

[8] Hajder S. R. (2011).

[9] Wrong analyses are marked with an asterisk (*).

[10] Farghaly A., and Shaalan K. (2009).

Beyond these cases, AraMorph outputs further unlikely candidate parses. For example, Buckwalter includes obsolete lexical items[11]. In fact, the fourth proposed analysis is derived from the verb *hajja* that is not used in Arabic[12]. Focusing now on the last four solutions (7-10), they correspond to different inflected forms of the verb *wahaja* depending on what word final vowels are interpolated in the original spelling:

Solution 7 * ***wahaj=tu***
* burn$_{.PST}$=I
*'I burn'

Solution 8 * ***wahaj=ta***
* burn$_{.PST}$=You$_{.M}$
*'You burn'

Solution 9 * ***wahaj=ti***
* burn$_{.PST}$=You$_{.F}$
*'You burn'

Solution 10 ***wahaj-at***
burn$_{.PST}$-she
'She burn'

The inflectional suffixes -tu, -ta, -ti and –at respectively convey 1S, 2SM, 2SF and 3SF. However, we know that the verb *wahaja* requires an inanimate subject. Therefore it cannot be inflected for 1S, 2SM and 2SF. To capture this restriction and cut down on parse overgeneration, one has to enforce further restrictions in compatibility tables, e.g. the verb's ability to accept nominative and accusative pronouns, and to select a rational subject. We then augmented verb entries with subcategorization information such as case assignment and the restriction on rational subjects. At the same time, it was necessary to update compatibility tables. Table 3 shows how many entries are contained in AraMorph's original dictionaries (Original), and how many entries form the current improved version of the same dictionaries (Plus). Note that the number of stems is smaller in Plus than in Original, due to removal of obsolete entries and a number of foreign names that are unlikely to be found in Arabic texts[13]. Table 4 shows compatibility rules for tables AB, AC and BC in both Original and Plus.

AraMorph	entries		
	Prefixes	dictStems	dictSuffixes
Original	299	38600	618
Plus	335	35475	876

Table 3 - Entries in AraMorph's dictionaries

AraMorph	Compatibility		
	Table AB	Table AC	Table BC
Original	1648	598	1285
Plus	2698	1295	2161

Table 4 - Entries in compatibility tables

Finally, Table 5 shows how many parses of the same text[14] are output by AraMorph (Original) and AraMorph Plus. Figures are higher in the former case, in spite of the parser's failure to recognize 656 word tokens, due to lexical gaps in the stem dictionary. In addition, AraMorph Original presents a number of spurious parses. In Plus, on the other hand, restrictions on word grammatical behavior help improve results, and the number of proposed parses significantly decreases, despite Plus more extensive coverage (0 "Not found" parses).

Aramorph	Arabic forms	parses	Not found
Original	9502	21544	656
Plus		20847	0

Table 5 - Arabic text parsing by Original and Plus AraMorph

In addition, original AraMorph presents severely underspecified parses especially concerning morphosyntactic features. By augmenting information in clitics dictionaries and updating compatibility tables, AraMorph Plus provides more thorough morphosyntactic features[15].

4 Conclusion and future research

Automatic text processing requires annotation of different levels of linguistic analysis: morphological, syntactic, semantic and pragmatic. For some languages, like English, it makes sense to analyze those levels in a serial way, by taking the output of an early level of analysis as the input of the ensuing level. Purpose of this article is to demonstrate that specific characteristics of Arabic appear to recommend a different approach. Inflectional, derivational and non-concatenative characteristics of Arabic morphology require interdependence and interaction between different levels of analysis for segmentation of spelled-out forms and their analysis to be adequate. This suggests that Arabic processing may require substantial revision of traditional NLP architectures. For improvement and future work, we plan to complete and refine language resources for Arabic. As a further step, we consider including other contextual factors, such as knowledge about the immediate syntactic context of a word token, as restrictions on diacriticization.

[11] Attia M., Tounsi, L., and Van Genabith J. (2010)
[12] Lisān al-arab. Volume 2, page 170.
[13] Lancioni et al. (2013).
[14] Badawī A. (1966).
[15] Nahli O. (2013).

Reference

Alansary S., Nagi M., and Adly N. (2009). *Towards analysing the international corpus of Arabic (ICA)*. In International conference on language engineering. Progress of Morphological Stage, Egypt. Pp. 241–245.

Alkuhlani S. and Habash N. (2012). *Identifying Broken Plurals, Irregular Gender, and Rationality in Arabic Text*. In Proceeding EACL '12 Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Pages 675-685.

Attia M., Tounsi, L., and Van Genabith J. (2010) *Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic*. Technical Report. The NCLT Seminar Series, DCU, Dublin, Ireland.

Attia M. (2008). *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation*. Ph.D. Thesis. The University of Manchester, Manchester, UK. Pages 35-39.

Attia M. (2002*). Implications of the Agreement Features in Machine Translation*. Phd Thesis. Faculty of Languages and Translation, Al-Azhar University, Cairo, Egypt.

Badawī A. (1966). *'aflūṭīn 'inda-l-'Arab*, Dār al-Nahḍat al-'arabiyya, Cairo.

Bahou Y., Belguith Hadrich L., Aloulou C., and Ben Hamadou A. (2006). *Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés* In Actes du 15e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA'06).

Boudlal A., Lakhouaja A., Mazroui, A., Meziane A., Ould Abdallahi Ould Bebah, M., and Shoul M. (2011). *Alkhalil MorphoSys: A Morphosyntactic analysis system for non-vocalized Arabic,* Seventh International Computing Conference in Arabic (ICCA 2011). Riyadh.

Buckwalter T. (2004*). Issues in Arabic orthography and morphology analysis.* COLING 2004, in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, edited by Ali Farghaly and Karine Megerdoomian, Association for Computational Linguistics, Stroudsburg PA, USA. Pages 31-34.

Dichy J. (1997). *Pour une lexicomatique de l'arabe: l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot*. Meta: journal des traducteurs / Meta: Translators' Journal, vol. 42, n° 2, pages 291-306.

Farghaly A., and Shaalan K. (2009). *Arabic Natural Language Processing: Challenges and Solutions*. Journal ACM Transactions on Asian Language Information Processing (TALIP), Volume 8 Issue 4, December; New York, USA.

Hajder S. R. (2011). *Adapting Standard Open-Source Resources To Tagging A Morphologically Rich Language: A Case Study With Arabic*. Proceedings of the Student Research Workshop associated with RANLP 2011, Hissar, Bulgaria. pages 127–132.

Jackendoff R. (2002). *Foundations of language, Brain, Meaning, Grammar, Evolution*. Published in the United States by Oxford University Press Inc., New York.

Kenneth R. B. (1998). *Arabic morphology using only finite-state operations*. In Proceeding Semitic '98 Proceedings of the Workshop on Computational Approaches to Semitic Languages. Pages 50-57.

Lancioni, G., Pepe, I., Silighini, A., Pettinari, V., Cicola, I., Benassi, L., & Campanelli, M. *Arabic Meaning Extraction through Lexical Resources: A General-Purpose Data Mining Model for Arabic Texts*. IMMM 2013 "The Third International Conference on Advances in Information Mining and Management". Copyright (c) IARIA, 2013. ISBN: 978-1-61208-311-7

Lisān al-arab, edited by Ḥaydar A. and ʿibrāhīm A. Dār al-kutub al-ʿilmiyyah, Beirut, Lebanon.

Manning Christopher D., and Schuetze H. (1999) *Foundations of Statistical Natural Language Processing*. The MIT Press Cambridge, Massachusetts, London, England.

Nahli O. (2013). *Computational contributions for Arabic language processing Part 1. The automatic morphologic analysis of Arabic texts*. In Studia graeco-arabica vol.3, Published by ERC Greek into Arabic Philosophical Concepts and Linguistic Bridges European Research Council Advanced Grant 249431, C. D'Ancona (a cura di), Pacini Editore, Pisa. Pages 195-206. ISSN 2239-012X.

Tsarfaty R., Seddah D., Kubler S., and Nivre J. (2013). *Parsing Morphologically Rich Languages: Introduction to the Special Issue*. Computational Linguistics, Vol. 39, No. 1: 15–22.

Zemirli Z., and Elhadj, Y.O.M. (2012). *Morphar+: an Arabic morphosyntactic analyzer*. In Proceedings of ICACCI. 2012, International Conference on Advances in Computing, Communications and Informatics, CHENNAI, India. ACM New York, NY, USA ©2012. Pages 816-823.

ItEM: A Vector Space Model to Bootstrap an Italian Emotive Lexicon

Lucia C. Passaro, Laura Pollacci, Alessandro Lenci
ColingLab, Dipartimento di Filologia, Letteratura e Linguistica
University of Pisa (Italy)
lucia.passaro@for.unipi.it, laurapollacci.pl@gmail.com,
alessandro.lenci@unipi.it

Abstract

English. In recent years computational linguistics has seen a rising interest in subjectivity, opinions, feelings and emotions. Even though great attention has been given to polarity recognition, the research in emotion detection has had to rely on small emotion resources. In this paper, we present a methodology to build emotive lexicons by jointly exploiting vector space models and human annotation, and we provide the first results of the evaluation with a crowdsourcing experiment.

Italiano. *Negli ultimi anni si è affermato un crescente interesse per soggettività, opinioni e sentimenti. Nonostante sia stato dato molto spazio al riconoscimento della polarità, esistono ancora poche risorse disponibili per il riconoscimento di emozioni. In questo lavoro presentiamo una metodologia per la creazione di un lessico emotivo, sfruttando annotazione manuale e spazi distribuzionali, e forniamo i primi risultati della valutazione effettuata tramite crowdsourcing.*

1 Introduction and related work

In recent years, computational linguistics has seen a rising interest in subjectivity, opinions, feelings and emotions. Such a new trend is leading to the development of novel methods to automatically classify the emotions expressed in an opinionated piece of text (for an overview, see Liu, 2012; Pang and Lee, 2008), as well as to the building of annotated lexical resources like SentiWordNet (Esuli and Sebastiani, 2006; Das and Bandyopadhyay, 2010), WordNet Affect (Strapparava and Valitutti, 2004) or EmoLex (Mohammad and Turney, 2013). Emotion detection can be useful in several applications, e.g. in Customer Relationship Management (CRM) it can be used to track sentiments towards companies and their services, products or others target entities. Another kind of application is in Government Intelligence, to collect people's emotions and points of views about government decisions. The common trait of most of these approaches is a binary categorization of emotions, articulated along the key opposition between POSITIVE and NEGATIVE emotions. Typically, then, these systems would associate words like "rain" and "betray" to the same emotion class in that they both evoke negative emotions, without further distinguishing between the SADNESS-evoking nature of the former and the ANGER-evoking nature of the latter. Emotion lexica, in which lemmas are associated to the emotions they evoke, are valuable resources that can help the development of detection algorithms, for instance as knowledge sources for the building of statistical models and as gold standards for the comparison of existing approaches. Almost all languages but English lack a high-coverage high-quality emotion inventory of this sort. Building these resources is very costly and requires a lot of manual effort by human annotators. On the other hand, connotation is a cultural phenomenon that may vary greatly between languages and between different time spans (Das and Bandyopadhyay, 2010), so that the simple transfer of an emotive lexicon from another language cannot be seen as nothing else than a temporary solution for research purposes.

Crowdsourcing is usually able to speed the process and dramatically lower the cost of human annotation (Snow et al., 2008; Munro et al, 2010). Mohammad and Turney (2010, 2013) show how the "wisdom of the crowds" can be effectively exploited to build a lexicon of emotion associations for more than 24,200 word senses. For the creation of their lexicon, EmoLex, they selected the terms from Macquarie Thesaurus (Bernard, 1986), General Inquirer (Stone et al.,1966), WordNet Affect Lexicon (Strapparava and Valitutti., 2004) and Google n-gram corpus (Brants and Franz, 2006) and they exploited a crowdsourcing experiment, in order to obtain, for every target term, an indication of

its polarity and of its association with one of the eight Plutchik (1994)'s basic emotions (see below). The methodology proposed by Mohammad and Turney (2010, 2013), however, cannot be easily exported to languages where even small emotive lexica are missing. Moreover, a potential problem of a lexicon built solely on crowdsourcing techniques is that its update requires a re-annotation process. In this work we're proposing an approach to address these issues by jointly exploiting corpus-based methods and human annotation. Our output is ItEM, a high-coverage emotion lexicon for Italian, in which each target term is provided of an association score with eight basic emotion. Given the way it is built, ItEM is not only a static lexicon, since it also provides a dynamic method to continuously update the emotion value of words, as well as to increment its coverage. This resource will be comparable in size to EmoLex, with the following advantages: i) minimal use of external resources to collect the seed terms; ii) little annotation work is required to build the lexicon; iii) its update is mostly automatized.

This paper is structured as follows: In section 2, we present ItEM by describing its approach to the seed collection and annotation step, its distributional expansion and its validation. Section 3 reports the results obtained from the validation of the resource using a crowdsourcing experiment.

2 ItEM

Following the approach in Mohammad and Turney (2010, 2013), we borrow our emotions inventory from Plutchik (1994), who distinguishes eight "basic" human emotions: JOY, SADNESS, ANGER, FEAR, TRUST, DISGUST, SURPRISE and ANTICIPATION. Positive characteristics of this classification include the relative low number of distinctions encoded, as well as its being balanced with respect to positive and negative feelings. For instance, an emotive lexicon implementing the Plutchik's taxonomy will encode words like "ridere" (laugh) or "festa" (celebration) as highly associated to JOY while words like "rain" (pioggia) or "povertà" (poverty) will be associated to SADNESS, and words like "rissa" (fight) or "tradimento" (betray) will be encoded as ANGER-evoking entries.

ItEM has been built with a three stage process: In the first phase, we used an online feature elicitation paradigm to collect and annotate a small set of emotional seed lemmas. In a second phase, we exploited distributional semantic methods to expand these seeds and populate ItEM. Finally, our automatically extracted emotive annotations have been evaluated with crowdsourcing.

2.1 Seed collection and annotation

The goal of the first phase is to collect a small lexicon of "emotive lemmas", highly associated to the one or more Plutchik's basic emotions. To address this issue, we used an online feature elicitation paradigm, in which 60 Italian native speakers of different age groups, levels of education, and backgrounds were asked to list, for each of our eight basic emotions, 5 lemmas for each of our Parts-of-Speech (PoS) of interest (Nouns, Adjectives and Verbs). In this way, we collected a lexicon of 347 lemmas strongly associated with one or more Plutchik's emotions. For each lemma, we calculated its emotion distinctiveness as the production frequency of the lemma (i.e. the numbers of subjects that produced it) divided by the number of the emotions for which the lemma was generated. In order to select the best set of seed to use in the bootstrapping step, we only selected from ItEM the terms evoked by a single emotion, having a distinctiveness score equal to 1. In addition, we expanded this set of seeds with the names of the emotions such as the nouns "gioia" (joy) or "rabbia" (anger) and their synonyms attested in WordNet (Fellbaum, 1998), WordNet Affect (Strapparava and Valitutti, 2004) and Treccani Online Dictionary (www.treccani.it/vocabolario).

Emotion	N. of seeds	Adj	Nouns	Verbs
Joy	61	19	26	19
Anger	77	32	30	16
Surprise	60	25	17	22
Disgust	80	40	21	25
Fear	78	37	20	27
Sadness	77	39	22	26
Trust	62	25	21	17
Anticipation	60	15	22	23

Table 1 Distribution of the seeds lemmas

Globally, we selected 555 emotive seeds, whose distribution towards emotion and PoS is described in Table 1.

2.2 Bootstrapping ItEM

The seed lemmas collected in the first phase have been used to bootstrap ItEM using a corpus-based model inspired to Turney and Littmann (2003) to automatically infer the semantic orientation of a word from its distributional similarity with a set of positive and negative paradigm

words. Even if we employ a bigger number of emotion classes, our model is based on the same assumption that, in a vector space model (Sahlgren, 2006; Pantel and Turney, 2010), words tend to share the same connotation of their neighbours. We extracted from the La Repubblica corpus (Baroni et al, 2004) and itWaC (Baroni et al., 2009), the list T of the 30,000 most frequent nouns, verbs and adjectives, which were used as target and contexts in a matrix of co-occurrences extracted within a five word window (±2 words, centered on the target lemma, before removing the words not belonging to T). Differently from the Turney and Littmann (2003)'s proposal, however, we did not calculate our scores by computing the similarity of each new vector against the whole sets of seed terms. On the contrary, for each <*emotion, PoS*> pair we built a centroid vector from the vectors of the seeds belonging to that emotion and PoS, obtaining in total 24 centroids. We constructed different word spaces according to PoS because the context that best captures the meaning of a word, differs depending on the word to be represented (Rothenhäusler and Schütze, 2007). Finally, our emotionality scores have been calculated on the basis of the distance between the new lemmas and the centroid vectors. In this way, each target term received a score for each basic emotion. In order to build the vector space model, we reweighted the co-occurrence matrix using the Positive Pointwise Mutual Information (Church and Hanks, 1990), which works well for both word–context matrices (Pantel & Lin, 2002a) and term–document matrices (Pantel & Lin, 2002b). In particular, we used the Positive PMI (PPMI), in which negative scores are changed to zero, and only positive ones are considered (Niwa & Nitta, 1994). We followed the approach in Polajnar and Clark (2014) by selecting the top 240 contexts for each target word. Finally, we calculated the emotive score for a target word as the cosine similarity with the corresponding centroid (e.g. the centroid of "JOY-nouns").

The output of this stage is a list of words ranked according to their emotive score. Appendix A shows the top most associated adjectives, nouns and verbs in ItEM. As expected, a lot of target words have a high association score with more than one emotive class, and therefore some centroids are less discriminating because they have a similar distributional profile. Figure 1 shows the cosine similarity between the emotive centroids: we can observe, for example, a high similarity between SADNESS and FEAR, as well as between SURPRISE and JOY. This is consistent with the close relatedness between these emotions.

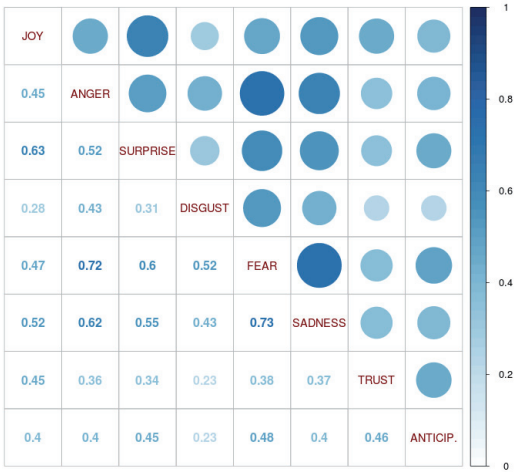


Figure 1 Cosine similarity between the emotive centroids

2.3 Validation

We evaluated our procedure using a two-step crowdsourcing approach: in the first step, for each <*emotion, PoS*> pair we ranked the target words with respect to their cosine similarity with the corresponding emotive centroid. We then selected the top 50 words for each centroid and we asked the annotators to provide an emotive score: Given a target word <*w*>, for each Plutchik's emotion <*e*>, three annotators were asked to answer the question "How much is <w> associated with the emotion <e>?". The annotators had to choose a score ranging from 1 (not associated) to 5 (highly associated). Since very often the words may be associated with more than one emotion, we wanted to estimate the average degree of association between the word and the various emotions. Empirically, we defined the best distinctiveness score *d* as follows:

$$d = max1 - max2me * (max1 - mn)$$

Where *max1* is the highest emotive association for the target word, *max2* is the second higher value, *me* is the association score between the target word and the target emotion, and *mn* is the average of the evaluations for the word across the emotion classes. This formula captures the intuition that a word is distinctive for a target emotion if its association degree with the target emotion is high as well as its association degree with the other classes is low. After ranking the words over this association score, we selected the top 10 distinctive nouns, adjectives and verbs for each <*emotion, PoS*> pair, in order to further expand the set of the seeds used to build the distributional space. For this second run, we removed the words belonging the top 10 of more than one emotion and we added the remaining

192 words to the set of the seeds used to build the centroid emotive vectors, using the procedure described in section 2.2. The second run allows us to evaluate the quality of the initial seeds and to discover new highly emotive words.

3 Results

We have evaluated the precision of our distributional method to find words correctly associated with a given emotion, as well the effect of the incremental process of seed expansion. In particular, we evaluated the top 50 nouns, adjectives and verbs for each emotion. Precision@50 has been calculated by comparing the vector space model's candidates against the annotation obtained with crowdsourcing. True positives (TP) are the words found in the top 50 neighbours for a particular emotion and PoS, for which the annotators provided a average association score greater than 3. False positives (FP) are the words found in the top 50 nouns, adjectives and verbs, but for which the aggregate evaluation of the evaluators is equal or lower than 3. Table 2 shows the Precision by emotion in the first run (P Run 1) and in the second one (P Run 2), calculated on a total of 1,200 target associations.

Emotion	P (Run 1)	P (Run 2)
Joy	**0.787**	0.767
Anger	0.813	**0.827**
Surprise	**0.573**	0.56
Disgust	**0.78**	0.753
Fear	0.673	**0.727**
Sadness	**0.827**	0.793
Trust	0.43	**0.5**
Anticipation	**0.557**	0.527
Micro AVG	0.68	**0.682**

Table 2 Precisionby Emotion (Runs 1 and 2)

If we analyze the same results by aggregating the Precision by PoS (Table 3), we can notice some differences between the first and the second run. Although overall there is a slight increase of the Precision score, this growth only affects verbs and adjectives. This is probably due to the way in which the noun seeds are distributed around the emotion centroids: a lot of them, in fact, are strongly associated to more than one emotion.

PoS	P (Run 1)	P (Run 2)
Adjectives	0.727	**0.735**
Nouns	**0.685**	0.675
Verbs	0.629	**0.635**

Table 3 Precision by PoS (Runs 1 and 2)

To appreciate the gain obtained in the second run, we analyzed the medium change of cosine similarity between the first and the second experiment, and we noticed that the true positives have, on average, a higher cosine similarity with the corresponding emotive centroid in the second run (cf. Table 4). This proves the positive effect produced by the new seeds discovered by the distributional model in the first run.

Emotion	CosR1	CosR2	CosR2-CosR1
Joy	0.564	0.595	+0.032
Anger	0.582	0.6	+0.018
Surprise	0.635	0.657	+0.022
Disgust	0.524	0.555	+0.034
Fear	0.616	0.613	-0.003
Sadness	0.612	0.648	+0.036
Trust	0.575	0.665	+0.103
Anticipation	0.54	0.563	+0.027
Macro Avg	0.581	0.612	+0.034

Table 4 Increase of cosine similarity

In general, the distributional method is able to achieve very high levels of precision, despite an important variance among emotion types. Some of them (e.g., ANTICIPATION) confirm to be quite hard, possibly due to a higher degree of vagueness in their definition that might also affect the intuition of the evaluators.

The results that we achieved for the different emotions and PoS show that additional research is needed to improve the seed selection phase, as well as the tuning of the distributional space.

4 Conclusion

What we are proposing with ItEM is a reliable methodology that can be very useful for languages that lack lexical resources for emotion detection, and that is at the same time scalable and reliable. Moreover, the resulting resource can be easily updated by means of fully automatic corpus-based algorithms that do not require further work by human annotators, a vantage that can turn out to be crucial in the study of a very unstable phenomenon like emotional connotation.

The results of the evaluation with crowdsourcing show that a seed-based distributional semantic model is able to produce high quality emotion scores for the target words, which can also be used to dynamically expand and refine the emotion tagging process.

Reference

Baroni M. and Lenci A. (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. In Computational Linguistics, 36 (4), pp. 673-721.

Baroni M., Bernardini S., Comastri F., Piccioni L., Volpi A., Aston G. and Mazzoleni M. (2004). Introducing the "la Repubblica" Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In Proceedings of LREC 2004.

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): 209-22

Bradley, M. and Lang P. (1999) Affective norms for english words (ANEW): Instruction manual and affective ratings. Technical Report, C-1, The Center for Research in Psychophysiology, University of Florida.

Brants T. and Franz A. (2006). Web 1t 5-gram version 1. Linguistic Data Consortium.

Das A. and Bandyopadhyay S. (2010). Towards the Global SentiWordNet. In Proceedings of PACLIC 2010, pp. 799-808.

Esuli A. and Sebastiani F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC 2006.

Fellbaum C. (1998).WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press

Liu B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

Mohammad S. M. and Turney P. D. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26-34.

Mohammad S. M. and Turney P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. In Computational Intelligence, 29 (3), pp. 436-465.

Munro R., Bethard S., Kuperman V., Lai V., Melnick R., Potts C., Schnoebelenand T., Tily H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.

Niwa, Y., & Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In Proceedings of the 15th International Conference On Computational Linguistics , pp. 304-309, Kyoto, Japan.

Pang B. and Lee L. (2008). Opinion mining and sentiment analysis. In Foundations and trends in Information Retrieval, 2 (1-2), pp. 1-135.

Pantel, P., & Lin, D. (2002a). Discovering word senses from text. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 613–619, Edmonton, Canada.

Pantel, P., & Lin, D. (2002b). Document clustering with committees. In Proceedings of the 25th Annual International ACM SIGIR Conference, pp. 199–206.

Plutchik R. (1994) The psychology and biology of emotion. Harper Collins, New York.

Polajnar T., and Clark S. (2014). Improving Distributional Semantic Vectors through Context Selection and Normalisation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), pp. 230-238, Gothenburg, Sweden, 2014.

Rothenhäusler, K. and Schütze, H. (2007) Part of Speech Filtered Word Spaces. In Proceedings of the Sixth International and Interdisciplinary Conference on Modeling and Using Context, Roskilde, Denmark, August 20-24, 2007.

Sahlgren M. (2006) The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. dissertation, Stockholm University.

Snow R., O'Connor, B. Jurafsky D, Ng, A. (2008) Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In Proceedings of EMNLP 2008, pp. 254-263.

Stone P., Dunphy D.C., Smith M.S. and Ogilvie D.M. (1966). The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge, MA.

Strapparava C. and Valitutti A. (2004) Wordnet-Affect: An affective extension of WordNet. Proceedings of LREC-2004, pp. 1083-1086.

Turney P. D and Pantel P. (2010). From frequency to meaning: vector space models for semantics. In Journal of Artificial Intelligence Research, 37, pp. 141-188.

Turney P.D. and Littman M.L. (2003) Measuring praise and criticism: Inference of semantic orientation from association. In ACM Transactions on Information Systems.

Appendix A: Top 5 adjectives, verbs and nouns for each emotion, with their association scores, calculated as the cosine similarity between the word and the corresponding centroid vector.

EMOTION	ADJECTIVES	COSINE	VERBS	COSINE	NOUNS	COSINE
JOY	gioioso (*joyful*)	0.85	rallegrare (*to make happy*)	0.6	gioia (*joy*)	0.83
	scanzonato (*easygoing*)	0.68	consolare (*to comfort*)	0.54	ilarità (*cheerfulness*)	0.73
	spiritoso (*funny*)	0.66	apprezzare (*to appraise*)	0.53	tenerezza (*tenderness*)	0.72
	scherzoso (*joking*)	0.65	applaudire (*to applaud*)	0.53	meraviglia (*astonishment*)	0.7
	disinvolto (*relaxed*)	0.62	rammentare (*to remind*)	0.53	commozione (*deep feeling*)	0.69
ANGER	insofferente (*intolerant*)	0.72	inveire (*to inveigh*)	0.59	impazienza (*impatience*)	0.8
	impaziente (*anxious*)	0.67	maltrattare (*totreatbadly*)	0.58	dispetto (*prank*)	0.76
	permaloso (*prickly*)	0.66	offendere (*to offend*)	0.56	rancore (*resentment*)	0.75
	geloso (*jealous*)	0.66	ingiuriare (*to vituperate*)	0.53	insofferenza (*intolerance*)	0.74
	antipatico (*unpleasant*)	0.65	bastonare (*to beat with a cane*)	0.52	antipatia (*impatience*)	0.74
SURPRISE	perplesso (*perplexed*)	0.81	stupefare (*to amaze*)	0.82	sgomento (*dismay*)	0.74
	sgomento (*dismayed*)	0.73	sconcertare (*to disconcert*)	0.81	trepidazione (*trepidation*)	0.74
	allibito (*shocked*)	0.73	rimanere (*to remain*)	0.79	turbamento (*turmoil*)	0.74
	preoccupato (*worried*)	0.72	indignare (*to makeindignant*)	0.74	commozione (*deep feeling*)	0.74
	sconvolto (*upset*)	0.72	guardare (*to look*)	0.73	presentimento (*presentiment*)	0.73
DISGUST	immondo (*dirty*)	0.6	scandalizzare (*to shock*)	0.63	fetore (*stink*)	0.84
	malsano (*unhealthy*)	0.58	indignare (*to makeindignant*)	0.53	escremento (*excrement*)	0.83
	insopportabile (*intolerable*)	0.58	disapprovare (*to disapprove*)	0.5	putrefazione (*rot*)	0.82
	orribile (*horrible*)	0.56	criticare (*to criticize*)	0.49	carogna (*lowlife*)	0.74
	indegno (*shameful*)	0.52	biasimare (*to blame*)	0.49	miasma (*miasma*)	0.74
FEAR	impotente (*helpless*)	0.6	stupefare (*to amaze*)	0.7	disorientamento (*disorientation*)	0.82
	inquieto (*restless*)	0.57	scioccare (*to shock*)	0.68	angoscia (*anguish*)	0.81
	infelice (*unhappy*)	0.55	sbalordire (*to astonish*)	0.68	turbamento (*turmoil*)	0.79
	diffidente (*suspicious*)	0.53	sconcertare (*to disconcert*)	0.66	prostrazione (*obeisance*)	0.79
	spaesato (*disoriented*)	0.53	disorientare (*to disorient*)	0.65	inquietudine (*apprehension*)	0.78
SADNESS	triste (*sad*)	0.8	deludere (*to betray*)	0.78	tristezza (*sadness*)	0.91
	tetro (*gloomy*)	0.65	amareggiare (*to embitter*)	0.75	sconforto (*discouragement*)	0.88
	sconsolato (*surrowful*)	0.62	angosciare (*to anguish*)	0.72	disperazione (*desperation*)	0.88
	pessimistico (*pessimistic*)	0.61	frustrare (*to frustrate*)	0.71	angoscia (*anguish*)	0.88
	angoscioso (*anguished*)	0.59	sfiduciare (*to discourage*)	0.71	inquietudine (*apprehension*)	0.87
TRUST	disinteressato (*disinterested*)	0.65	domandare (*to ask*)	0.64	serietà (*seriousness*)	0.91
	rispettoso (*respectful*)	0.65	dubitare (*to doubt*)	0.59	prudenza (*caution*)	0.9
	laborioso (*hard-working*)	0.64	meravigliare (*to amaze*)	0.58	mitezza (*mildness*)	0.89
	disciplinato (*disciplined*)	0.63	rammentare (*to remind*)	0.56	costanza (*tenacity*)	0.89
	zelante (*zealous*)	0.62	supporre (*to suppose*)	0.56	abnegazione (*abnegation*)	0.88
ANTICIPATION	inquieto (*agitated*)	0.7	sforzare (*to force*)	0.56	oracolo (*oracle*)	0.77
	ansioso (*anxious*)	0.58	confortare (*to comfort*)	0.56	premonizione (*premonition*)	0.74
	desideroso (*desirous*)	0.56	degnare (*to deign*)	0.55	preveggenza (*presage*)	0.73
	entusiasta (*enthusiastic*)	0.56	distogliere (*to deflect*)	0.55	auspicio (*auspice*)	0.72
	dubbioso (*uncertain*)	0.55	appagare (*to satiate*)	0.54	arcano (*aracane*)	0.71

Somewhere between Valency Frames and Synsets. Comparing Latin *Vallex* and Latin WordNet

Marco Passarotti, Berta González Saavedra, Christophe Onambélé Manga
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 – 20123 Milan, Italy
{marco.passarotti, berta.gonzalezsaavedra,
christophe.onambele}@unicatt.it

Abstract

English. Following a comparison of the different views on lexical meaning conveyed by the Latin WordNet and by a treebank-based valency lexicon for Latin, the paper evaluates the degree of overlapping between a number of homogeneous lexical subsets extracted from the two resources.

Italiano. *Alla luce di un confronto tra gli approcci al significato lessicale realizzati dal WordNet latino e da un lessico di valenza prodotto sulla base di due treebank latine, l'articolo descrive la valutazione del grado di sovrapposizione tra alcuni sottoinsiemi lessicali omogenei estratti dalle due risorse lessicali.*

1 Introduction

Several lexical resources are today available for many languages, ranging from dictionaries to wordnets, ontologies, valency lexica and others. Although such resources deal with the same basic constituents, i.e. lexical entries, these are organized according to different criteria, corresponding to different views on lexicon and, in particular, on lexical meaning.

On the one hand, a widespread approach to lexical meaning comes from the basic assumption of frame semantics (Fillmore, 1982), according to which the meaning of some words can be fully understood only by knowing the frame elements that are evoked by those words. Following such an assumption, there is a large use of the concept of *valency* and of labels for semantic roles in lexical resources. The degree of semantic granularity of the set of semantic roles used is what mostly distinguishes resources like PropBank, VerbNet and FrameNet one from the other.

On the other hand, a lexical resource largely used in both theoretical and computational linguistics is WordNet, which is centred on the idea of synonymy in the broad sense. Words are included in *synsets*, which are sets of words "that are interchangeable in some context without changing the truth value of the proposition in which they are embedded" (from the glossary of WordNet: http://wordnet.princeton.edu).

Despite their differences, these two views are not incompatible. Over the last decade, several attempts at linking different lexical resources together have been launched. One of the best known projects is Semlink, which makes use of a set of mappings to link PropBank, VerbNet, FrameNet and WordNet (Palmer, 2009). Pazienza et alii (2006) study the semantics of verb relations by mixing WordNet, VerbNet and PropBank. Shi and Mihalcea (2005) integrate FrameNet, VerbNet and WordNet into one knowledge-base for semantic parsing purposes.

Regarding the relations between valency lexica and wordnets, Hlavácková (2007) describes the merging of the Czech WordNet (CWN) with the database of verb valency frames for Czech VerbaLex, whose lexical entries are related to each other according to the CWN synsets. Hajič et alii (2004) use CWN while performing the lexico-semantic annotation of the Prague Dependency Treebank for Czech (PDT), which is in turn exploited to improve the quality and the coverage of CWN. In order to pick out the semantic constraints of the verbal arguments in the Polish WordNet (PolNet), the valency structure of verbs is used as a property of verbal synsets, because it is "one of the formal indices of the meaning (it is so that all members of a given synset share the valency structure)" (Vetulani and Kochanowski, 2014, page 402).

Despite a centuries-long tradition in lexicography, the development of state-of-the-art computational lexical resources for Latin is still in its infancy. However, some fundamental resources were built over the last decade. Among them are a WordNet and a treebank-based valency lexicon. In this paper, we present the first steps towards a comparison of these two resources, by evaluating the degree of overlapping of a number of their lexical subsets.

2 The Lexical Resources

2.1 The Latin Valency Lexicon *Vallex*

The Latin valency lexicon *Vallex* (LV; González Saavedra and Passarotti, forthcoming) was developed while performing the semantic annotation of two Latin treebanks, namely the *Index Thomisticus* Treebank, which includes works of Thomas Aquinas (Passarotti, 2014), and the Latin Dependency Treebank, which features works of different authors of the Classical era (Bamman and Crane, 2006). All valency-capable lemmas occurring in the semantically annotated portion of the two treebanks are assigned one lexical entry and one valency frame in LV.

The structure of the lexicon resembles that of the valency lexicon for Czech *PDT-Vallex* (Hajič et al., 2003). On the topmost level, the lexicon is divided into lexical entries. Each entry consists of a sequence of frame entries relevant for the lemma in question. A frame entry contains a sequence of frame slots, each corresponding to one argument of the given lemma. Each frame slot is assigned a semantic role. The set of semantic roles is the same used for the semantic annotation of the PDT (Mikulová et al., 2005). Since the development of the lexicon is directly related to textual annotation, the surface form of the semantic roles run across during the annotation is recorded as well.

Presently, LV includes 983 lexical entries and 2,062 frames: 760 verbs (1,728 frames), 161 nouns (263 frames), 60 adjectives (68 frames), and 2 adverbs (3 frames).

2.2 The Latin WordNet

The Latin WordNet (LWN; Minozzi, 2010) was built in the context of the MultiWordNet project (Pianta et al., 2002), whose aim was to build a number of semantic neworks for specific languages aligned with the synsets of Princeton WordNet (PWN). The language-specific synsets are built by importing the semantic relation among the synsets for English provided by PWN.

At the moment, LWN includes 8,973 synsets and 9,124 lemmas (4,777 nouns; 2,609 verbs; 1,259 adjectives; 479 adverbs).

3 Comparing the Lexical Resources

3.1 Method

To provide a basic understanding of the differences and similarities between the views on lexical meaning pursued by LV and LWN, we evaluate the degree of overlapping between some lexical subsets extracted from the two resources.

The lexical subsets of LWN that we use are the synsets, while for LV they are groups of words (lemmas) that share the same properties of arguments at frame entry level. For each subset extracted from LV we calculate its degree of overlapping with the synsets of LWN. The maximum overlapping holds when all the words belonging to the same LV subset do occur in the

same synset(s) of LWN. Conversely, the minimum overlapping holds when no word of an LV subset shares the same synset with any of the other words of the same LV subset, i.e. when all the words of a LV subset are "single" words. Our starting point, thus, are the LV subsets, whose contents are compared with the synsets of LWN.

We use two metrics to evaluate the results: (a) the number of single words, and that of couples, triples … *n*-tuples of words in the LV subset that share the same LWN synset(s) (if the same word occurs in more *n*-tuples, the *n*-tuple with the higher value of *n* is considered); (b) the number of words in the LV subset that share the same LWN synset with *n* words of the same group ("connection degree").

Loosely speaking, a good overlapping degree between an LV subset and the LWN synsets is given by (a) a low percentage of singles, (b) a high number of couples and *n*-tuples (this being as more meaningful as the value of *n* is higher) and (c) a high number of words with high connection degree.

3.2 Selecting the *Vallex* Subsets

LV subsets include words that share the same properties of arguments at frame entry level. We use frame entries instead of lexical entries because the frame is the level of the lexical entry that is mostly bound to meaning, a frame entry usually corresponding to one of the word's senses. We focus on verbal entries only, as verbs are the most valency-capable words and the best represented PoS in LV.

Three selection criteria for LV subsets are at work: the quality of the arguments (i.e. their semantic role), their quantity and their surface form. For reasons of space, the groups discussed here are only a small selection of those that we built. In particular, we focus on a number of "semantically rich" frame entries, which feature such semantic roles (and some morphological features of them) that are expected to select verbs with a substantial degree of common semantic properties. According to these criteria, we selected the following LV subsets:

(A) frame entries with three arguments: (a) an Actor, a Patient and a Direction-To (A_P_TO) and (b) an Actor, a Patient and an Addressee (A_P_AD);

(B) the subsets in (A) are further specified by the following: an Actor, a Patient and a Direction-To expressed by a prepositional phrase headed by the preposition (a) *ad* (A_P_TO-ad) and (b) *in* (A_P_TO-in); (c) an Actor, a Patient and an Addressee expressed by a noun in dative (A_P_AD-dat);

(C) frame entries featuring a Patient expressed (a) by a verbal phrase (P-VP) and (b) by a conjunction phrase (P-VP-conj), the latter being a subset of the former.

3.3 Results and Discussion

For each LV subset, table 1 shows the number of its members (column "W[ords]"), the percentage of them occurring in LWN ("Cover[age]"), the number of single words ("Singles") and their percentage ("S%"). For instance, there are five singles in the A_P_TO subset: *conduco* (*to drive*), *educo,-ere* (*to lead out*), *immergo* (*to dip*), *instigo* (*to incite*), *termino* (*to limit*)[1].

Table 1 shows also the number of couples ("Couples") and *n*-tuples ("*n*ples") for each LV subset. For instance, the A_P_AD subset features one sextuple, i.e. six members of this subset share the same LWN synset: *doceo* (*to teach*), *exhibeo* (*to present*), *offero,-erre* (*to offer*), *ostendo*, (*to show*), *praebeo* (*to offer*), *praesto* (*to offer*).

Finally, table 1 shows the connection degree values (columns "1" to "10"). For instance, there are two words (*do - to give* - and *offero,-erre - to*

[1] Given the difference in size between LV and LWN , we considered only those subsets having a coverage ≥ 0.7. The English translations provided here report the sense of the Latin word in the frame concerned. For instance, *termino* has two main senses, which are conveyed by two different frames: A_P (*to mark the boundaries of*) and A_P_TO (*to limit*).

offer -) in the A_P_AD subset that share the same synset at least once with nine different words belonging to the same synset. The nine words sharing the same synset with *do* are: *attribuo* (*to assign*), *dedo (to consign*), *largior* (*to donate*), *mitto* (*to send*), *offero,-erre* (*to offer*), *perhibeo* (*to present*), *praebeo* (*to offer*), *refero* (*to report*), *tribuo* (*to assign*).

According to the results, we can organize the LV subsets into three groups by overlapping degree. Low overlapping is shown by the subsets that include the Direction-To argument, A_P_TO-in and A_P_TO-ad presenting lower overlapping with LWN than A_P_TO. The two subsets featuring a Patient expressed by a VP (P-VP and P-VP-conj) show medium overlapping. The highest overlapping degree holds when the subsets with the Addressee argument are concerned (A_P_AD and A_P_AD-dat).

The results show a correspondence between the level of granularity of the semantic roles of the LV frame entries and the overlapping degree. Since Actor and Patient are quite semantically poor labels and they are common to all A_P_TO and A_P_AD subsets, it is the more fine-grained (and more strictly selecting) meaning of the Addressee than that of the Direction-To argument to make the A_P_AD subsets more overlapping with LWN than the A_P_TO ones.

Another aspect that biases the overlapping degree between LV and LWN are some morphological features of the semantic roles. A Patient expressed by a verbal phrase performs a quite strict selection of the verbs that can have such a construction. These verbs must be able to subcategorize a Patient that is an event or a state expressed by a verb. Several of them tend to be verbs of perception and cognition, like for instance verba dicendi, putandi and sentiendi. In fact, one of the quadruples of the P-VP subset includes four verba putandi: *cogito* (*to think*), *credo* (*to believe*), *opinor* (*to suppose*), *suspicor* (*to suspect*).

4 Conclusion and Future Work

Although a valency lexicon like LV accounts for the different senses that one word may have by assigning it different frame entries, these are not as much semantically defined as the LWN synsets are. However, there is a certain degree of correspondence between these two resources: the more/less fine-grained a frame-based LV subset is, the higher/lower its overlapping with the LWN synsets. For instance, LV includes 1,060 frame entries of verbs formed by an Actor and a Patient: such a subset is both too large and semantically coarse-grained to allow for a sufficient overlapping with the LWN synsets. For this reason, while evaluating the overlapping degree between LV and LWN, we have first focussed on a number of "semantically rich" LV subsets. The evaluation metrics that we used are still very simple. The values of the *n*-tuples must be weighted at evaluation stage (one sextuple is "heavier" that one triple) and the lexical subsets of LWN must be extended beyond synonymy, by exploiting also other relations between words, like hyperonymy and hyponymy.

LV_Subset	W	Cover	Singles	S%	Couples	3ples	4ples	5ples	6ples	1	2	3	4	5	6	7	8	9	10
A_P_TO	24	0.833	5	0.25	16	1	-	-	-	8	4	2	1	-	-	-	-	-	-
A_P_AD	55	0.8	4	0.091	49	9	7	4	1	9	6	6	6	2	4	3	1	2	-
A_P_TO-ad	21	0.905	9	0.474	7	1	-	-	-	5	2	1	1	-	-	-	-	-	-
A_P_TO-in	17	1	9	0.529	4	-	-	-	-	6	1	-	-	-	-	-	-	-	-
A_P_AD-dat	35	0.8	5	0.178	31	5	8	2	-	3	2	6	6	3	3	-	1	-	-
P-VP	100	0.71	19	0.268	89	17	8	1	-	13	7	11	7	3	5	3	3	-	1
P-VP-conj	30	0.833	6	0.24	30	5	2	-	-	7	4	6	1	-	-	1	-	-	-

Table 1. Coverage, Singles and *n*-tuples, Connection Degree

References

D. Bamman, D. and G. Crane. 2006. The design and use of a Latin dependency treebank. J. Nivre and J. Hajič (eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, 67-78.

C. Fillmore. 1982. Frame semantics. *Linguistics in the Morning Calm*. Hanshin Publishing Co., Seoul, 111-137.

B. González Saavedra B. and M. Passarotti. Forthcoming. Verso un lessico di valenza del latino empiricamente motivato. *Atti del Workshop SLI "Dati empirici e risorse lessicali", La Valletta, Malta, 25 Settembre 2015*.

J. Hajič, M. Holub, M. Hučínová and M. Pavlík. 2004. Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank. *Proceedings of the Workshop on "Building Lexical Resources from Semantically Annotated Corpora" at LREC 2004*, 25-30.

J. Hajič, J. Panevová, Z. Urešová, A. Bémová, V. Kolárová and P. Pajas. 2003. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. J. Nivre and E. Hinrichs (eds.), *Proceedings of the second workshop on treebanks and linguistic theories*, 57-68.

D. Hlavácková. 2007. The Relations between Semantic Roles and Semantic Classes in VerbaLex. P. Sojka and A. Horák (eds.), *RASLAN 2007 Recent Advances in Slavonic Natural Language Processing*, 97.

M. Mikulová et alii. 2005. Annotation on the tectogrammatical layer in the Prague Dependency Treebank. The Annotation Guidelines. Available at http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf.

S. Minozzi. 2010. The Latin WordNet project. In Latin Linguistics Today. Latin Linguistics Today. P. Anreiter and M. Kienpointner (eds.), *Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik, 4.-9. April 2009*. Innsbrucker Beiträge zur Sprachwissenschaft, Innsbruck, 707-716.

M. Palmer. 2009. Semlink: Combining English lexical resources. *Proceedings of the 5th. International Workshop on Generative Approaches to the Lexicon (GL2009)*, 9-15.

M. Passarotti. 2014. From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin. K. Zervanou and C. Vertan (eds.), *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014. April 26, 2014. Gothenburg, Sweden*, 100-109.

M.T. Pazienza, M. Pennacchiotti and F.M. Zanzotto. 2006. Mixing wordnet, verbnet and propbank for studying verb relations. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, 1372-1377.

E. Pianta, L. Bentivogli and C. Girardi. 2002. MultiWordNet: developing an aligned multilingual database. *Proceedings of the first international conference on global WordNet*, Vol. 152, 55-63.

L. Shi and R. Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. *Computational linguistics and intelligent text processing*. Springer, Berlin Heidelberg, 100-111.

Z. Vetulani and B. Kochanowski. 2014. "PolNet-Polish WordNet" project: PolNet 2.0-a short description of the release. *Proceedings of the Global Wordnet Conference*, 400-404.

SentIta and Doxa: Italian Databases and Tools for Sentiment Analysis Purposes

Serena Pelosi
Department of Political, Social
and Communication Science
University of Salerno
spelosi@unisa.it

Abstract

English. This reserach presents SentIta, a Sentiment lexicon for the Italian language, and Doxa, a prototype that, interacting with the lexical database, applies a set of linguistic rules for the Document-level Opinionated teXt Analysis. Details about the dictionary population, the semantic analysis of texts written in natural language and the evaluation of the tools will be provided in the paper.

Italiano. *Questa ricerca presenta SentIta, un lessico dei Sentimenti per l'Italiano, e Doxa, un prototipo che, interagendo con il lessico, applica una serie di regole linguistiche per l'analisi di testi contenenti opinioni. Dettagli in merito al popolamento del dizionario, all'analisi semantica di testi scritti in linguaggio naturale e alla valutazione degli strumenti utilizzati saranno forniti nell'articolo.*

1 Introduction

Through online customer review systems, Internet forums, discussion groups and blogs, consumers are allowed to share positive or negative information than can influence in different ways the purchase decisions and model the buyer expectations, above all with regard to experience goods (Nakayama et al., 2010); such as hotels (Ye et al., 2011), restaurants (Zhang et al., 2010), movies (Duan et al., 2008), books (Chevalier and Mayzlin, 2006) or videogames (Zhu and Zhang, 2006).

The consumers, as Internet users, can freely share their thoughts with huge and geographically dispersed groups of people, competing, this way, with the traditional power of marketing and advertising channels.

Differently from the traditional word-of-mouth, which is usually limited to private conversations, the user generated contents on Internet can be directly observed and described by the researchers.

The present paper will provide in Section 2 a concise overview on the most popular techniques for both sentiment analysis and polarity lexicon propagation. Afterward, it will describe in Section 3 the method used to semi-automatically populate SentIta, our Italian sentiment lexicon and in Section 4 the rules exploited to put the words' polarity in context . In the end, Sections 5 will describe our opinion analyzer Doxa, that performs document-level sentiment analysis, sentiment role labeling and feature-based opinion summarization.

2 State of the Art

The most used approaches in sentiment analysis include, among others, the lexicon-based methods, which are grounded on the idea that the text semantic orientation can be inferred by the orientation of words and phrases it contains.

In literature, polarity indicators are usually adjectives or adjective phrases (Hatzivassiloglou and McKeown, 1997; Hu and Liu, 2004; Taboada et al., 2006); but recently also the use of adverbs (Benamara et al., 2007), nouns (Vermeij, 2005; Riloff et al., 2003) and verbs (Neviarouskaya et al., 2009) became really common.

Among the most popular lexicons for the sentiment analysis, WordNet-Affect (Strapparava et al., 2004), SentiWordNet (Esuli and Sebastiani, 2006) and SentiFul (Neviarouskaya et al., 2011) deserve to be cited. Because the largest part of the state of the art lexicons focuses on the English language, Italian lexical databases are mostly created by translating and adapting the English ones (Steinberger et al., 2012; Baldoni et al., 2012; Basile and Nissim, 2013; Hernandez-Farias et al., 2014).

As regards the works on lexicon propagation, we mention thesauri-based works (Kim and Hovy,

2004; Esuli and Sebastiani, 2006; Hassan and Radev, 2010; Maks and Vossen, 2011), corpus-based approaches (Turney, 2002; Baroni and Vegnaduzzo, 2004; Qiu et al., 2009; Wawer, 2012) and morphological strategies (Moilanen and Pulman, 2008; Ku et al., 2009; Neviarouskaya, 2010; Wang et al., 2011).

Due to the strong impact of the syntactic structures in which the lemmas occur on the resulting polarity of the sentences, we can find in literature many studies on contextual polarity shifters, e.g. Choi and Cardie (2008), Benamara et al. (Benamara et al., 2012) on negation; Kennedy and Inkpen (2006) and Polanyi and Zaenen (2006) on intensification; Taboada et al. (2011) and Narayanan et al. (2009) on irrealis markers and conditional tenses.

3 Prior Polarity: Building and Propagating Italian Lexical Resources

In this work we created SentIta, a Sentiment lexicon for the Italian language, that has been semi-automatically generated on the base of the richness of the Italian lexical databases of Nooj[1] (Silberztein, 2003; Vietri, 2014) and the Italian Lexicon-grammar (LG) resources[2] (Elia et al., 1981; Elia, 1984).

The tagset used for the Prior Polarity annotation (Osgood, 1952) of the lexical resources is composed of four tags (POS *positive*; NEG *negative*; FORTE *intense* and DEB *weak*), that combined together generate an evaluation scale that goes from -3 (+NEG+FORTE) to +3 (+POS+FORTE) and a strength scale that ranges from -1 (+DEB) to +1 (+FORTE). Neutral words have been excluded from the lexicon.

Details about the lexical asset available for the Italian language is summarized in Table 1.

Because hand-built lexicons are definitely more accurate than the automatically-built ones, especially in cross-domain sentiment analysis tasks (Taboada et al., 2011; Bloom, 2011), we started the creation of the SentIta database with the manual tagging of part of the lemmas contained in the Nooj Italian dictionaries. The adjectives and the bad words have been manually extracted and evaluated starting from the Nooj databases, preserving their inflectional (FLX) and derivational (DRV) properties.

Category	Entries	Example
Adjectives	5.383	*allegro*
Adverbs	3.626	*tristemente*
Compound Adv	793	*a gonfie vele*
Idioms	552	*essere in difetto*
Nouns	3.122	*eccellenza*
Psych Verbs	635	*amare*
LG Verbs	879	*prendersla*
Bad words	189	*leccaculo*
Tot	15.179	-

Table 1: Composition of SentIta

Compound adverbs [3] (Elia, 1990), idioms[4] (Vietri, 1990; Vietri, 2011), psych verbs[5] and other LG verbs[6] (Elia et al., 1981; Elia, 1984) have been manually weighted starting from the Italian LG tables, in order to maintain the syntactic, semantic and transformational properties connected to each one of them.

However, to manually draw up a complete dictionary is a strong time-consuming activity, therefore we automatically derived the remaining parts of the lexicon exploiting a set of morphological rules.

To be more precise, the Prior Polarity has been assigned to adverbs ending in *-mente*, "-ly" (Ricca, 2004) on the base of the morpho-phonological relations with their known adjectival bases (e.g. *allegro* and *allegramente*), with 0.99 Precision and 0.88 Recall. In a similar way, with a Precision of 0.93 an a Recall of 0.72, we made a set of quality nouns inherit the semantic tags of the qualifier adjectives with which they were related (e.g. *eccellente* and *eccellenza*) using a set of 21 suffixes for the noun formation from qualifier adjectives [7]. In

[1]The Nooj software and the Italian module of Nooj are available for download at `www.nooj-association.org`.

[2]LG tables anavailable for consultation at `dsc.unisa.it/composti/tavole/combo/tavole.asp`.

[3]LG classes of compound adverbs: PC, PDETC, PAC, PCA, PCDC, PCPC, PCONG, CAC, CPC, PV, PF, PCPN, PEW. This list comprises 70 discourse operators, able to summarize (e.g. *in parole povere*), invert (e.g. *nonostante ciò*), confirm (e.g. *in altri termini*), compare (e.g. *allo stesso modo*) and negate (e.g. *neanche per sogno*) the opinion expressed in the previous sentences of the text.

[4]LG classes of frozen sentences with *essere* as support verb: CEAC, EAA, EAPC, ECA, PECO.

[5]LG classes of psych verb: 41, 42, 43, 43B.

[6]Other LG classes of verbs: 2, 2A, 4, 9, 10, 11, 18, 20I, 20NR, 20UM, 21, 21A, 22, 23R, 24, 27, 28ST, 44, 44B, 45, 45B, 47, 47B, 49, 50, 51, 53, 56.

[7]The suffixes for the quality noun formation (Rainer, 2004) are -edine, *-edine*, *-età*, *-izie*, *-ela*, *-udine*, *-ore*, *-(z)ione*, *-anza*, *-itudine*, *-ura*, *-mento*, *-izia*, *-enza*, *-eria*, *-ietà*,

the end, we also collected a list of 37 prefixes (Iacobini, 2004), that, as morphological Contextual Valence Shifters (CVS), are able to negate (e.g. *anti-*, *contra-*, *non-*, ect) or to intensify/downtone (e.g. *arci-*, *semi-*, ect) the orientation of the words in which they appear. They directly work on opinionated documents, in order to make the machine understand the actual orientation of the words occurring in real texts, also when their morphological context shifts the polarity of the words listed in the dictionaries.

4 Contextual Polarity: Rules for the Sentence-level Sentiment Analysis

Grounding a sentiment analysis tool only on the simple lexical valence of negative or positive words can become misleading in all the cases in which the sentence or the discourse context shifts the valence of individual terms.

In order to annotate real texts with the proper semantic orientation, we took advantage of the finite-state technology, thanks to which we could systematically recall and modify the prior polarities expressed in the dictionaries on the base of the syntactic contexts in which the words occur.

Among the most used CVS, we took into account linguistic phenomena like Intensification, Negation, Modality and Comparison.

In this work the sentence annotations is not performed through mathematical computations; instead, it is grounded on the semantic labels attributed to each one of the embedded nodes of the finite-state automata (FSA), which contain, just as boxes, all the syntactic structures that should obtain the same score. This choice is due to the fact that the effects of the interactions between word polarities and CVS on the semantics of phrases and sentences are various, but regular. As an example, negation[8] does not always switch the polarity of the modified words: as it can be observed in Table 2, there are cases in which it is only shifted.

Regarding intensification, we considered the cases in which intensifiers and downtoners modify the opinionated words; superlative, and repetition of positive or negative words. Basically, adjectives can only modify the intensity of nouns, while adverbs intensify or attenuate adjectives, verbs and other adverbs.

Because intensification and negation can also appear together in the same sentence, we took into account this eventuality by weighting firstly the intensification and then the negation.

Rules	Example	Score
Adj^{+2}	*bello*	+2
Adj^{+2}+Adj^{+2}	*bello, bello*	+3
Adj^{+2}+Adv^{+}	*molto bello*	+3
Adj^{+2}+Sup	*bellissimo*	+3
Adj^{+2}+Adv^{-}	*poco bello*	-2
Adj^{+2}+Neg	*non bello*	-2
Adj^{+2}+Neg + Adv^{+}	*non molto bello*	-1
Adj^{+2}+Neg+Sup	*non bellissimo*	-1
Adj^{+2}+Neg+Adv^{-}	*non poco bello*	+1

Table 2: An example of intensification and negation rules on a positive adjective

As concerns modality (Benamara et al., 2012), we contemplated modal verbs (e.g. *dovere*, *potere*), conditional tenses (Narayanan et al., 2009) and doubt or necessity adverbs (e.g. *più o meno*, *di sicuro*).

In order to detect similarities and differences between two or more objects, we listed a set of comparative opinion indicators; namely, comparative frozen sentences of the type *N0 Agg come C1* (Vietri, 1990); some simple comparative sentences that involve the expressions *meglio di*, *inferiore a*; and the comparative superlative (e.g. *il più*, *il peggiore*).

5 Doxa: a Document-level Opinionated teXt Analyzer

Using the command-line program noojapply.exe, we built a Java prototype by which users can automatically apply the resources formalized in Nooj to every kind of text.

Doxa, in its standard version, is a Document-level Opinionated teXt Analyzer that evaluates customer reviews, by summing up the values corresponding to every sentiment expression and, then, normalizes the result for the total number of

-aggine, *-ia*, *-ità*, *-ezza*, *-igia*, *-(z)a*. They generally make the new words simply inherit the orientation of the derived adjectives. Exceptions are *-edine* and *-eria* that almost always shift the polarity of the quality nouns into the weakly negative one (-1), e.g. *faciloneria* "slapdash attitude". Also the suffix *-mento* differs from the others, in so far it belongs to the derivational phenomenon of the deverbal nouns of action (Gaeta, 2004). It has been possible to use it in this work noun derivation by using the past participles of the verbs listed in the adjective dictionary of sentiment.

[8] As negation indicators we took into account negative operators (e.g. *non*, "not", *mica*), negative quantifiers (e.g. *nessuno*, "nobody" *niente*, *nulla*, "nothing") and lexical negation (e.g. *senza*, "without", *mancanza di*, *assenza di*, *carenza di*, "lack of") (Benamara et al., 2012).

sentiment expressions contained in the text. In the end, it compares the resulting values with the stars that the opinion holders gave to their reviews and provides statistics about the opinions expressed in each case (Maisto and Pelosi, 2014a).

The test dataset used to evaluate the performances of Doxa contains Italian opinionated texts in the form of users reviews and comments from e-commerce and opinion websites. It lists 600 texts units (300 positive and 300 negative for each product class) and refers to experience and search goods, for all of which different websites have been exploited[9].

It the sentence-level sentiment analysis Doxa achieved an average Precision of 0.75 and a Recall of 0.73, and in the document-level classification an average Precision of 0.74 and a Recall of 0.97.

Although the document-level analysis provides important information regarding the consumers needs and expectations, for companies it is crucial to discern the aspects of the products that must be improved, or whether the opinions extracted online by the sentiment analysis applications are relevant to the products or not. That is why we designed new sentiment analysis modules for Doxa: a sentiment role labeling module (Pelosi et al., 2015 in press) and a feature-based sentiment analyzer (Maisto and Pelosi, 2014b).

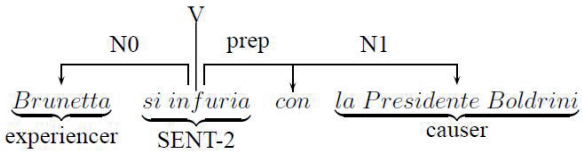


Figure 1: Sentiment Role Labeling

In the first task the achieved F-score was 0.71 in a Twitter dataset and 0.76 in a free web news headings dataset provided by DataMediaHub. As shown in Figure 1, the tool is based on the matching between the definitional syntactic structures, attributed to the each one of the 28 LG class of Italian verbs, and the semantic information attached in the sentiment database to every lexical entry. The semantic roles evoked by this selection of verbs pertain to three frames: *Sentiment (experiencer, causer), Opinion (holder, target) and Physical act (agent, patient)*.

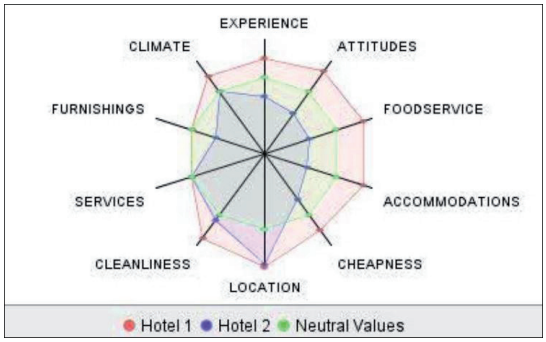


Figure 2: Feature-based module of Doxa

In the feature based sentiment analysis (Maisto and Pelosi, 2014b), which allows the comparison between more than one object on the base of different kind of aspects that characterize them (Figure 2), we achieved 0.81 F-score on a corpus of hotels reviews.

Because sometimes a lexicon is not enough to correctly extract and classify the product features, domain-specific FSA have been formalized in order to make the analyses adequate to the corpus (e.g. *L'hotel vicinissimo alla metro.*, BENEFIT TYPE="LOCATION" SCORE="3").

6 Conclusion

In the present paper we underlined that the social and economic impact of the online customer opinions and the huge volume of raw data available on the web, concerning users point of views, offer new opportunities both to marketers and researchers.

Indeed, sentiment analysis applications, able to go deep in the semantics of sentences and texts, can play a crucial role in tasks like web reputation monitoring, in social network analysis, in viral tracking campaigns, etc...

Therefore, we presented SentIta, a semi-automatic Italian Lexicon for Sentiment Analysis, and Doxa, a Document-level Opinionated teXt Analyzer that exploits finite-state technologies to go through the subjective dimension of user generated contents.

[9]Books: www.amazon.it, www.qlibri.it, Movies: www.mymovies.it, www.cinemalia.it, www.filmtv.it, www.filmscoop.it, Hotels: www.tripadvisor.it, www.expedia.it, www.venere.com, it.hotels.com, www.booking.com and Videogames: www.amazon.it, Cars: www.ciao.it, Smartphones: www.tecnozoom.it, www.ciao.it, www.amazon.it, www.alatest.it.

References

Matteo Baldoni, Cristina Baroglio, Viviana Patti, and Paolo Rena. 2012. From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, 6(1):41–54.

Marco Baroni and Stefano Vegnaduzzo. 2004. Identifying subjective adjectives through web-based mutual information. 4:17–24.

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.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*.

Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2012. How do negation and modality impact on opinions? pages 10–18.

Kenneth Bloom. 2011. Sentiment analysis based on appraisal theory and functional local grammars.

Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. pages 793–801.

Wenjing Duan, Bin Gu, and Andrew B Whinston. 2008. The dynamics of online word-of-mouth and product salesan empirical investigation of the movie industry. *Journal of retailing*, 84(2):233–242.

Annibale Elia, Maurizio Martinelli, and Emilio D'Agostino. 1981. *Lessico e Strutture sintattiche. Introduzione alla sintassi del verbo italiano.* Napoli: Liguori.

Annibale Elia. 1984. Le verbe italien. *Les complétives dans les phrases à un complément*.

Annibale Elia. 1990. *Chiaro e tondo: Lessico-Grammatica degli avverbi composti in italiano*. Segno Associati.

Andrea Esuli and Fabrizio Sebastiani. 2006. Determining term subjectivity and term orientation for opinion mining. 6:2006.

Livio Gaeta. 2004. Nomi d'azione. *La formazione d elle parole in italiano. Tübingen: Max Niemeyer Verlag*, pages 314–51.

Ahmed Hassan and Dragomir Radev. 2010. Identifying text polarity using random walks. pages 395–403.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics.

Irazú Hernandez-Farias, 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 EVALITA2014*, pages 75–81.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760.

Claudio Iacobini. 2004. Prefissazione. *La formazione delle parole in italiano. Tübingen: Max Niemeyer Verlag*, pages 97–161.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. page 1367.

Lun-Wei Ku, Ting-Hao Huang, and Hsin-Hsi Chen. 2009. Using morphological and syntactic structures for chinese opinion analysis. pages 1260–1269.

Alessandro Maisto and Serena Pelosi. 2014a. Feature-based customer review summarization. In *On the Move to Meaningful Internet Systems: OTM 2014 Workshops*, pages 299–308. Springer.

Alessandro Maisto and Serena Pelosi. 2014b. 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.

Isa Maks and Piek Vossen. 2011. Different approaches to automatic polarity annotation at synset level. pages 62–69.

Karo Moilanen and Stephen Pulman. 2008. The good, the bad, and the unknown: morphosyllabic sentiment tagging of unseen words. pages 109–112.

Makoto Nakayama, Norma Sutcliffe, and Yun Wan. 2010. Has the web transformed experience goods into search goods? *Electronic Markets*, 20(3-4):251–262.

Ramanathan Narayanan, Bing Liu, and Alok Choudhary. 2009. Sentiment analysis of conditional sentences. pages 180–189.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *ICWSM*.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2011. Sentiful: A lexicon for sentiment analysis. *Affective Computing, IEEE Transactions on*, 2(1):22–36.

Alena Neviarouskaya. 2010. Compositional approach for automatic recognition of fine-grained affect, judgment, and appreciation in text.

Charles E Osgood. 1952. The nature and measurement of meaning. *Psychological bulletin*, 49(3):197.

Serena Pelosi, Annibale Elia, and Alessandro Maisto. 2015 (in press). Towards a lexicon-grammar based framework for nlp: an opinion mining application. In *Recent Advances in Natural Language Processing 2015 - RANLP 2015 Proceedins*.

Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. pages 1–10.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. 9:1199–1204.

Franz Rainer. 2004. Derivazione nominale deaggettivale. *La formazione delle parole in italiano*, pages 293–314.

Davide Ricca. 2004. Derivazione avverbiale. *La formazione delle parole in italiano*, pages 472–489.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25–32. Association for Computational Linguistics.

Max Silberztein. 2003. Nooj manual. *Available for download at:* `www.nooj4nlp.net`.

Josef Steinberger, Mohamed Ebrahim, Maud Ehrmann, Ali Hurriyetoglu, Mijail Kabadjov, Polina Lenkova, Ralf Steinberger, Hristo Tanev, Silvia Vázquez, and Vanni Zavarella. 2012. Creating sentiment dictionaries via triangulation. *Decision Support Systems*, 53(4):689–694.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.

Maite Taboada, Caroline Anthony, and Kimberly Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genova, Italy*, pages 427–432.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. pages 417–424.

MJM Vermeij. 2005. The orientation of user opinions through adverbs, verbs and nouns. In *3rd Twente Student Conference on IT, Enschede June*. Citeseer.

Simonetta Vietri. 1990. On some comparative frozen sentences in italian. *Lingvisticæ Investigationes*, 14(1):149–174.

Simonetta Vietri. 2011. On a class of italian frozen sentences. *Lingvisticæ Investigationes*, 34(2):228–267.

Simona Vietri. 2014. The italian module for nooj. In *In Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it 2014*. Pisa University Press.

Xin Wang, Yanqing Zhao, and Guohong Fu. 2011. A morpheme-based method to chinese sentence-level sentiment classification. *Int. J. of Asian Lang. Proc.*, 21(3):95–106.

Aleksander Wawer. 2012. Extracting emotive patterns for languages with rich morphology. *International Journal of Computational Linguistics and Applications*, 3(1):11–24.

Qiang Ye, Rob Law, Bin Gu, and Wei Chen. 2011. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, 27(2):634–639.

Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. 2010. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 1462–1470. Association for Computational Linguistics.

Feng Zhu and Xiaoquan Zhang. 2006. The influence of online consumer reviews on the demand for experience goods: The case of video games. *ICIS 2006 Proceedings*, page 25.

Le scritture brevi dello storytelling: analisi di case studies di successo

Maria Laura Pierucci
Università di Macerata
marialaurapierucci@gmail.com

Abstract

English This paper presents an analysis of successful storytelling case studies. Their strategies and techniques of branding will be analyzed with an interdisciplinary approach, grounded on pragmatics of communication, within the conceptual framework of 'scritture brevi' as set in www.scritturebrevi.it.

Italiano *Il contributo presenta l'analisi di case study di storytelling di successo. La prospettiva è interdisciplinare: muovendo da premesse di pragmatica della comunicazione, si indagano strategie e tecniche di branding attraverso l'uso della categoria concettuale delle 'scritture brevi' come in www.scritturebrevi.it.*

1 Introduction

Una tradizione ormai consolidata di studi semiotici, di linguistica cognitiva e di psicologia indica la mente umana come 'narrante'. Così Roland Barthes (1969: 7): "[…] il racconto è presente in tutti i tempi, in tutti i luoghi, in tutte le società; il racconto comincia con la storia stessa dell'umanità; non esiste, non è mai esistito in alcun luogo un popolo senza racconti […] il racconto è là come la vita".
La riflessione sulle strutture della narratività iniziata da Propp e dai formalisti russi negli anni '20 del Novecento venne ripresa nella seconda metà del secolo scorso da strutturalisti come Barthes, Todorov e Genette, fra gli altri, che diedero il loro contributo all'elaborazione di una teoria narrativa a partire proprio dal presupposto che il raccontare, e il raccontarsi, siano fenomeni costanti nella storia dell'uomo, comportamenti all'essere umano connaturati.

2 Storytelling fra passato e presente

È con l'era digitale che lo storytelling si lega nella prassi al marketing non convenzionale, in una complementarietà disciplinare che contraddistingue sempre più questo tipo di studi. Internet e la Rete, infatti, si sono subito dimostrati strumenti di straordinaria efficacia per coinvolgere il consumatore in una comunicazione interattiva e personalizzata (Collesei/Casarin/Vescovi 2001).
"Contributi di recente formalizzazione", spiegano Russo Spena/Colurcio/Melia (2013: 99), "evidenziano la portata innovativa di tale strumento [lo storytelling, NdA] soprattutto nella declinazione che individua in Internet l'infrastruttura portante e nelle virtual communities i driver essenziali per potenziare il contributo della dinamica narrativa al consolidamento e allo sviluppo delle relazioni sociali ed emozionali finalizzate alla *brand loyalty*".
I case studies, presi in esame a seguire, si contraddistinguono proprio per l'impiego di strategie di comunicazione che, facendo leva sullo storytelling e su un impiego efficace delle 'sue' scritture brevi, hanno potenziato il proprio brand.

2.1 Lo storytelling e le sue 'scritture brevi'

La prospettiva scientifica con la quale usiamo l'etichetta 'scritture brevi' è quella definita da Chiusaroli (2012a, 2012b, 2014a, 2014b), che la propone "come categoria concettuale e metalinguistica per la classificazione di forme grafiche come abbreviazioni, acronimi, segni, icone, indici e simboli, elementi figurativi, espressioni testuali e codici visivi per i quali risulti dirimente il principio della 'brevità' connesso al criterio dell''economia'. In particolare sono comprese nella categoria Scritture Brevi tutte le manifestazioni grafiche che, nella dimensione sintagmatica, si sottraggono al principio della linearità del significante, alterano le regole morfotattiche convenzionali della lingua scritta, e intervengono nella costruzione del messaggio nei termini di 'riduzione, contenimento, sintesi' indotti dai supporti e dai contesti. La categoria ha applicazione nella sincronia e nella diacronia linguistica, nei sistemi standard e non standard, negli ambiti generali e specialistici".
L'applicazione di tale etichetta si intende non solo nel senso stretto, come sopra specificato, ma anche come macro-contenitore, laboratorio e os-

servatorio scientifico dei fenomeni della lingua del web, in considerazione del fatto che le campagne di strategia del brand esaminate sono state pensate per l'ecosistema digitale, nel senso di community di soggetti che interagiscono e si scambiano informazioni, accrescendo conoscenze e contatti con lo scopo di migliorare la loro esistenza e soddisfare le loro necessità.
A partire dalle premesse teoriche della linguistica pragmatica, con particolare attenzione agli studi di analisi conversazione (da Austin e Searle e la teoria degli atti linguistici, passando per Grice, Halliday fino a Berretta e Bazzanella) e della pragmatica della comunicazione (Watzlawick/Helmick Beavin/Jackson 1971), i confini disciplinari si dilatano fino a ricomprendere le riflessioni di Lambert (2006), Bran (2010), Malita/Martin (2010) e, in particolare, Fog/Budtz/Yakaboylu (2005) e Brown/Groh/Prusak/Denning (2005) che sottolineano come la narrazione del brand si costruisca attraverso il dialogo fra gli interlocutori, intercettati e coinvolti grazie a precise strategie di engagement, nel contesto digitale.
Quella *social* è, infatti, una dimensione in cui, dopo una fase fondamentale di listening del proprio target, lo storytelling viene ideato e realizzato per poi essere alimentato di scrittura e riscrittura.
Anche quando si tratti di visual storytelling, la forma scrittoria, creata e formulata dal copywriter, mantiene la sua funzione di strumento di condensazione e formalizzazione del messaggio.
Si badi bene che condensazione non vuol dire necessariamente abbreviazione. Si tratta, infatti, di forme 'brevi' di scrittura nelle quali la brevità, come dimostrato da Chiusaroli (2012b), non inficia i livelli di informatività.
A volte, come nel primo caso di studio che presentiamo, può anzi essere 'forma' e 'sostanza' allo stesso tempo, 'il mezzo e il fine' per dirla ancora con Chiusaroli (a questo proposito, si veda in particolare il blog www.scritturebrevi.it).

2.2 Storytelling nell'era digitale: case studies e la funzione delle 'scritture brevi'

Prendiamo in esame la campagna di comunicazione promossa nel 2014 dalla Visa, la multinazionale finanziaria, strutturata mediante visual storytelling e il cui claim (il *core message*, per dirla in termini tecnici) era "Let's go do something" viralizzato su canali social (Facebook, Instagram, Twitter, Youtube, Vine, Google+) tramite l'hashtag #GoInSix in cui il numero 6 (per una corretta formazione dell'hashtag che non prevede l'inserimento di cifre o simboli, pena la decadenza della sua funzionalità, si veda Chiusaroli 2014b e Pierucci 2015) indica in secondi la durata dei teaser pubblicitari, la quantità di foto per album e di parole per post.
La brevità come esercizio della 'lingua' del web: un tweet ha la lunghezza massima di 140 caratteri; la cosiddetta generazione Millennium ha imparato a caricare su Vine filmati di 6 secondi; è di 8 parole la lunghezza media dei commenti (in lingua inglese) postati dai consumatori.
Per tornare a #go*insix*, con testi (headline abbinate a video o foto) come 'Music under the moon sounds sweeter', 'Sorry Nonna, your secret is out', 'Stand on the shoulders of ocean', l'invito di Visa era a contribuire con "stories in six": la sfida a rispettare il limite di sei, che fossero secondi, foto o parole, ha fatto guadagnare all'azienda 330 milioni di 'earned impressions' (contatti spontanei generati dal passaparola) in un anno e con 36.838 interazioni (like, commenti e condivisioni per post) quella di Visa è diventata la community con maggior engagement (coinvolgimento) nel settore dei servizi finanziari, risultando seconda in assoluto nella più generale categoria del 'Lifestyle'.
'Nati per proteggere' è invece il claim dello storytelling promosso da Axa Assicurazioni sia nel 2014 che nel 2015; 'Raccontaci la tua storia di protezione' è la call to action cui hanno risposto in 351. Migliaia le visualizzazioni.
Secondo quanto indicato in Fog et al. (2005), gli elementi dello storytelling sono il messaggio, cioè il tema principale attorno al quale si costruisce la storia; il conflitto, che è la forza trainante di una buona storia; i personaggi, ciascuno con un ruolo ben definito e nel quale il destinatario possa immedesimarsi; e infine, la linea narrativa, vale a dire la trama, il plot, che può avere andamento differente a seconda che il racconto si ascriva ad un genere, come quello tragico ad esempio, oppure ad un altro, come nel caso di una storia d'amore. Come sottolineato da Lambert (2006) e Bran (2010), non devono mancare le leve del coinvolgimento emotivo. Nel caso di digital storytelling, quando il racconto si struttura per immagini o filmati, la brevità - come abbiamo visto nel primo caso studio - rimane un elemento fondamentale che viene scandito dal pacing, ossia l'uso del ritmo, della musica, della voce.
Il progetto è attivo sui social veicolato dall'hashtag #natiperproteggere ed è legato ad un concorso: come si legge nel regolamento, fra i

criteri 'premianti' c'è la 'funzionalità in termini di impatto emotivo, condivisione e notiziabilità'. In sostanza, i presupposti per un buon digital storytelling.
Case study di visual storytelling da oltre mezzo milione di visualizzazioni su Youtube è quello realizzato nel 2010 da Ogilvy & Mather per la catena alberghiera Shangri-La: un filmato della durata di 3 minuti, completamente decontestualizzato rispetto al brand che lo ha commissionato, in cui i passaggi della narrazione sono sottolineati solo dalla colonna sonora, e nel quale la headline (un altro esempio di 'scritture brevi' di 54 caratteri) compare in chiusura anzi che in esergo, "To embrace a stranger as one's own. It's in our nature". In termini di pragmatica linguistica il focus della struttura informativa è dislocato a sinistra: una forma marcata dal punto di vista dell'ordine delle parole la cui efficacia è massimizzata dalla concisione.
Di matrice culturale, infine, è il progetto di branding attraverso storytelling promosso nel 2015 dal Macerata Opera Festival in collaborazione con il blog di Francesca Chiusaroli e Fabio Massimo Zanzotto, scritturebrevi.it.
L'ente lirico marchigiano ha impiegato le tecniche di narrazione in ambito social, specificamente sulla piattaforma da 140 caratteri, raccontando le opere in cartellone (Rigoletto, Cavalleria Rusticana e Pagliacci, Bohème) con brani scelti dai libretti, abbinati a foto dal vivo delle Prime durante il loro svolgimento all'Arena Sferisterio di Macerata. Uno storytelling lanciato con hashtag 'breve' #nutrimilive, forma contratta di due hashtag, quello ufficiale della 51° stagione lirica #nutrimilanima e di #live inteso come 'spettacolo dal vivo', proposto sotto forma di live tweeting che ha creato, in seno alla community di Scritture Brevi, tre meta-testi poi cristallizzati grazie allo strumento dello Storify.
Insieme agli hashtag #Rigoletto, #CavalleriaRusticana e #Bohème, #nutrimilive ha totalizzato 924 tweet (fonte: Topsy) in corrispondenza delle tre serate di live tweeting, portando l'opera lirica sui social con un racconto fatto di immagini e 'scritture brevi'.

3 Conclusion

In questo lavoro abbiamo analizzato alcuni case study di storytelling e le loro 'scritture brevi' (claim, headline, hashtag) attraverso i quali quei racconti sono stati pensati e diffusi nel web per essere intercettati nel mare magnum dei contenuti digitali.
Abbiamo visto come lo strumento principe del marketing narrativo trovi la sua massima espressione in una coniugazione sapiente di testo (breve) ed immagini, riuscendo così a tradursi in un efficace strumento di branding per aziende ma anche per enti culturali.
Le regole da rispettare sono le stesse oggi come nel passato più remoto: il racconto parla dell'uomo e all'uomo e, in virtù del fatto che i mercati sono conversazioni (Cleutrain Manifesto 1999), vive una nuova stagione di successo grazie all'ecosistema digitale.

Reference

John L. Austin. 1962. *How to do things with words*. Oxford University Press, Oxford.

Roland Barthes. 1969. *Introduzione*. AAVV. *L'analisi del racconto*. Bompiani, Milano: 7-46.

Carla Bazzanella. 1994. *Le facce del parlare. Un approccio pragmatica all'italiano parlato*. La Nuova Italia, Firenza-Roma.

Carla Bazzanella. 2005. *Linguistica e pragmatica del linguaggio*. Laterza, Roma-Bari.

Émile Benveniste. 1966. *Problèmes de linguistique génerale*. Gallimard, Paris.

Ramona Bran. 2010. *Message in a bottle. Telling stories in a digital world*. In *Procedia Social and Behavioral Sciences*, 2: 1790-1793.

John Seely Brown, Katalina Groh, Larry Prusak and Steve Denning. 2005. *Storytelling in organizations. Why storytelling is transforming 21th century organizations and management*. Elsevier Butterworth Heinemann, Burlington.

Francesca Chiusaroli. 2012a. *Scritture brevi oggi: tra convenzione e sistema*. In Francesca Chiusaroli and Fabio Massimo Zanzotto (eds.). *Scritture brevi di oggi*, Quaderni di Linguistica Zero, 1. Università degli studi di Napoli "L'Orientale", Napoli: 4-44.

Francesca Chiusaroli and Fabio Massimo Zanzotto. 2012b. *Informatività e scritture brevi del web*. In Francesca Chiusaroli and Fabio Massimo Zanzotto (eds.). Scritture brevi nelle lingue moderne. Quaderni di Linguistica Zero, 2. Università degli studi di Napoli "L'Orientale", Napoli: 3-20.

Francesca Chiusaroli 2014a, *Sintassi e semantica dell'hashtag: studio preliminare di una forma di Scritture Brevi*, in R. Basili, A. Lenci, B. Magnini (eds.), *The First Italian Conference on Computational Linguistics, CLiC-it 2014 – Proceedings, 9-10 December 2014*. Pisa University Press, Pisa, vol. I: 117-121.

Francesca Chiusaroli. 2014b. *Scritture Brevi di Twitter: note di grammatica e di terminologia*. In Vincenzo Orioles, Raffaella Bombi and Marika Brazzo (eds.). *Metalinguaggio. Storia e statuto dei costrutti della linguistica*. Il Calamo, Roma: 435-448.

Umberto Collesei, Francesco Casarin and Tiziano Vescovi. 2001. *Internet e i cambiamenti nei comportamenti di acquisto del consumatore*. In *Micro & Macro marketing*, 1: 33-50.

David Crystal. 2004. Language and the Internet. Cambridge University Press, Cambridge.

Klaus Fog, Christian Budtz and Baris Yakaboylu. 2005. *Storytelling. Branding in practise.* Springer, Berlin-Heidelberg.

Joe Lambert. 2006. *Digital storytelling cookbook*. Digital Diner Press, Berkeley.

Gerard Genette. 1966. *Figures*. Seuil, Paris.

Gottschall, Jonathan. 2012. *The storytelling animal. How stories make us human*. Houghton Mifflin Harcourt, Boston.

Algirdas J. Greimas. 1966. *Sémantique structurale*. Larousse, Paris.

Herbert Paul Grice. 1989. *Studies in the way of words*. Harvard University Press, Cambridge (MA).

Michael A. K. Hallyday. 1985. *Spoken and written language*. Oxford University Press, Oxford.

Laura Malita and Catalin Martin. 2010. *Digital storytelling as web passport to success in the 21th century*. In *Procedia social and behavioral sciences*, 2: 3060-3064.

Maria Laura Pierucci. In press. *Categorie per il dizionario di Scritture Brevi: l'hashtag*.

Vladimir J. Propp. [1928] 2000. *Morfologia della fiaba: con un intervento di Claude Lévi-Strauss e una replica dell'autore*. Einaudi, Torino.

John R. Searle. 1969. *Speech Acts.* Cambridge University Press, Cambridge.

Tiziana Russo Spena, Maria Colurcio and Monia Melia. 2013. *Storytelling e web communication*. In Mercati e competitività. Franco Angeli, Milano: 97-117.

Tezvetan Todorov. 1965. *Théorie de la littérature*. Seuil, Paris.

Paul Watzlawick, Helmick Beavin, Janet Jackson and Don D. 1971. *Pragmatica della comunicazione umana. Studio dei modelli interattivi delle patologie e dei paradossi*. Astrolabio, Roma.

Tracking the Evolution of Written Language Competence: an NLP–based Approach

Stefan Richter•, Andrea Cimino◇, Felice Dell'Orletta◇, Giulia Venturi◇
•University of Leipzig (Germany)
hewuri@gmail.com
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it

Abstract

English. In this paper, we present an NLP-based innovative approach for tracking the evolution of written language competence relying on different sets of linguistic features that predict text quality. This approach was tested on a corpus essays written by Italian L1 learners of the first and second year of the lower secondary school.

Italiano. *In questo articolo, presentiamo un metodo innovativo per monitorare l'evoluzione delle competenze di scrittura basato su tecnologie del linguaggio che sfruttano caratteristiche linguistiche predittive della qualità del testo. Questo approccio è stato testato su un corpus di produzioni scritte di apprendenti l'italiano L1 del primo e secondo anno della scuola secondaria di primo grado.*

1 Introduction and Background

Using automatic techniques to trace the learning progress of students starting from their written productions is receiving growing attention in many different research fields and for different purposes. Two different perspectives are taken into account: i.e. the analysis of the *form* and of the *content* of texts. The first scenario is mainly addressed within the writing research community where the learning progress is framed as an analysis aimed at detecting linguistic predictors of written quality across grade levels. Using Natural Language Processing (NLP) tools, different set of features (e.g. grammar features, errors, measures of lexical complexity) are automatically extracted from corpora of student essays to investigate how they relate to writing quality (Deane and Quinlan, 2010) or to other literacy processes such as reading (Deane, 2014). Human ratings of essay writing quality are also used to develop Automatic Essay Scoring systems mostly of L2 essays (Attali and Burstein, 2006). For what concerns the analysis of *content* of texts, traditional Knowledge Tracing systems are based on a framework for modeling the process of student learning while completing a sequence of assignments (Corbett and Anderson, 1994). These systems rely on a correctness value of each assignment given by a teacher. More recently, the Knowledge Tracing framework started to be explored by the Machine Learning (ML) community[1] in Adaptive E–learning scenarios. Different ML approaches have been devised to build statistical models of student knowledge over time in order to predict how students will perform on future interactions and to provide personalized feedback on learning (Piech et al., 2015; Ekanadham and Karklin, 2015).

Both the evaluation of *form* and *content* of a text share a common starting point: they imply a human 'commitment'. In the first case, it is assumed that the analyzed essays are manually scored according to the writing quality level, in the second case, the statistical models are trained on student exercises indicating whether or not the exercise was answered correctly.

In this paper, we present an innovative approach for tracking the evolution of written language competence using NLP techniques and relying on not-scored essays. Our approach focuses on the analysis of *form* but we combined for the first time the methods developed to tackle the form and content evaluation. Namely, we automatically extracted from written essays linguistic predictors of text quality that we used as features of a machine learning classifier to trace student developmental growth over the time. We tested this method on a corpus of written essays of Italian L1 learners collected in the first and second year of the lower secondary school. The use of not-scored essays is

[1] http://dsp.rice.edu/ML4Ed_ICML2015

one of the main novelty making our approach particularly suited for less resourced languages such as the Italian language, as far as corpora of L1 students are concerned.

2 Our Approach

Our approach of tracking the evolution of written language competence of L1 learners is based on the assumption that given a set of chronologically ordered essays written by the same student a document d_j should show a higher written quality level with respect to the ones written previously. Following this assumption, we consider the problem of tracking the evolution of a student as a classification task. Given two essays d_i and d_j written by the same student, we want to classify whether $t(d_j) > t(d_i)$, where $t(d_i)$ is the time in which the document d_i was written.

For this purpose, we built a classifier operating on morpho-syntactically tagged and dependency parsed essays which assigns to each pair of documents (d_i, d_j) a score expressing its probability of belonging to a given class: 1 if $t(d_j) > t(d_i)$, 0 otherwise. Given a training corpus, the classifier builds all possible pairs (d_i, d_j) of documents written by the same student. For each pair of documents (d_i, d_j), two feature vectors (V_{d_i}, V_{d_j}) are extracted. Exploiting these two vectors, $V_{d_i,d_j} = V_{d_i} - V_{d_j}$ is computed. Since many machine learning algorithms assume that input data values are in a standard range, we finally calculated V'_{d_i,d_j} obtained by scaling each component in the range $[0, 1]$. The classifier was trained and tested on the corpus described in section 3, it uses the features described in section 4 and linear Support Vector Machines (SVM) using LIBSVM (Chang and Lin, 2001) as machine learning algorithm.

3 Corpus

We relied on CItA (*Corpus Italiano di Apprendenti L1*), the first corpus of essays written by Italian L1 learners in the first and second year of lower secondary school which has been manually annotated with grammatical, orthographic and lexical errors (Barbagli et al., 2015). The corpus contains 1,352 texts written by 156 students and collected in 7 different lower secondary schools in Rome: 3 schools (77 students) are located the historical center and 4 schools (79 students) in suburbs. CItA contains two different types of essays differing with respect to the prompt, i.e the prompts assigned individually by each teacher during each school year and a prompt common to all schools that was assigned at the end of the first and second year. It is also accompanied by a questionnaire containing a set of questions referring to the student background (e.g. questions about the student family, about the native language spoken at home, etc.). This makes possible to investigate whether and to which extent some of the student background information are related to the observed language competence evolution. The main peculiarity of the corpus is its diachronic nature. Even though the contained essays were not manually scored, the covered temporal span makes CItA particularly suitable for tracking the evolution of the written language competence over the time.

4 Features

Our approach relies on multi–level linguistic features, both automatically extracted and manually annotated in CItA. A first set of features was extracted from the corpus morpho-syntactically tagged by the POS tagger described in (Dell'Orletta, 2009) and dependency-parsed by the DeSR parser using Multi-Layer Perceptron (Attardi et al., 2009). They range across different linguistic description levels and they qualify lexical and grammatical characteristics of a text. These features are typically used in studies focusing on the "form" of a text, e.g. on issues of genre, style, authorship or readability (see e.g. (Biber and Conrad, 2009; Collins-Thompson, 2014; Cimino et al., 2013; Dell'Orletta et al., 2014)). The second set of features refers to the errors manually annotated. Also these features range across different linguistic description levels.

Raw and Lexical Text Features *Sentence Length* and *Token Length*: calculated as the average number of words and characters. *Basic Italian Vocabulary rate features*: these features refer to the internal composition of the vocabulary of the text. To this end, we took as a reference resource the Basic Italian Vocabulary by De Mauro (1999), including a list of 7000 words highly familiar to native speakers of Italian. *Words Frequency class*: this feature refers to the average class frequency of all lemmas in the document. The class frequency for each lemma was computed exploiting the *2010-news-1M* corpus (Quasthoff et al., 2006), using the following function: $C_{cw} = \lfloor \log_2 \frac{freq(MFL)}{freq(CL)} \rfloor$, where MFL is the

most frequent lemma in the corpus and CL is the considered lemma. *Type/Token Ratio*: this feature refers to the ratio between the number of lexical types and the number of tokens.

Morpho-syntactic Features *Language Model probability of Part-Of-Speech unigrams*: this feature refers to the distribution of unigram Part-of-Speech. *Lexical density*: this feature refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text. *Verbal mood*: this feature refers to the distribution of verbs according to their mood.

Syntactic Features *Unconditional probability of dependency types*: this feature refers to the distribution of dependency relations. *Parse tree depth features*: this set of features captures different aspects of the parse tree depth and includes the following measures: a) the depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to some leaf; b) the average depth of embedded complement chains governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers; c) the probability distribution of embedded complement chains by depth. *Verbal predicates features*: this set of features ranges from the number of verbal roots with respect to number of all sentence roots occurring in a text to their arity. The arity of verbal predicates is calculated as the number of instantiated dependency links sharing the same verbal head. *Subordination features*: these features include a) the distribution of subordinate vs main clauses, b) their relative ordering with respect to the main clause, c) the average depth of chains of embedded subordinate clauses and d) the probability distribution of embedded subordinate clauses chains by depth. *Length of dependency links*: the length is measured in terms of the words occurring between the syntactic head and the dependent.

Annotated Error Features These features refer to the distribution of different kinds of errors manually annotated in CItA: a) *grammatical errors*, e.g. wrong use of verbs, preposition, pronouns; b) *orthographic errors*, e.g. inaccurate double consonants (e.g: *tera* instead of *terra*, *subbito* instead of *subito*); c) *lexical errors*, i.e. misuse of terms.

5 Experiments and Discussion

The system was evaluated with a weighted 7-fold cross validation in which every fold is represented by a different school. It follows that in each experiment the test set is composed by the school documents which are not included in the corresponding training set. The accuracy for each fold is calculated in terms of F-Measures. The final score is the weighted average with respect to the number of student of each school.

Feature	Correlation
9 most correlated features used for feature selection	
Frequency class of verbs	0.212
Percentage of auxiliary verbs in first person plural	-0.168
Number of tokens	0.164
Number of sentences	0.162
Percentage of prepositional dependency relation	0.153
Percentage of auxiliary dependency relation	-0.137
Percentage of auxiliary verbs in indicative	-0.136
Type/Token ratio (first 200 tokens)	0.130
Average of characters per token	0.126
Correlation of manually annotated errors	
Grammatical errors per word	-0.103
Orthographic errors per word	-0.119
Lexical errors per word	-0.162

Table 1: Correlations between features and the chronological order of the texts

Four different sets of experiments were devised to test the performance of our system. The experiments differ with respect to the temporal span between the two compared documents used in training and test sets. In the first experiment all pairs of texts written by the same student are used as training and test set, which means that the sets contain pairs of documents with all possible temporal distances (from the minimum to the maximum distance). In the second experiment we compared only the texts written in two different years, so that at least one year occurs between the documents. In the third experiment the pairs used in the training and test sets contain the first and the penultimate text written by the same student, whereas in the last experiment the first and the last text of a student were compared. Thus the time period between the texts is the maximum possible, i.e. two years. Every experiment was performed using all features described in section 4 and using only a subset of features resulting from the feature selection process. These features were selected by calculating the correlation between all features (with the exclusion of the *Annotated Error features*) and the chronological order of the texts of each student. For these experiments we selected the nine

	F-Score for each school							Weighted average F-Score
	1	2	3	4	5	6	7	
all texts								
All Features	73.0	68.0	56.5	59.1	64.8	51.8	64.0	**62.7**
Feature Selection	67.3	70.9	50.2	71.4	55.9	57.4	59.5	**61.2**
Feature Selection + Errors	67.3	70.5	54.5	73.4	56.2	57.5	59.2	**61.6**
different years								
All Features	78.1	70.5	52.3	68.5	68.0	44.3	76.7	**64.1**
Feature Selection	77.9	77.4	48.4	67.5	63.6	57.5	59.1	**64.8**
Feature Selection + Errors	77.4	78.2	50.2	67.7	63.6	57.5	58.5	**64.6**
first and penultimate text								
All Features	84.0	92.6	73.9	61.9	55.6	56.5	64.3	**71.7**
Feature Selection	92.0	96.3	65.2	95.2	72.2	58.7	71.4	**79.8**
Feature Selection + Errors	92.0	96.3	70.2	96.3	72.8	62.4	71.4	**81.2**
first and last text								
All Features	100.0	96.3	87.0	81.8	76.3	95.8	78.6	**89.3**
Feature Selection	76.0	96.3	52.2	90.9	78.9	100.0	82.1	**82.8**
Feature Selection + Errors	78.2	96.3	55.2	89.7	80.7	100.0	82.3	**84.1**

Table 2: Results of experiments.

most correlated features corresponding to different linguistic phenomena, reported in Table 1.

The results of all experiments are shown in Table 2. As a general remark, we can note that the bigger the temporal span between the tested documents, the bigger the achieved accuracy. This is due to the fact that the growth of the student writing quality is related to the temporal span. The best accuracy is achieved in the *first and last text* experiment (89.2%) using all features. Since the last text is the Common Prompt written at the end of the second year, this result can be biased by the features capturing prompt-dependent characteristics rather than the language competence evolution. Therefore the result could indicate an overfitting of the model. This assumption is supported by the accuracy achieved in the *first and penultimate text* experiment using all features (71.7%). In this case, the prompts of the written essays differ from school to school.

The *Feature Selection* rows report the results obtained after the feature selection process. Even though in these experiments we considered only nine features (vs. the total number of about 150 features), we can note a general improvement in particular for what concerns the *first and penultimate text* experiment (about 8% points of improvement). These results demonstrate that these nine features are able to capture the evolution of the written language competence at different level of linguistic description. The main competence improvement captured by these features refer to: the use of verbs, in terms of both the frequency class of used verbs (during the language competence evolution the students tend to use less frequent verbs) and the verb structures produced by the students, as it is suggested by the occurrence of features capturing the use of the auxiliary verbs; basic characteristics of the sentence, such as the sentence and word length; and features referring to lexical richness (the type/token ratio feature). Interestingly, these features are in line with the results obtained by socio-pedagogical studies reported in (Barbagli et al., 2014). It is noticeable that the results of the third school are significantly the lowest ones when feature selection is used. This is due to the fact that the nine selected features do not significantly change in the student essays over the time for this school. Further investigations is part of our current studies where we are combining student background information with the competence evolution.

The *Feature Selection + Errors* rows show the results obtained using the manually annotated errors combined with the nine selected features. As we can note, in almost all cases we obtained only a small improvement with respect to the feature selection results. This result is of pivotal importance demonstrating that the written language competence is mainly captured by relying on features that refer to the essay linguistic structure rather then by focusing on errors (also when manually annotated). This is in line with the observation of De Mauro (1977) who claims that, in particular for what concerns orthographic errors, the language competence is not related with the orthography correctness.

References

Y. Attali and J. Burstein. 2006. Automated Essay Scoring with e–rater V.2. *Journal of Technology, Learning, and Assessment*, 4(3).

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia.

A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2014. Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado. *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it)*, 9–10 December, Pisa, Italy.

A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2015. CItA: un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1 Annotato con Errori. *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC-it)*, 2–3 December, Trento, Italy.

D. Biber and S. Conrad. 2009. Genre, Register, Style. Cambridge: CUP.

A. Cimino, F. Dell'Orletta, G. Venturi, and S. Montemagni. 2013. Linguistic Profiling based on Generalpurpose Features and Native Language Identification. *Proceedings of Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, June 13, pp. 207-215.

C. Chang and C. Lin. 2001. LIBSVM: a library for support vector machines. *Software available at* `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

K. Collins-Thompson. 2014. Computational Assessment of text readability. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, 165:2, John Benjamins Publishing Company, 97-135.

A.T. Corbett and J.R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user–adapted interaction*, 4(4), 253–278.

P. Deane. 2014. Using Writing Process and Product Features to Assess Writing Quality and Explore How Those Features Relate to Other Literacy Tasks. *ETS Research Report Series*.

P. Deane and T. Quinlan. 2010. What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, 2(2), 151–177.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.

F. Dell'Orletta, M. Wieling, A. Cimino, G. Venturi, and S. Montemagni. 2014. Assessing the Readability of Sentences: Which Corpora and Features. *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014)*, Baltimore, Maryland, USA.

T. De Mauro. 1977. *Scuola e linguaggio*. Editori Riuniti, Roma.

T. De Mauro. 1999. *Grande dizionario italiano dell'uso (GRADIT)*. Torino, UTET.

C. Ekanadham and Y. Karklin. 2015. T-SKIRT: Online Estimation of Student Proficiency in an Adaptive Learning System. *Proceedings of the 31st International Conference on Machine Learning*.

C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein. 2015. Deep Knowledge Tracing. *ArXiv e-prints:1506.05908 2015*.

U. Quasthoff, M. Richter, and C. Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. *Proceedings of the fifth international Language Resources and Evaluation Conference (LREC-06)*, Genoa, Italy.

Learning Grasping Possibilities for Artifacts: Dimensions, Weights and Distributional Semantics

Irene Russo*, Irene De Felice **
Istituto di linguistica Computazionale "A . Zampolli" CNR *
University of Pisa **
{irene.russo,irene.defelice}@ilc.cnr.it

Abstract

English. In this paper we want to test how grasping possibilities for concrete objects can be automatically classified. To discriminate between objects that can be manipulated with one hand and the ones that require two hands, we combine conceptual knowledge about the situational properties of the objects, which can be modeled with distributional semantic methodologies, and physical properties of the objects (i.e. their dimensions and their weights), which can be found on the web through crawling.

Italiano. *In questo articolo vogliamo testare come le possibilità di manipolazione degli oggetti concreti possano essere classificate automaticamente. Per distinguere tra oggetti che possono essere manipolati con una mano e oggetti che richiedono due mani, combiniamo conoscenza concettuale sulle proprietà situazionali dell'oggetto - rappresentandola secondo il paradigma della semantica disribuzionale - con le proprietà fisiche degli oggetti (le loro dimensioni e il loro peso) estratte dal web mediante crawling.*

1 Introduction

Distributional semantic models of word meanings are based on representations that want to be cognitively plausible and that, as a matter of fact, have been tested to produce results correlated with human judgments when concepts similarity and automatic conceptual categorizations are the aim of the experiment (Erk, 2012; Turney and Pantel, 2010).
These approaches share the idea that two nominal concepts are similar and can be clustered in the same group if the corresponding lexemes occur in comparable linguistic contexts.
Their success is also due to the expectations of the Natural Language Processing (henceforth NLP) community: both for count and predictive models of distributional semantics (Baroni et al. 2014), the core idea is that encyclopedic knowledge packed in a big corpus can improve the performance in tasks such as word sense disambiguation.
However, purely textual representations turn out to be incomplete because in language learning and processing human beings are exposed to perceptual stimuli paired with linguistic ones: the old AI dream to ground language in the world requires the mapping between these two sources of knowledge.
One of the aim of this paper is to understand how much physical knowledge can be retrieved in language. Can distributional representations of concrete nouns be helpful for the automatic classification of objects, when grasping possibilities are the focus? Could they help to discriminate between objects that can be manipulated with one hand and the ones that require two hands? More generally, how much knowledge about the physical world can be found in language?
Inspired by the cognitive psychology literature on the topic, in this paper artifactual categories are theorized as situated conceptualization where physical and situational properties meet (Barsalou 2002). These situational properties describe a physical setting or event in which the target object occurs (as *grocery store, fruit basket, slicing, picnic* for *apple*). In an action-based categorization of objects, these kinds of properties function as a complex relational system, which links the physical structure of the object, its use, the background settings, and the design history (Chaigneau et al. 2004). Situational properties can be derived from distributional semantic models, where each

co-occurrence vector approximates the encyclopedic knowledge about its referent.

A complementary, but more action-oriented idea, is the psychological notion of affordance as the possibilities for actions that every environmental object offers (Gibson 1979). Conceptual information concerning objects affordances can be partially acquired through language, considering verb-direct object pairs as the linguistic realizations of the relations between the actions that can be performed by an agent, and the objects involved in those actions. Affordance verbs, intended as verbs that select a distinctive action for a specific object, can be discovered through statistical measures in corpora (Russo et al. 2013).

The main assumption of this paper is that the primary affordance for grasping of an artifact largely depends on its physical properties, in particular dimensions and weight. Such features are found in e-commerce websites. Extracting these values for many similar items, for example for all instances of "plate", may help to automatically represent average dimensions for that object. However, combining this knowledge with situational properties of objects modeled as distributional semantics vectors can help understanding if they can be combined. This issue is relevant for the implementation of a module that automatically classifies grasping possibilities for objects in embodied robotics.

The paper is structured as follow: section 2 reports on the manual annotation of grasping possibilities for a set of 143 artifacts, discussing the definition of the gold standard that will be the dataset for classification experiments in section 3. Section 4 presents conclusions and ideas for future work.

2 Manual Annotation of Grasping Possibilities

Concerning grasping possibilities for concrete objects, we expect as relevant several features. First of all, objects dimensions strongly influence the type of grasp afforded by objects. For instance, we are likely to grasp a tennis ball with a whole hand, but a soccer ball with two hands: the difference between the two spheres clearly is in their diameter.

Heavy objects require a type of grasp different from the one required by the light ones. Apart from these features, we should also consider more subjective factors, such as culture, past experience with objects, or intentions. This is particularly evident for artifacts and tools, that are the kind of objects most typically involved in manipulation and grasping and that often have a part that is specifically designed (or more suited than others) for grasping, for its shape and conformation, such as a handle (which we may call affording parts; cf. De Felice, 2015; in press). However, such parts (e.g. the handle of a cup) are usually grasped when the agents intention is to use the object for its canonical function (e.g. to drink from the cup), whereas in other cases it may be ignored and a different grasp could be performed (e.g. the whole cup might be taken from the above if we simply wanted to displace it).

Therefore, we can individuate at least four different grasp types afforded by concrete entities (cf. infra): the undifferentiated one-handed or two-handed grasps; a grasp by part, i.e. directed to a specific part of the object; a grasp with instrument, for substances, aggregates or every sort of things usually manipulated with some other object.

In order to obtain a gold standard annotation of artifacts grasping possibilities, we first searched WordNet 3.0 for all the nouns that have artifact as hyperonym, obtaining a list of 1510 synsets. From this list, we chose the nouns that have enough pictures as products sold on amazon.com, since it was our intention to extract objects dimensions from this website for classification experiments (cf. 3). We selected the nouns for which at least 15 pages about that object sold on amazon.com were homogeneous - i.e. they contain objects of the same type- reducing noise caused by the crawling strategy. We obtained a total number of 143 nouns. Then, for each of these nouns, we manually annotated the type of grasp afforded by the object, according to the following classes:

- One-handed grasp: this kind of grasp is for objects that have no handles or protruding parts suited for the grasp, and that can be grasped by using only one hand. The size of two of the objects dimensions (length, width or thickness) usually does not exceed the maximum span of a hand with at least two fingers bent in order to grasp and hold something. E.g.: bowl, bottle, candle, shell, necklace, clothes peg.

- Two-handed grasp: this kind of grasp is for objects that have no handles or protruding

Table 1: Number of items per classes in the gold standard.

class	#nouns
onehand	43
onehandORpart	1
oneORtwohand	25
part	23
twohand	73
twohandORpart	3

parts suited for the grasp, and that are usually grasped with two hands, because their size exceeds the maximum span of a single hand. E.g.: board, soccer ball, player piano, table, computer.

- Grasp by part: this kind of grasp is for: (i) small or large objects that have a part specifically designed for the grasping; (ii) entities that have a well identifiable part that, even if it is not specifically designed for this specific purpose, is more suited than others for the grasping thanks to its shape and conformation. E.g. knife, jug, axe, trolley, bag.

- Grasp with instrument: this kind of grasp is mainly for substances, aggregates, and entities which cannot be (or are usually not) controlled without using some other object (an instrument, generally a container). E.g. water, broth, flour, bran, sand.

For several objects more than one grasping possibility is plausible, depending on the size (a plate can be small or big) or on the availability of a container (sand can be grasped by hand).
The dataset of 143 nouns have been annotated by two annotators and the inter-annotator agreement was 0.66. Since we need a gold standard for experiments, we managed disagreements reaching a consensus on every noun.
The gold standard contains items assigned to 6 classes, distributed as in Table 1.

3 Semantic and physical knowledge about artifacts: guessing grasping possibilities

The way humans can grasp an object can be designed as a function that depends on multiple variables, such as the presence of affording parts (i.e. handle for bag), its shape, its dimensions, its weight and the final aim of the action of grasping, modeled here as part of the situational properties. In this paper we want to test which one of these features can help in classifying artifacts that have been manually annotated according to 6 categories (see par. 2). In particular we experiment with a combination of 4 features provided for each noun:

- distributional semantics information from two corpora (GoogleNews and instructables.com) obtained with word2vec toolkit (Mikolov et al. 2013);

- average dimensions (height, length and depth) for each object, obtained crawling at least 15 pages per object from amazon.com;

- average weight for each object, obtained crawling at least 15 pages per object from amazon.com;

- co-occurrence matrix in the corpus instructables.com with nouns that are affording parts, extracting the syntactic pattern AFFORDING PART NOUN of ARTIFACT (e.g. "handle of the bag").

Because all the big corpora available contain in general news or web crawled texts that don't mention concrete actions and concrete objects so often, we choose to build a smaller but coherent corpus of do-it-yourself instructions, with the assumption that it will contain frequent instances of concrete language.
We crawled from the website instructables.com all the titles and descriptions for the projects available online in six categories (e.g. technologies, workshop, living, food, play, outside). Cleaned of the html code, the instructables.com corpus has 17M tokens; each project was parsed with the Stanford parser (de Marneffe and Manning 2008). To test if a do-it-yourself instructions corpus is useful with respect to a generic one, we represent each noun in the following experiment as a vector extracted from GoogleNews with word2vec toolkit (Mikolov et al. 2013) but also as a vector extracted from the instructables.com corpus trained with the same toolkit. These are the purely textual representations we experimented with; to complete this knowledge we added extracted information about dimensions, weight and affording parts for 143 objects.
The list of objects' parts that afford grasping and

Table 2: Precision and recall for 8 combinations of features for the 6 classes dataset.

features	Precision	Recall
instructables.com	0.113	0.336
GoogleNews	0.113	0.336
weight	0.364	0.406
dimensions	0.413	0.517
dimensions+weight	0.561	0.531
affording parts	0.25	0.399
instructables.com + all	0.443	0.552
GoogleNews + all	0.458	0.559

Table 3: Precision and recall for 8 combinations of features on two-classes dataset ("onehand" VS "twohand").

features	Precision	Recall
GoogleNews	0.846	0.846
weight	0.715	0.714
dimensions	0.851	0.846
dimensions+weight	0.831	0.802
affording parts	0.63	0.615
GoogleNews + all	0.846	0.846

are component of the pattern extracted for the feature "affording parts" has been derived with a psycholinguistic test (De Felice 2015). Thirty students of the University of Pisa were interviewed and presented with 42 images of graspable entities. For each picture, they were asked to describe in the most detailed way how they would have grasped the object represented. Among the objects depicted, there were 31 artefacts. From the interviews recorded for these artefacts, we extracted all nouns denoting objects' parts that were named as possible target of the grasp (e.g. the handle for the bag, the cup or the ladle). The list of 78 nouns was then translated in English.

3.1 Classification Experiment

The experiment is based on a multi-label classification, since our dataset consists of 143 nouns denoting artifacts, annotated according to 6 categories. The implementation of Support Vector Multi-Classification is based on LibSVM software (Chang and Lin 2001) in WEKA with 10 fold cross-validation. Table 2 reports the results in terms of precision and recall. The best perfomance depends on information about average dimensions and weight of the objects. Distributional semantics vectors seems useless.

The overall performance is influenced by the fact that some classes are small in the gold standard. For this reason, we experimented with the same features including just the 91 nouns that belong to the "onehand" or "twohand" classes. In Table 3, results show again that dimensions and dimensions plus weight produce good results (with "dimensions" as the best feature), even if they do not improve the performance when combined with distributional vectors that in this case are useful per se. Again, affording parts co-occurences produce the worst performance, mainly because the list of affording parts was originally derived for only 31 artefacts, and not for all the objects considered in our experiment.

4 Conclusions and Future Works

In this paper we test how distributional representations of nouns denoting artifacts can be combined with physical information about their dimensions and weights automatically extracted from an e-commerce website and with co-occurrence information about their affording parts as found in a corpus of do-it-yourself instructions. The starting hypothesis - concerning grasping possibilities as basic manipulative actions for object - was that they are conceptually a combination of situational and physical properties.

As a consequence, we expect the best performance from a mixed features models. This hypothesis is not confirmed; for the two-classes dataset ("onehand" VS "twohand") both physical knowledge and distributional semantics vectors give good resultsbut they don't improve the classifier's performance when combined.

These results are in line with the current trend to mix textual and visual features from computer vision algorithms (Bruni et al. 2012) in order to go beyond the limitations of purely textual semantic representations that cannot encode information about colors, dimensions, shapes etc. As future work we plan to integrate the features used for the experiment in this paper with representations of words as bag of visual words derived from the scale-invariant feature transform (SIFT) algorithm (Lowe 1999) that in computer vision helps to detect and describe local features in images.

References

Barsalou, L.W. 2002. Being there conceptually: simulating categories in preparation for situated action. *Representation, Memory, and Development: Essays in Honor of Jean Mandler*,1–15.

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

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 2012. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. *Proceedings of ACM Multimedia*, 1219-1228.

Chaigneau, S.E., Barsalou, L.W., and Sloman, S. 2004. Assessing the causal structure of function. *Journal of Experimental Psychology: General*, 133: 601-625.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1-27:27. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Katrin Erk. 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. Language and Linguistics Compass, 6(10):635-653.

De Felice, I. in press. Objects' parts afford action: evidence from an action description task. In V. Torrens (ed.), Language Processing and Disorders. Newcastle: Cambridge Scholars Publishing.

De Felice, I. 2015. Language and Affordances. PhD thesis, University of Pisa, Italy.

Gibson, J. J. 1979. The Ecological Approach to Visual Perception. Boston: Houghton Mifflin.

Lowe, D.G. 1999. Object recognition from local scale-invariant features. *International Conference on Computer Vision*, pp. 1150-1157.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In COLING Workshop on Cross-framework and Cross-domain Parser Evaluation.

Russo, I., De Felice, I., Frontini, F., Khan, F., and Monachini, M. 2013. (Fore)seeing actions in objects. Acquiring distinctive affordances from language. In B. Sharp, and M. Zock (eds.), Proceedings of The 10th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2013 (Marseille, France, 15-17/10/2013), 151-161.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: vector space models of semantics. J. Artif. Int. Res., 37(1):141-188, January.

Experimenting the use of *catenae* in Phrase-Based SMT

Manuela Sanguinetti
Dipartimento di Informatica,
Università degli Studi di Torino
manuela.sanguinetti@unito.it

Abstract

English. Following recent trends on hybridization of machine translation architectures, this paper presents an experiment on the integration of a phrase-based system with syntactically-motivated bilingual pairs, namely the so-called *catenae*, extracted from a dependency-based parallel treebank. The experiment consisted in combining in different ways a phrase-based translation model, as typically conceived in Phrase-Based Statistical Machine Translation, with a small set of bilingual pairs of such catenae. The main goal is to study, though still in a preliminary fashion, how such units can be of any use in improving automatic translation quality.

Italiano. *L'integrazione di conoscenza linguistica all'interno di sistemi di traduzione automatica statistica é un trend diffuso e motivato dal tentativo di combinare le migliori caratteristiche dei sistemi basati su regole con approcci puramente statistici e basati su corpora. Il presente lavoro si inserisce all'interno di queste ricerche e costituisce uno studio preliminare sull'applicazione di una nozione sintattica basata su dipendenze, quella delle cosiddette "catenae", all'interno di una tipica architettura di traduzione statistica.*

1 Introduction

The hybridization of machine translation systems in order to benefit from both statistical-based and linguistically-motivated approaches is becoming a popular trend in translation field. Such trend is well described in a number of surveys (Costa-Jussá and Farrús, 2014; Costa-Jussá and Fonollosa, 2015) and witnessed by recent initiatives in NLP community, such as the HyTra workshop series[1]. The motivations to this choice can be manifold, but essentially lie in the need to either reduce the costs - both in terms of time and resources - of building a fully rule-based system, or to integrate statistical models or SMT outputs with linguistic knowledge, as this could be useful to capture complex translation phenomena that data-driven approaches cannot handle properly.
Such phenomena are often called translational divergences, or even *shifts* (Catford, 1965), and usually involve a large number of linguistic and extra-linguistic factors.

Our main research interest is the study of such shifts, in particular from a syntactic point of view, and of how such linguistic knowledge could be of any use to overcome the current shortcomings in machine translation.
The preliminary experiment presented here is therefore guided by the second motivation mentioned above: our basic assumption is that supplementing translation models in classical Phrase-Based Statistical Machine Translation (PBSMT) with syntactically-motivated units extracted from parallel treebanks can lead to improvements in machine translation accuracy. This was already demonstrated, for example, in Tinsley (2009), where syntactic contituents were used to improve the translation quality of a PBSMT system. However, instead of a constituency paradigm, we focused on a more dependency-oriented syntactic unit, namely the one of *catena*. The choice of a dependency-paradigm in general is mainly dictated by the acknowledged fact that dependencies can better represent linguistic phenomena typical of morphologically rich and free-word order languages (see e.g. (Covington, 1990; Goldberg et al., 2013)). On the other hand, to capture translation shifts of various nature, it is necessary to consider a syntactic unit that goes beyond the single

[1] http://www.hyghtra.eu/workshop.html

node, as also recently pointed out, e.g., in Deng et al. (2015); hence the introduction of the notion of *catena* in our study.
In order to verify our assumption, we carried out a preliminary experiment performing several translation tasks, with Italian and English as language pair. For this purpose, a typical phrase-based SMT system was built, using for training the translation model various combinations of baseline SMT configurations and pairs of catenae automatically extracted from a parallel treebank, i.e. ParTUT, and then automatically aligned.

The remainder of this paper is thus organized as follows: Section 2 introduces the notion of catena, in Section 3 we describe our use of catenae in this experiment, while in Section 4 we describe the training configurations chosen and discuss the results.

2 Catenae: a brief introduction

A large number of contributions, in MT, provided some hints on the need to infer complex translational patterns - often encoded by one-to-many or many-to-many alignments - by including a more extensive hierarchical notion that goes beyond the mere word level. In constituency frameworks, such notion is fully covered by syntactic phrases, or constituents, while in dependency contexts - where this is not explicitly defined - a number of different approaches have been proposed to tackle the problem; Ding and Palmer (2004) (and follow-up works) proposed the extraction and learning of the so-called *treelets*, which refer to any arbitrary dependency subgraph that does not necessarily goes down to some leaf. Recently though, a new unit type has been defined in dependency framework, which, to a certain extent, linguistically justifies and formalizes the abovementioned notion of treelet (originally conceived for computational purposes only). This is the notion of *catena* (Latin for "chain", pl. *catenae*). In Kiss (2015), a catena is defined as:

> *a single w(ord) or a set of w(ords) C such that for all w in C, there is another w' in C that either immediately dominates or is dominated by w. According to this definition, any given tree or any given subtree of a tree qualifies as a catena.*

As a result, catena is claimed to be more inclusive than constituents, as it does not require the unit to include all the nodes that are dominated. Because of the dominance constraint, however, it cannot be compared to a string either.

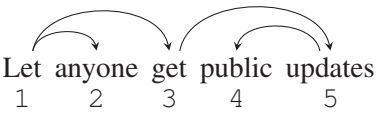


Figure 1: Example of catena.

Figure 1 shows an example of a sentence represented in an unlabelled dependency graph where each word is assigned an identifier (`1, 2, 3, 4, 5`). In the sentence, 15 distinct catenae can be identified (including single nodes)[2]: `[1]`, `[2]`, `[3]`, `[4]`, `[5]`, `[1 2]`, `[1 3]`, `[3 5]`, `[4 5]`, `[1 2 3]`, `[1 3 5]`, `[3 4 5]`, `[1 2 3 5]`, `[1 3 4 5]`, and `[1 2 3 4 5]` (i.e. the whole dependency graph).

A catena may thus include both contiguous and non-contiguous sequences of words, such as *Let get* or *Let get updates*; however, this is not the case for the string "*Let anyone get public*", since there is no direct path to the word "*public*".

The usefulness of catenae in theoretic accounts of complex linguistic phenomena has already been widely shown in literature (Osborne, 2005; Osborne et al., 2011; Osborne and Putnam, 2012; Simov and Osenova, 2015). And to our knowledge, only a few NLP studies (even beyond the bare MT field) exploited this syntactic unit for some practical purpose. The only study we are aware of so far is that of Maxwell et al. (2013), who present an approach based on catenae to *ad hoc* Information Retrieval. It is our opinion, however, that even translation issues can be tackled by integrating such inclusive notion; catenae can be used, for example, to explain and properly identify those cases of one-to-many or many-to-many correspondences, typical of several translation shifts, such as different underlying syntactic structures, MWEs or idioms. For this reason we attempted to exploit them in this experimental study, among other purposes.

3 Catenae extraction and alignment

The first preprocessing step in this experiment consists in the extraction of the possible catenae

[2] In accordance with the convention used in (Osborne et al., 2012), the words that form a catena are listed in a left-to-right order, following their linear order in the sentence.

from parse trees of a parallel treebank. The resource we used for this purpose is ParTUT, a recently developed parallel treebank for Italian, English and French[3] (Sanguinetti and Bosco, 2014). The whole treebank currently comprises an overall amount of 148,000 tokens, with approximately 2,200 sentences in the Italian and English sections respectively, and 1,050 sentences for French.

For this experiment, we used the Europarl section of the treebank, retaining only the sentence pairs that have a direct correspondence (1:1), hence using a set of 376 pairs with an average of 10K tokens per language. To each monolingual file, formatted in CoNLL, of this parallel set we then applied the script for the extraction of catenae.

The script basically performs a depth-first search into the dependency tree, and for each node *w* recursively detects all the possible catenae starting from *w* to the nodes that, directly or indirectly, it dominates. The output file thus provides for each sentence a sequence of such catenae (one per line).

Although the parallel sentences perfectly match with each other, this is not obviously the case for catenae as well. For this reason we carried out a further preprocessing step that entailed the automatic alignment of the output English and Italian files containing such catenae. The alignment was performed considering catenae as if they were sentences, thus using the Microsoft Sentence Bilingual Aligner[4] (Moore, 2002) as alignment tool, and setting a high-probability threshold (0.99) in order to have a more accurate - though far more reduced[5] - pairs of parallel catenae. The set obtained in this step consists of about 1,700 pairs (set A), which was further filtered to obtain a separate subset of pairs - 778 in total - where each catena has a 7-token maximum length (set B). Such subset was created so as to be used in a different training configuration during the translation step (see next section).

Once extraction and alignment steps were completed, we proceeded with the translation tasks, as detailed in the next section.

[3] http://www.di.unito.it/~tutreeb/partut.html

[4] Downloadable here: http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[5] The extremely smaller amount of aligned catenae may also be explained by the fact that the order in which the source and target sentences (and catenae, in our case) are listed impacts on the amount and quality of the final alignments.

4 Using catenae in PBSMT

To perform the task, we used Moses (Koehn et al., 2007) as translation toolkit, and set up the system so as to train multiple models, that correspond to the baseline model and to the baseline model augmented with catenae in two different ways.

4.1 Data

Because of its size and availability, the Europarl-v7 parallel corpus (Koehn, 2005) was used for training and testing the system.

To train the baseline translation model, we used a set of 100K parallel sentences, that, however, reduced to an amount of approximately 85K after cleaning up the corpus (we just retained the sentence pairs of up to a 50-tokens length), while we retrieved a far smaller set for tuning (850 sentences) and a set of 1000 sentences for testing.

As we built a system for both translation directions, the language model was computed for both languages using the entire monolingual sets on the English and Italian sides of the corpus (around 1.9M sentences each).

4.2 Experimental setup

The baseline system was built using the basic phrase-based model, which typically does not make any explicit use of linguistic information.

For language modeling, we opted for the trigram option using the IRSTLM toolkit (Federico et al., 2008).

The translation model was computed using the default settings provided by the system guidelines. Word alignment was performed with GIZA++ (Och and Ney, 2003) and 'grow-diag-final-and' as symmetrization heuristic, while a default length of 7 was kept for phrases.

This model, however, was also adapted so as to be configured with three different options:

- to be trained with phrase pairs only (BASELINE)
- to be trained by adding to the baseline training corpus the set A of aligned catenae described in Section 3 (BASELINE+TRAIN)
- to be trained with a combination of multiple sources, i.e. extending Moses' phrase table with the set B of aligned catenae mentioned in Section 3 (BASELINE+CAT)

source sentence	*sia l' Islam che il mondo cristiano sostengono i diritti delle donne*
reference	*for Islam and Christianity both uphold the rights of women*
BASELINE	*both Islam that the Christian world are the rights of women*
BASELINE+TRAIN	*both Islam and Christianity support the rights of women*

Table 1: Translation example.

The second and third configurations were obtained using a simple approach, i.e. concatenating the bilingual catenae *a*) to the training files (BASELINE+TRAIN), and *b*) to the list of the corpus-extracted phrase pairs (BASELINE+CAT); as also suggested in (Bouamor et al., 2012).

The final translation outputs were then evaluated with BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) scores, and results are discussed in the next section.

4.3 Results

The findings emerged from the final evaluation, as also reported in Table 2, show very different results both according to the type of configuration used and to the translation direction. However, from such diversified outputs, relevant data can be highlighted.

Such relevance mainly consists in the improvement of translation quality when simply augmenting the training corpus with other external data (BASELINE+TRAIN). As a matter of fact, although such improvement is far from significant in terms of BLEU score in Italian-to-English translation, its NIST counterpart, together with the overall quality of English-to-Italian translation show more encouraging results, with an increase from 6.2410 to 6.2599 in NIST score for the first case, and a 0.02 and 0.03 points in BLEU and NIST scores respectively, for the the second one. Table 1 shows an example translation of an Italian sentence comparing BASELINE and BASELINE+TRAIN outputs.

A small improvement is also reported in the NIST score of the Italian-to-English model when adding a set of bilingual catenae into the phrase table (BASELINE+CAT). This case as well may not be particularly significant in itself, though however encouraging, considering the small amount of data that was added with respect to the baseline system. On the other hand, such set does not seem to affect at all the English-to-Italian model. As a matter of fact, it produces the same hypothesis translation than the one produced with the baseline configuration, and both translations are reported to have a lower translation quality with respect to the first system pair, despite the same amount of training data was used in both directions, even for the language modeling. Such result can be probably explained with some error in the tuning process, while the overall lower quality may be explained, we hypothesize, as an effect of translating into a morphologically richer language - though more in-depth studies should be carried out to support this hypothesis.

		BLEU	**NIST**
It-En	BASELINE	0.2610	6.2410
	BASELINE+TRAIN	**0.2621**	**6.2599**
	BASELINE+CAT	0.2609	6.2582
En-It	BASELINE	0.2241	5.9161
	BASELINE+TRAIN	0.2427	6.2194
	BASELINE+CAT	0.2241	5.9161

Table 2: Experimental evaluation of Italian-to-English and English-to-Italian translation quality under a baseline PBSMT system, and other two PBSMT systems integrated with catenae.

5 Conclusions

The paper presented a small experiment on the combined use of linguistic knowledge - in the form of syntactically-motivated translation units - and statistical model provided by state-of-the-art machine translation techniques. The results reported here are to be considered preliminary, as they suffer from the absence of systematic procedures and data that could not have been applied so far due to lack of time and proper resources. Still, considering these shortcomings, translation evaluation, at least in one direction, produced promising results. There is however a lot of work to do to under this respect in order to effectively improve translation quality with the help of such linguistic information; for example by scaling up this experiment using a larger set of external data, or using different training configurations, so as to have multiple

sources of comparison for final assessments and considerations.

References

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multiword expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), May.

John C. Catford. 1965. *A Linguistic Theory of Translation: An Essay on Applied Linguistics*. Oxford University Press.

Marta R. Costa-Jussá and Mireia Farrús. 2014. Statistical machine translation enhancements through linguistic levels: a survey. *ACM Computing Surveys (CSUR)*, 46:1–28, January.

Marta R. Costa-Jussá and José A.R. Fonollosa. 2015. Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32:3—10, July.

Michael A. Covington. 1990. Parsing discontinuous constituents in dependency grammar. *Comput. Linguist.*, 16(4):234–236, December.

Dun Deng, Nianwen Xue, and Shiman Guo. 2015. Harmonizing word alignments and syntactic structures for extracting phrasal translation equivalents. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST9)*, pages 1–9.

Duan Ding and Martha Palmer. 2004. Automatic learning of parallel dependency treelet pairs. In *Proceedings of the 1rst International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 233–243.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pages 22–26.

Yoav Goldberg, Yuval Marton, Ines Rehbein, and Yannick Versley, editors. 2013. *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*. Association for Computational Linguistics, Seattle, Washington, USA, October.

Tibor Kiss. 2015. *Syntax - Theory and Analysis. Vol. 2*. Walter de Gruyter GmbH & Co KG, Berlin.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

K. Tamsin Maxwell, Jon Oberlander, and W. Bruce Croft. 2013. Feature-based selection of dependency paths in ad hoc information retrieval. In *ACL '13*, pages 507–516.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA-02)*, pages 135–144.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 1(29):19–51.

Timothy Osborne and Michael Putnam. 2012. Constructions are catenae: Construction grammar meets dependency grammar. *Cognitive Linguistics*, 23:165–215.

Timothy Osborne, Michael Putnam, and Thomas Gross. 2011. Bare phrase structure, label-less trees, and specifier-less syntax: Is minimalism becoming a dependency grammar? *The Linguistic Review*, 28:315–364.

Timothy Osborne, Michael Putnam, and Thomas Gross. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15:354–396.

Timothy Osborne. 2005. Beyond the constituent: A dependency grammar analysis of chains. *Folia Linguistica*, 39:251–297.

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manuela Sanguinetti and Cristina Bosco. 2014. Parttut: The turin university parallel treebank. In Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi, editors, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, pages 51–69.

Kiril Simov and Petya Osenova. 2015. Catena operations for unified dependency analysis. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 320–329.

John Tinsley. 2009. Resourcing machine translation with parallel treebanks. phd thesis.

Cross-language projection of multilayer semantic annotation in the NewsReader Wikinews Italian Corpus (WItaC)

Manuela Speranza, Anne-Lyse Minard
Fondazione Bruno Kessler, Trento
{manspera,minard}@fbk.eu

Abstract

English. In this paper we present the annotation of events, entities, relations and coreference chains performed on Italian translations of English annotated texts. As manual annotation is a very expensive and time-consuming task, we devised a cross-lingual projection procedure based on the manual alignment of annotated elements.

Italiano. *In questo articolo descriviamo l'annotazione degli eventi, delle entità, delle relazioni e delle catene di coreferenza realizzata su traduzioni in italiano di testi inglesi già annotati. Essendo l'annotazione manuale un compito molto dispendioso, abbiamo ideato una procedura di proiezione interlinguale basata sull'allineamento degli elementi annotati.*

1 Introduction

The NewsReader Wikinews Italian Corpus (WItaC) is a new Italian annotated corpus consisting of English articles taken from Wikinews[1] and translated into Italian by professional translators.

The English corpus was created and annotated manually within the NewsReader project,[2] whose goal is to build a multilingual system able to reconstruct storylines across news articles in order to provide policy and decision makers with an overview of what happened, to whom, when, and where. Semantic annotations in the NewsReader English Wikinews corpus span over multiple levels, including both intra-document annotation (entities, events, temporal information, semantic roles, and event and entity coreference) and cross-document annotation (event and entity coreference). As manual annotation is a very expensive and time-consuming task, we devised a procedure to automatically project the annotations already available in the English texts onto the Italian translations, based on the manual alignment of the annotated elements in the two languages.

The English corpus, taken directly from Wikinews, together with WItaC, being its translation, ensures access to non-copyrighted articles for the evaluation of the NewsReader system and the possibility of comparing results in the two languages at a finegrained level.

WItaC aims at being a reference for the evaluation of storylines reconstruction, a task requiring several subtasks, e.g. semantic role labeling (SRL) and event coreference. In addition, it is part of a cross-lingually annotated corpus,[3] thus enabling for experiments across different languages.

The remainder of this article is organized as follows. We review related work in Section 2. In Section 3 we present the annotations available in the English corpus used as the source for the projection of the annotation. In Section 4 we detail some adaptations of the guidelines specific for Italian. In Sections 5 and 6 we describe the annotation process and the resulting WItaC corpus. Finally, we conclude presenting some future work.

2 Related work

A number of semantically annotated corpora are available for English, whereas most other languages are under-resourced. As far as Italian is concerned, WItaC is the first corpus offering annotations of entities, events, and event factuality, together with semantic role labeling and cross-document coreference annotation.

For entities and entity coreference, the reference Italian corpus is I-CAB (Magnini et al., 2006),

[1] Wikinews (http://en.wikinews.org) is a collection of multilingual online news articles written collaboratively in a wiki-like manner.

[2] http://www.newsreader-project.eu/

[3] The NewsReader consortium has annotated also the Spanish and Dutch translations of the same Wikinews articles.

which is annotated with entities, time expressions (following the TIMEX2 standard), and intra-document entity coreference; for cross-document person entity coreference, we refer to CRIPCO (Bentivogli et al., 2008). Regarding temporal information and event factuality, two annotated corpora are available: respectively, the EVENTI corpus (Caselli et al., 2014), used as the evaluation dataset for the EVENTI task at Evalita 2014 and annotated with events, time expressions (TIMEX3), temporal signals, and temporal relations, and FactIta Bank (Minard et al., 2014), a subsection of EVENTI annotated with event factuality.

To the best of our knowledge there exist no other Italian corpora with semantic role labeling and event cross-document coreference annotation. The reference corpus for SRL in English is the CoNLL-2008 corpus (Surdeanu et al., 2008). For cross-document coreference, the ECB+ corpus (Cybulska and Vossen, 2014) has recently been created extending the ECB corpus.

The method we propose for cross-lingual annotation projection taking advantage of the alignment between texts in two different languages is similar to other methods used, for example, to build annotated corpora with semantic roles (Padó and Lapata, 2009), temporal information (Spreyer and Frank, 2008; Forascu and Tufi, 2012), and coreference chains (Postolache et al., 2006). However, previous work is based on the use of corpora aligned at the word level either manually, which is very time-consuming, or automatically, which is error prone. On the other hand, our method envisages a manual alignment at the markable level, where the extent of each element is annotated on the translated text and then aligned to the English annotated element on a semantic rather than syntactic basis.

3 Annotation available in the English source corpus

The NewsReader Wikinews English corpus contains intra-document semantic annotation and cross-document coreference annotation.

3.1 Annotation at document level

The annotation is based on the NewsReader guidelines (Tonelli et al., 2014) and was performed using the CAT tool (Bartalesi Lenzi et al., 2012). The first five sentences (including the headline) of each document contain the following annotations: markables, relations, and intra-document coreference.

Markable annotation. Textual realizations of entity instances, referred to as entity mentions, are the portions of text in which entity instances of different types (people, organizations, locations, financial entities, and products) are referenced within a text. Each entity mention is described through that portion of text (extent) and two optional attributes, i.e. syntactic head and syntactic type.

The textual realization of an event, the event mention, can be a verb, a noun, a pronoun, an adjective, or a prepositional construction. It is annotated through its extent and a number of attributes, e.g. predicate (lemma), part-of-speech, and factuality. Factuality attributes (van Son et al., 2014) of an event include time, certainty and polarity.

The annotation of temporal expressions is based on the ISO-TimeML guidelines (ISO, 2012) , and thus includes durations, dates (e.g. the document creation time), times, and sets of times, with the following attributes : type, normalized value, anchorTimeID (for anchored temporal expressions), and beginPoint and endPoint (for durations).

Numerical expressions include percentages, amounts described in terms of currencies, and general amounts. Temporal signals, inherited from ISO-TimeML, make explicit a temporal relation. Similarly, causal signals (C-SIGNALS) indicate the presence of a causal relation between two events (e.g. *because of*, *since*, *as a result*, and *the reason why*).

Relation annotation. Based on the TimeML approach (Pustejovsky et al., 2003), temporal relations (e.g. 'before', 'after', 'includes', and 'ends') are used to link two event mentions, two temporal expressions or an event mention and a temporal expression. The annotation of subordinating relations also leans on TimeML, although its scope was reduced to the annotation of reported speech.

In addition, explicit causal relations between causes and effects denoted by event mentions have been annotated taking into consideration the *cause*, *enable*, and *prevent* categories of causation, and grammatical relations have been created for events that are semantically dependent on another event, to link them to their governing content verb/noun.

Semantic role labeling is modeled through the HAS_PARTICIPANT relation, a one-to-one relation linking an event mention to an entity mention playing a role in the event. PropBank (Bonial et al., 2010) is used as the reference framework for the assignment of the semantic role to each relation.

Intra-document event and entity coreference. The annotation of coreference chains that link different mentions to the same instance is based on the REFERS_TO relation.

Entity instances are described through the non text-consuming ENTITY tag and the two attributes entity type and tag descriptor; similarly, event instances are described through the non text-consuming EVENT tag and the two attributes event class and tag descriptor.

3.2 Annotation at corpus level

Annotation at the corpus level (Speranza and Minard, 2014), performed using the CROMER tool (Girardi et al., 2014), relies on the creation of corpus instances (both entities and events) and on links holding between each mention and the corpus instance it refers to. Corpus instances are described through a unique instance ID and the DBpedia URI (when available). Annotation consists of:

- cross-document entity coreference in the first five sentences;
- cross-document entity and event coreference in the whole document for a subset of 44 seed entities (i.e., annotation and coreference of all mentions referring to the seed entities and of the events of which the entities are participants).

4 Italian language specific annotations

We adopted the NewsReader guidelines already available for English with some minor language specific adaptations, as described in detail in Speranza et al. (2014). For this reason the data on inter-annotator agreement provided for English by van Erp et al. (2015) can be used as a reference.

For the annotation of clitics, which do not exist in English, we decided to leave the annotation at the word level, rather than split it into smaller units, so as to be consistent with annotations on existing corpora, e.g. I-CAB (Magnini et al., 2006). So in the case of a token composed of a verb (i.e. an event mention) and a clitic corresponding to a pronominal mention of a markable entity, the whole token was annotated both as an entity and as an event. The syntactic head attribute of the entity mention, having as value the clitic, and the predicate attribute of the event mention, having as value the verbal root, contribute to distinguish the two annotated elements (see [1]).

(1) *Aveva già deciso di dargli un aiuto* ('He had already decided to give him some help')

EVENT_MENTION: [dargli], `pred` "dare"
ENTITY_MENTION: [dargli], `head` "gli"

As Italian, unlike English, is a null-subject language where clauses lacking an explicit subject are permitted, we devised specific guidelines that allowed us to straightforwardly align English pronouns to Italian null subjects. In particular, null subjects having finite verb forms as predicates and referring to existing entity instances (see [2]) were marked through the creation of an empty (i.e. non text-consuming) ENTITY_MENTION tag, which was then linked to other markables following the guidelines for regular text consuming entity mentions; in addition, annotators filled the tag descriptor attribute with a human friendly name and the sentence number (e.g. "He-LuiS2" for the null subject in [2]).

(2) *Obama fece un discorso. [Ø] Disse che [...]* ('Obama gave a speech. [He] said that [...]')

The annotation of modals for Italian is based on It-TimeML, where they are marked as events like all other verbs.[4]

5 Annotation of WItaC

The method we propose for the annotation of the Italian corpus consists of cross-lingual projection of annotation from a source corpus to a target corpus; it enabled us to reduce the effort by approximately three times. The annotation was performed in five steps starting with a file containing the source English annotated text and the Italian translation aligned at the sentence level.

1. Mention annotation. The first step of the annotation, performed using the CAT tool, consisted of the identification and annotation of all markable extents.

2. Alignment. The use of CAT, which is highly customizable, enabled us to set up the alignment between Italian and English markables by simply adding to the Italian markables a new attribute which takes as value the ID of a different markable. Annotators filled this attribute with the corresponding English markable by using drag-and-drop. In some cases it was also necessary to mark the attributes and/or relations that should not be imported (by writing a note in the comment attribute), or to create extra relations.[5] If a mention had no equivalent, annotators filled in the values of the attributes

[4] In WItaC modals are also linked to their governing verb through a grammatical link.

[5] No exceptions were needed for aligning null subjects.

and created the relations in which it was involved and, if it did not already exist, the instance to which it referred.

3. Automatic projection. The automatic projection was performed using a Python script working on the XML files produced by the CAT annotation tool. For each article, the script takes as input the file containing both the English fully annotated text and the Italian text on which the annotated markables have been aligned. It produces as output a file in which the Italian text has been enriched with the annotations imported from English, i.e. the event instances, the entity instances, the relations (including the REFERS_TO relation which models intra-document coreference), and the values of the non-language-specific attributes (unless a specific comment is present).

4. Manual revision. Manual revision consists of an overall check of the annotations imported automatically; in particular, it involves the annotation of the language specific attributes and the deletion of the relations that had been marked as non-importable (using the CAT tool).

5. Projection of cross document coreference. The projection of the cross-document annotation consists of importing coreference from the English corpus taking advantage of the alignment performed in the second step and extending the entity and event instances by importing the IDs of the English instances and their DBpedia URIs.

6 Dataset Description

WItaC is composed of 120 articles. In Table 1 we give the size of the whole corpus and the size of the "first 5 sentences" section, i.e. the subsection annotated with markables, relations, intra-document coreference and cross-document entity coreference. In total 6,127 markables have been annotated in Italian; of these, 5,580 are aligned to English markables while 547 have no English correspondent.

	Whole corpus		First 5 sentences	
	Ita.	Eng.	Ita.	Eng.
# files	120	120	120	120
# sentences	1,845	1,797	597	597
# tokens	44,540	40,231	15,676	13,981

Table 1: Italian and English corpus size

Exploiting the alignment, relations and attributes have been imported automatically. For only 5.7% of the markables the attributes could not be projected (e.g. two events with different PoS). In Table 2 we present the number of markables and relations annotated in the Italian corpus. Out of the total 2,709 entity mentions, 56 are null subjects aligned with English pronominal entity mentions.

Markables		Relations	
EVENT_MENTION	2,208	SLINK	220
ENTITY_MENTION	2,709	TLINK	1,711
TIMEX3	507	CLINK	61
VALUE	415	GLINK	300
SIGNAL	253	HAS_PART	1,865
C-SIGNAL	35		
Total	6,127	Total	4,157
Instances		**Coreference chains**	
EVENT INSTANCE	1,773	REFERS_TO	3,054
ENTITY INSTANCE	1,281		
Total	3,054		

Table 2: Annotations in the first five sentences

As a result of the projection of event and entity cross-document coreference chains from English, WItaC contains 740 entity instances and 887 event instances annotated at the corpus level. Annotation by projection enables us to also have cross-lingual annotation, which means that the instances are shared between English and Italian.

7 Conclusions and future work

We have presented WItaC, a new corpus consisting of Italian translations of English texts annotated using a cross-lingual projection method. We acknowledge some influence of English in the translated texts (for instance, we noticed an above-average occurrence of noun modifiers, as in "dipendenti Airbus") and in the annotation (for instance, annotators might have been influenced by English in the identification of light verb constructions in the Italian corpus). On the other hand, this method enabled us not only to considerably reduce the annotation effort, but also to add a new cross-lingual level to the NewsReader corpus; in fact, we now have two annotated corpora, in English and Italian, in which entity and event instances (in total, over 1,600) are shared.

In the short-term we plan to manually revise the projected relations and add the language-specific attributes. We also plan to use the corpus as a dataset for a shared evaluation task and afterwards we will make it freely available from the website of the HLT-NLP group at FBK[6] and from the website of the NewsReader project.

[6] http://hlt-nlp.fbk.eu/technologies.

Acknowledgments

This research was partially funded the EU NewsReader project (FP7-ICT-2011-8 grant 316404).

References

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Luisa Bentivogli, Christian Girardi, and Emanuele Pianta. 2008. Creating a gold standard for person cross-document coreference resolution in italian news. In *LREC Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. 2010. PropBank Annotation Guidelines, Version 3.0. Technical report, Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder. `http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf`.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI EValuation of Events and Temporal INformation at Evalita. In *Proceedings of the First Italian Conference on Computational Linguistic CLiC-it 2014 & the Fourth International Workshop EVALITA 2014 Vol. II: Fourth International Workshop EVALITA 2014*, pages 27–34.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Corina Forascu and Dan Tufi. 2012. Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. 2014. CROMER: A Tool for Cross-Document Event and Entity Coreference. In *Proceedings of the 9th Internation Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

ISO. 2012. *ISO 24617-1: Language Resource Management. Semantic Annotation Framework (SemAF). Time and Events (SemAF-Time, ISO-TimeML)*. ISO International Standard.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

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*, pages 260–264.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual Annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research*, 36(1):307–340, September.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring Coreference Chains through Word Alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, pages 28–34.

Manuela Speranza and Anne-Lyse Minard. 2014. NewsReader Guidelines for Cross-Document Annotation. Technical Report NWR2014-9, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2014/12/NWR-2014-9.pdf`.

Manuela Speranza, Ruben Urizar, and Anne-Lyse Minard. 2014. NewsReader Italian and Spanish specific Guidelines for Annotation at Document Level. Technical Report NWR2014-6, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2014/02/NWR-2014-61.pdf`.

Kathrin Spreyer and Anette Frank. 2008. Projection-based Acquisition of a Temporal Labeller. In *Proceedings of IJCNLP*, pages 489–496, Hyderabad, India, January.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proceedings*

of the Twelfth Conference on Computational Natural Language Learning, CoNLL '08, pages 159–177, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. 2014. NewsReader Guidelines for Annotation at Document Level. Technical Report NWR2014-2-2, Fondazione Bruno Kessler. http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf.

Marieke van Erp, Piek Vossen, Rodrigo Agerri, Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Egoitz Laparra, Itziar Aldabe, and German Rigau. 2015. Annotated Data, version 2. Technical Report D3-3-2, VU Amsterdam. http://www.newsreader-project.eu/files/2012/12/NWR-D3-3-2.pdf.

Chantal van Son, Marieke van Erp, Antske Fokkens, and Piek Vossen. 2014. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Parsing Events: a New Perspective on Old Challenges

Rachele Sprugnoli[1], Felice Dell'Orletta[2], Tommaso Caselli[3], Simonetta Montemagni[2], Cristina Bosco[4]
[1]Fondazione Bruno Kessler - Università di Trento
[2]Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR, Pisa
[3]VU Amsterdam, [4]Dipartimento di Informatica - Università di Torino
sprugnoli@fbk.eu, felice.dellorletta@ilc.cnr.it,
t.caselli@vu.nl
simonetta.montemagni@ilc.cnr.it, bosco@di.unito.it

Abstract

English. The paper proposes a new evaluation exercise, meant to shed light on the syntax-semantics interface for the analysis of written Italian and resulting from the combination of the EVALITA 2014 dependency parsing and event extraction tasks. It aims at investigating the cross-fertilization of tasks, generating a new resource combining dependency and event annotations, and devising metrics able to evaluate the applicative impact of the achieved results.

Italiano. *L'articolo propone un innovativo esercizio di valutazione focalizzato sull'interfaccia sintassi-semantica per l'analisi dell'italiano scritto che combina i task di EVALITA 2014 su parsing a dipendenze ed estrazione di eventi. Il suo contributo consiste nell'approfondire la combinazione di task che spaziano tra diversi livelli di analisi, nello sviluppo di nuove risorse con annotazione a dipendenze e basata su eventi, e nella proposta di metriche che valutino l'impatto applicativo dei risultati ottenuti.*

1 Introduction

Since the '90s, evaluation campaigns organized worldwide have offered to the computational linguistics community the invaluable opportunity of developing, comparing and improving state-of-the-art technologies in a variety of NLP tasks. ACE[1], MUC[2], CoNLL[3] and SemEval[4] are probably the best-known series of evaluation campaigns that covered syntactic and semantic tasks for English as well as for other languages (e.g. Spanish, Arabic, Chinese). For Italian, EVALITA campaigns[5] have been organized since 2007 around a set of evaluation exercises related to the automatic analysis of both written text and speech.

Over the years, many challenging tasks have been proposed with the aim of advancing state-of-the-art technologies in different NLP areas: to mention only a few, dependency parsing (Nivre et al., 2007), (Bosco and Mazzei, 2013), textual entailment (Bos et al., 2009), frame labeling (Basili et al., 2013) and cross-document event ordering (Minard et al., 2015), all requiring cutting-edge methods and techniques as well as innovative approaches.

Following the fact that, in recent years, research is moving from the analysis of grammatical structure to sentence semantics, the attention in evaluation campaigns is shifting towards more complex tasks, combining syntactic parsing with semantically-oriented analysis. The interest of composite and articulated tasks built by combining basic tasks also lies at the applicative level, since Information Extraction architectures can realistically be seen as integrating components which carry out distinct basic tasks.

Starting from the analysis of the results achieved for individual tasks in EVALITA 2014 and illustrated in Attardi et al. (2015), this paper represents a first attempt of designing a complex shared task for the next EVALITA edition, resulting from the combination of the dependency parsing and event extraction tasks for the analysis of Italian texts. Such a complex task is expected to shed new light onto old challenges by: a.) investigating whether and how the cross-fertilization of tasks can make the evaluation campaign more application-oriented, while also improving individual task results; b.) generating a new resource combining dependency

[1] http://www.itl.nist.gov/iad/mig/tests/ace/
[2] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html
[3] http://ifarm.nl/signll/conll/
[4] http://aclweb.org/aclwiki/index.php?title=SemEval_Portal
[5] http://www.evalita.it

and event annotation; and, c.) devising evaluation metrics more oriented towards the assessment of the applicative impact of the achieved results.

2 Motivation and Background

In recent years, syntactic and semantic dependency parsing have seen great advances thanks to the large consensus on representation formats and to a series of successful evaluation exercises at CoNLL (Surdeanu et al., 2008; Hajič et al., 2009) and SemEval (Oepen et al., 2014; Oepen et al., 2015). However, access to the content, or meaning, of a text has not reached fully satisfactory levels yet. Current developments of data-driven models of parsing show that the recovery of the full meaning of text requires simultaneous analysis of both its grammar and its semantics (Henderson et al., 2013), whose interaction is still not well understood and varies cross-linguistically.

Since the CoNLL 2008 shared task (Surdeanu et al., 2008) much research has focused on the development of systems able either to jointly perform syntactic and semantic dependency tasks or to tackle them independently by means of pipelines of NLP modules specialized in the various subtasks (first full syntactic parsing and then semantic parsing). Insights on the linguistic relatedness of the two tasks derived from the comparison of joint and disjoint learning systems results. Another example is the SemEval 2010 "Task 12: Parser Evaluation using Textual Entailments (Yuret et al., 2010)" (PETE), aimed at recognizing textual entailment based on syntactic information only and whose results highlighted semantically relevant differences emerging from syntax. The evaluation exercise is closer to an extrinsic evaluation of syntactic parsing by focusing on semantically relevant differences.

At EVALITA 2014, two evaluation exercises for the analysis of written text, Dependency Parsing (Bosco et al., 2014) and EVENTI (Caselli et al., 2014), have provided separate evaluations of these two levels of analysis: syntax and semantics, respectively. The relation between the two levels of analysis was investigated in the Dependency Parsing task by setting up a semantically-oriented evaluation assessing the ability of participant systems to produce suitable and accurate output for Information Extraction. Based on measures such as Precision, Recall and F1, this evaluation has been carried out against a subset of 19 semantically-loaded dependency relations (e.g. subject, direct object, adjectival complement and temporal modifier among others). On the other hand, in the EVENTI exercise, syntactic information was considered to play a relevant role for at least two of the subtasks: event detection and classification (subtask B) and temporal relation identification and classification (subtask C).

Dependency parsing is now a key step of analysis from which higher-level tasks (e.g. semantic relations, textual entailment, temporal processing) can definitely benefit. Event Extraction is a high-level semantic task which is strictly connected to morphology and syntax both for the identification of the event mentions and for their classification. Event Extraction differs from standard semantic parsing as not all event mentions have semantic dependencies and it involves a wider range of linguistic realizations (such as verbs, nouns, adjectives, and prepositional phrases) some of which have not been taken into account so far in standard semantic parsing tasks. Despite the recognized influence of one level of analysis on the other, no systematic bi-directional analysis has been conducted so far. To gain more insight on the syntax-semantics interface more focused and complex evaluation exercises need to be setup and run.

In this paper we propose a new evaluation exercise, named "Parsing Events", which aims at shedding new light on the syntax-semantics interface in the analysis of Italian written texts by investigating whether and to what extent syntactic information helps improving the identification and classification of events, and conversely whether and to what extent semantic information, event mentions and classes, improve the identification and classification of dependency relations.

3 Task Description

Parsing Events will qualify as a new evaluation exercise for promoting research in Information Extraction and access to the text meaning for Italian. The exercise, which will start from previous research and datasets for Dependency Parsing and Temporal Processing of Italian, aims at opening a new perspective for what concerns the evaluation of systems to be carried out both at a high level, targeting complex Information Extraction architectures, and at a low level, as single components. The Parsing Events exercise will be thus articulated as follows: a main task, joint dependency parsing and event extraction, and two subtasks, dependency

parsing and event extraction, respectively.

Main task - Joint Dependency Parsing and Event Extraction: The main task will test systems for Dependency Parsing and Event Extraction. Systems have to determine dependency relations based on the ISDT[6] (Bosco et al., 2013) scheme and identify all event mentions as specified in the EVENTI annotation guidelines (Caselli and Sprugnoli, 2014). This will imply to identify the event mentions and fill the values of target attributes. To better evaluate the influence of syntactic information in Event Extraction, the set of event attributes which will be evaluated will be extended to include CLASS, TENSE, ASPECT, VFORM, MOOD and POLARITY. Participants will be given annotated data with both syntactic and event annotations for training. Ranking will be performed on the F1 score of a new evaluation measure based on Precision and Recall for event class and dependency relation.

Subtask A - Dependency parsing The subtask on Dependency Parsing will be organized as a classical dependency parsing task, where the performance of different parsers can be compared on the basis of the same set of test data provided by the organizers. The main novelty of this task with respect to the traditional dependency parsing task organized in previous EVALITA campaigns is that available information will also include event-related information.

Subtask B - Event extraction The Event Extraction subtask will be structured as the Subtask B of the EVENTI 2014 evaluation (Caselli et al., 2014). Participants will be asked to identify all event mentions according to the EVENTI annotation guidelines. The set of event attributes which will be evaluated is extended as described in the Main Task. The main innovation with respect to the original task is that participants will be provided with dependency parsing data both in training and test. Systems will be ranked according to the attribute CLASS F1 score.

3.1 Annotation and Data Format

In the spirit of re-using available datasets, the annotation efforts for the Parsing Events task will be mainly devoted to the creation of a new test set, called Platinum data, which will contain manual annotation for both dependency parsing and events. The size of the Platinum data will be around 10k-20k tokens. The annotation of the dataset will be conducted by applying the ISDT guidelines for the dependency parsing information and the EVENTI guidelines for events. An innovative aspect of the Platinum data concerns the text genres. To provide a more reliable evaluation, the Platinum data will consist of newspaper articles and biographies from Wikipedia[7].

The training data (Gold data) will be based on the EVENTI and the Dependency Parsing data. A subset of 27,597 tokens between the two datasets perfectly overlaps, thus making already available Gold annotations. Given that the focus of the evaluation exercise is on the reciprocal influence of the two basic tasks, we will provide the missing annotations on the remaining parts (i.e. 102,682 tokens for the EVENTI dataset and 160,398 tokens for the Dependency Parsing dataset) by means of automatically generated annotation, i.e. Silver data. Silver data have already been successfully used to extend the size of training data in previous evaluation exercises (e.g. TempEval-3). Furthermore, we plan to extend the set of overlapping Gold data by manual revision.

Training data will be distributed in a unified representation format compliant with the CoNLL-X specifications (Buchholz and Marsi, 2006) and extended for the encoding of event information which will be annotated in terms of standard IOB representation as exemplified in Figure 1 (the example is taken from the overlapping portion of the training data of the two task at EVALITA 2014). Event annotation (last seven columns) is concerned with the following information types: event extent, class, tense, aspect, vform, mood and polarity.

The test set for the main task will be distributed in the same format of the training dataset providing participants with pre-tokenized, POS-tagged and lemmatized data. This distribution format will be adopted also for the two subtasks. In addition to the information regarding tokens, POS tags and lemmas, Gold data for events will be available for the dependency parsing subtask, while Gold data for dependency parsing will be available for the event extraction subtask.

Systems will be required to produce a tab-delimited file. Systems participating to the main task will provide in output the extended CoNLL-X format including the information for the event an-

[6] http://medialab.di.unipi.it/wiki/ISDT

[7] The biographical data are part of the multilingual parallel section (Italian / English) of TUT (ParTUT http://www.di.unito.it/~tutreeb/partut.html).

1	Sulle	su	E	EA	num=p|gen=f	10	prep	_	_	_	_	_	_	_	_	_	_	_	
2	cause	causa	S	S	num=p|gen=f	1	pobj	_	_	B-EVENT	B-OCCURRENCE	B-NONE	B-NONE	B-NONE	B-NONE	B-POS			
3	dell'	di	E	EA	num=s|gen=n	2	prep	_	_	_	_	_	_	_	_	_	_	_	
4	incidente		incidente		S	S	num=s|gen=m	3	pobj	_	_	_B-EVENT		B-OCCURRENCE		B-NONE	B-NONE	B-NONE	B-NONE	B-POS
5	la	il	R	RD	num=s|gen=f	6	det	_	_	_	_	_	_	_	_	_				
6	Procura	Procura	S	SP	_	10	nsubj	_	_	_	_	_	_	_	_	_				
7	di	di	E	E	_	6	prep	_	_	_	_	_	_	_	_	_				
8	Marsala	Marsala	S	SP	_	7	pobj	_	_	_	_	_	_	_	_	_				
9	ha	avere	V	VA	num=s|per=3|mod=i|ten=p	10	aux	_	_	_	_	_	_	_	_	_				
10	aperto	aprire	V	V	num=s|mod=p|gen=m	0	ROOT	_	_	B-EVENT	B-ASPECTUAL		B-PRESENT		B-PERFECTIVE		B-NONE	B-NONE	B-POS
11	un'	uno	R	RI	num=s|gen=f	12	det	_	_	_	_	_	_	_	_	_				
12	inchiesta		inchiesta		S	S	num=s|gen=f	10	dobj	_	_	B-EVENT	B-OCCURRENCE		B-NONE	B-NONE	B-NONE	B-NONE	B-NONE	B-POS
13	.	.	F	FS	_	10	punct	_	_	_	_	_	_	_	_	_				

Figure 1: Example of a complete annotated sentence with syntactic and event information.

notation as shown in Figure 1. Systems taking part to the individual subtasks will provide in output the relevant fields: head token id, and the dependency linking the token under description to its head, for the dependency parsing subtask; the event extent and associated attributes for the event extraction subtask.

4 Evaluation and Discussion

Evaluation of systems is not a trivial issue. For the evaluation of participating systems we foresee at the moment different evaluation metrics for each task, described below.

Main Task: The main task aims at evaluating the bi-directional influence of syntactic and semantic information. We are then proposing a hybrid measure which takes into account the correctness of the event class and that of the dependency label. We propose the following definitions of Precision, Recall, and F1:

- Precision: the ratio between the tokens with correct event class and labeled dependency from the system, tp_i, and all tokens marked as event by the system (tp_i and fp_i): $\frac{tp_i}{tp_i+fp_i}$;
- Recall: the tokens with correct event class and labeled dependency from the system, tp_i, and the number of positive examples in the Gold data (tp_i plus false negatives fn_i) : $\frac{tp_i}{tp_i+fn_i}$
- F1: the mean of Precision and Recall calculated as follows: $\frac{2 Precision Recall}{Precision+Recall}$

Subtask A: Similarly to the dependency parsing task presented in EVALITA 2014, in addition to the standard accuracy dependency parsing measures, i.e. Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS), we will provide an alternative and semantically-oriented metric to assess the ability of the parsers to produce reliable and accurate output for Information Extraction applications. As in EVALITA 2014, we will select a set of dependency relations and for these relations the parser accuracy will be evaluated using Precision, the ratio of correct relations extracted over the total of extracted relations; Recall, the ratio of correct relations extracted over the relations to be found (according to the gold standard); and F-Measure. Differently from EVALITA 2014, for this semantically-oriented evaluation we will focus on dependency relations involved in the syntax of event structures.

Subtask B: Following the EVENTI evaluation exercise, the Event Extraction subtask will be evaluated by applying the adapted TempEval-3 scorer (UzZaman et al., 2013; Caselli et al., 2014). We will evaluate i.) the number of the elements correctly identified and if their extension is correct, and ii.) the attribute values correctly identified. As for the first aspect, we will apply standard Precision, Recall and F1 scores. Strict and relaxed (or partial) match will be taken into account. On the other hand, attribute evaluation will be computed by means of the attribute F1 score (UzZaman et al., 2013), which measures how well a system identified the element and corresponding attributes' values.

For the evaluation of subtask results, participants will be asked to submit different runs, carried out with and without the information from the other subtask: i.e. Dependency Parsing will be carried out with and without event information, and Event Extraction will be carried out with and without dependency information. This type of contrastive evaluation highlights one of the main novelties of the proposed complex task, which is not only aimed at assessing the performance of participating systems and ranking achieved results, but also at investigating impact and role of different types of information on each task depending on the adopted algorithm. A shared task organized along these lines thus creates the prerequisites for a more accurate error analysis and will possibly open up new directions of research in tackling old challenges.

References

Giuseppe Attardi, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli. 2015. State of the art language technologies for Italian: The EVALITA 2014 perspective. *Intelligenza Artificiale*, 9(1):43–61.

Roberto Basili, Diego De Cao, Alessandro Lenci, Alessandro Moschitti, and Giulia Venturi. 2013. EVALITA 2011: The frame labeling over Italian texts task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, Lecture Notes in Computer Science, pages 195–204. Springer Berlin Heidelberg.

Johan Bos, Fabio Massimo Zanzotto, and Marco Pennacchiotti. 2009. Textual entailment at EVALITA 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*.

Cristina Bosco and Alessandro Mazzei. 2013. The EVALITA dependency parsing task: from 2007 to 2011. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, Lecture Notes in Computer Science, pages 1–12. Springer Berlin Heidelberg.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69.

Cristina Bosco, Felice DellOrletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. In *Proceedings of the Fourth International Workshop EVALITA 2014*, pages 1–8.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Tommaso Caselli and Rachele Sprugnoli. 2014. EVENTI Annotation Guidelines for Italian v.1.0. Technical report, FBK and TrentoRISE.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI. EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*, pages 27–34.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 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.

James Henderson, Paola Merlo, Ivan Titov, and Gabriele Musillo. 2013. Multilingual joint parsing of syntactic and semantic dependencies with a latent variable model. *Computational Linguistics*, 39(4):949–998.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Rubén Urizar. 2015. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan T. McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 915–932.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Yi Zhang. 2014. Semeval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinkova, Dan Flickinger, Jan Hajic, and Zdenka Uresova. 2015. Semeval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, Colorado, June. Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Association for Computational Linguistics.

Nashaud UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9. Association for Computational Linguistics, Atlanta, Georgia, USA.

Deniz Yuret, Aydin Han, and Zehra Turgut. 2010. Semeval-2010 task 12: Parser evaluation using textual entailments. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 51–56. Association for Computational Linguistics.

Generalization in Native Language Identification: Learners versus Scientists

Sabrina Stehwien, Sebastian Padó
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{sabrina.stehwien,sebastian.pado}@ims.uni-stuttgart.de

Abstract

English. Native Language Identification (NLI) is the task of recognizing an author's native language from text in another language. In this paper, we consider three English learner corpora and one new, presumably more difficult, scientific corpus. We find that the scientific corpus is only about as hard to model as a less-controlled learner corpus, but cannot profit as much from corpus combination via domain adaptation. We show that this is related to an inherent *topic bias* in the scientific corpus: researchers from different countries tend to work on different topics.

Italiano. *La Native Language Identification (NLI) permette di riconoscere la lingua madre di un autore utilizzando il testo scritto in un' altra lingua. In questo lavoro utilizziamo tre collezioni di testi prodotti da apprendenti di inglese e un nuovo corpus scientifico, presumibilmente più difficile. In realtà, il corpus scientifico risulta essere difficile da modellare quanto un corpus di apprendimento meno controllato; tuttavia, a differenza di questi, esso non beneficia della combinazione di diversi corpora con metodi di domain adaptation. Questo limite è legato ad un'intrinseca specializzazione degli argomenti del corpus scientifico: ricercatori di paesi diversi tendono a trattare argomenti diversi.*

1 Introduction

Native Language Identification (NLI) is the task of recognizing an author's native language (L1) from text written in a second language (L2). NLI is important for applications such as the detection of phishing attacks (Estival et al., 2007) or data collection for the study of L2 acquisition (Odlin, 1989). State-of-the-art methods couch NLI as a classification task, where the classes are the L1 of the author and the features are supposed to model the effects of the author's L1 on L2 (*language transfer*). Such features may be of varying linguistic sophistication, from function words and structural features (Tetreault et al., 2012) on one side to N-grams over characters, words and POS tags (Brooke and Hirst, 2011; Bykh and Meurers, 2012) on the other side.

Like in many NLP tasks, there are few large datasets for NLI. Furthermore, it is often unclear how well the models really capture the desired language transfer properties rather than *topics*. The widely-used International Corpus of Learner English (ICLE, Granger et al. (2009)) has been claimed to suffer from a *topic bias* (Brooke and Hirst, 2011): Authors with the same L1 prefer certain topics, potentially due to the corpus collection strategy (from a small set of language courses). As a result, Brooke and Hirst (2013) question the generalization of NLI models to other corpora and propose the use of domain adaptation. In contrast, Bykh and Meurers (2012) report their ICLE-trained models to perform well on other learner corpora.

This paper extends the focus to a novel corpus type, non-native scientific texts from the ACL Anthology. These are substantially different from learner corpora: (a) most authors have a good working knowledge of English; and (b) due to the conventions of the domain, terminology and structure are highly standardized (Hyland, 2009; Teufel and Moens, 2002). A priori, we would believe that NLI on the ACL corpus is substantially more difficult.

Our results show, however, that the differences between the ACL corpus and the various learner corpora are more subtle: The ACL corpus is about as difficult to model as some learner corpora. However, generalization to the ACL corpus is more difficult, due to its idiosyncratic topic biases.

Corpus	# Docs/L1	Avg # Tokens/Doc	Type
TOEFL11	1100	348	Learner
ICLE	251	612	Learner
Lang-8	176	731	Learner
ACL	54	3850	Science

Table 1: Statistics on datasets

2 Datasets

We consider three learner corpora plus one scientific corpus, described below. We consider the 7 languages that are in the intersection of all datasets (DE, ES, FR, IT, JP, TR, ZH). To obtain a balanced setup comparable across corpora, we determined for each corpus the language with fewest documents, and randomly sampled that number of documents from the other languages (cf. Table 1).

TOEFL11. The TOEFL11 corpus (Blanchard et al., 2013) consists of texts that learners of English with mixed proficiency wrote in response to prompts during TOEFL exams.

ICLE. is the oldest and best-researched NLI corpus, a collection of essays written by students with a high intermediate to advanced level of English.

Lang-8. Lang8, introduced in (Brooke and Hirst, 2011) is a web-scraped version of the Lang-8 website[1] where learners of English post texts for correction by native speakers. Although it counts as a learner corpus, it is much less controlled.

ACL. Adapting a method proposed by Lamkiewicz (2014), we extracted a dataset for NLI from the 2012 release of the ACL Anthology Network Corpus (Radev et al., 2013), covering 25,000 papers from the Proceedings of ACL and other ACL-sponsored conferences and workshops. The dataset was extracted according to the e-mail domains of the authors, which were assumed to correspond to their native countries.[2] A document was included if and only if all the e-mail addresses had the same domain. Furthermore, we removed all the headers, acknowledgments and reference sections, since these often contain information on the authors' home country or L1.[3]

[1] `http://www.cs.toronto.edu/~jbrooke/Lang8.zip`

[2] While this heuristic would fail for countries with a high influx of foreign researchers, like the US, it seems reasonable for countries like Turkey and Japan. Manual evaluation of a small sample showed its precision to be >95%.

[3] The data is available at `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/NLI2015.html`

3 Models

Our NLI models uses binary features consisting of recurring unigrams and bigrams as proposed by Bykh and Meurers (2012).[4] An n-gram is recurring if it occurs in more than two documents of the same class. As multi-class classifier, we use the LIBLINEAR Support Vector Machine implementation (Chang et al., 2008) with default parameters.

Our standard models are simply trained on one corpus. However, since our focus will be on cross-corpus experimentation, we directly describe the two domain adaptation methods with which we will experiment to improve generalization. We will use the standard terminology of *source* for the (main) training domain and *target* for the testing domain.

Feature Augmentation. Daumé III's (2007) simple but effective domain adaptation method augments the feature space of the problem and can be applied as a preprocessing step to any learning algorithm. It allows feature weights to be learned per domain, by mapping input feature vector onto triplicate version of themselves. The first version of each feature, the "general" version, is identical to the original feature. The second version, the "source" version, is identical to the general version for all instances from the source domain, and zero for all instances from the target domain; vice versa for the third version, the "target" version.

Marginalized Stacked Denoising Autoencoders. Glorot et al. (2011) propose Stacked Denoising Autoencoders (SDAs) for domain adaptation, multilayer networks that reconstruct input data from a corrupted input by learning intermediate (hidden) layers. The intuition is that the intermediate layers model the relevant properties of the input data without overfitting, providing robust features that generalize well across domains. Chen et al. (2012) propose a marginalized SDA (mSDA) which makes the model more efficient while preserving accuracy.

Formally, the input data $\mathbf{x}$ is partially and randomly corrupted into $\tilde{\mathbf{x}}$, e.g., by setting some values to zero. The autoencoder leans a hidden representation from which x is reconstructed: $g(h(\tilde{\mathbf{x}})) \approx \mathbf{x}$. The objective is to minimize the reconstruction error $\ell(\mathbf{x}, g(h(\tilde{\mathbf{x}})))$. We set the corruption probability $p = 0.9$ and the number of layers $l = 1$ in line with previous work. If there are many more features than data points, Chen et al. (2012) use the

[4] We refrain from using structural features, concentrating on model generalization when using simple lexical features.

Name	Training Data	Test Data
SRC-only	TOEFL11	ICLE Lang-8 ACL
TGT-only	ICLE (2/3) Lang-8 (2/3) ACL (2/3)	ICLE Lang-8 ACL
CONCAT / FA / mSDA -big	TOEFL11 + ICLE (2/3) TOEFL11 + Lang-8 (2/3) TOEFL11 + ACL (2/3)	ICLE Lang-8 ACL
CONCAT / FA / mSDA -small	TOEFL11 + ICLE (1/3) TOEFL11 + Lang-8 (1/3) TOEFL11 + ACL (1/3)	ICLE Lang-8 ACL

Table 2: Model configurations

x most frequent features. The data D is sliced into $\frac{D}{x} = y$ partitions and mSDA is performed on each partition y_i by decoding $g(h(\mathbf{y}_i)) \approx \mathbf{x}$. We set x to 5000 and concatenate the learned intermediate layer units with the original features.

4 Experiments and Results

4.1 Experimental Setup

Table 2 shows all model configurations that we consider. In the SRC-only model, we use the full TOEFL-11 – our largest corpus – as training corpus and test on the other three corpora. The in-domain model (TGT-only), trains and tests always on the same corpora. The next set of models (CONCAT-big, FA-big, mSDA-big) all combine TOEFL-11 as source corpus with two thirds of a target domain corpus, using different combination methods (plain concatenation or the two domain adaptation methods). The final set of models is parallel the previous set, but uses just one third of the target corpora, to assess the influence of the amount of training data.

In all cases except SRC-only, we perform 3-fold cross-validation. We report accuracy, and test statistical significance using the Chi-squared test with Yates' continuity correction (Yates, 1984). Due to the balanced nature of our corpora, the frequency (and random) baselines are at 1/7 = 14.3%.

4.2 Main Experimental Results

The main results are shown in Table 3. The SRC-only results show that the only corpus for which an NLI model trained on TOEFL performs reasonably well is, unsurprisingly, its "nearest neighbor" ICLE, while performance on Lang-8 and ACL is poor. However, even on ICLE, performance remains below 80%. In contrast, the TGT-only results show that reasonable NLI results (generally >80%) can be obtained for each domain if there is target data to train on. Notably, the ACL corpus is easier to model than the Lang-8 corpus despite its special status as a scientific corpus and despite its much smaller size. Yet, the results (except for the very easy ICLE) generally leave room for improvement.

Model \ Test data	ICLE	Lang-8	ACL
SRC-only	79.5	57.7	49.5
TGT-only	96.1	77.1	**85.7**
CONCAT-big	94.4	80.0	75.1
FA-big	97.0***	84.1*	81.2
mSDA-big	**98.9*****	**90.0*****	**88.4****
CONCAT-small	92.5	75.5	68.8
FA-small	96.0***	77.9	74.6
mSDA-small	**98.6*****	86.8***	**86.0*****

Table 3: Classification accuracies. Bold indicates results not significantly different from best result for each test set ($p<0.05$). Significant improvements over results in previous row marked by asterisks (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

SRC plus a lot of TGT data. The next group of results (*-big*) shows what happens when the SRC data and all available TGT data are combined. This experiment establishes an upper bound of performance for when a lot of target domain data is available. Simple CONCATentation does not perform well, with degradation compared to TGT-only for ICLE and, with a notably large slump, ACL. Feature Augmentation works to some extent, but mSDA improves the results much more substantially, to almost 90% accuracy and above, and yielding the largest improvements over TGT-only (ICLE: +2.8%, Lang-8: +12.9%, ACL:+2.7%). We surmise that FA is handicapped by the relatively small sizes of Lang-8, and the very small size of the ACL corpus, which are "overpowered" by the large TOEFL-11 dataset.

SRC plus some TGT data. The final group of results (*-small*) shows the results of combining the SRC data with only half the available TGT data. In comparison to *-big*, the performance drops substantially for CONCAT and FA, but only somewhat for mSDA (ICLE: -0.3%, Lang-8: -3.2%, ACL: -2.4%; difference statistically significant only for Lang-8). This indicates that mSDA can take advantage of relatively small target domain datasets.

Summary. Domain adaptation, in particular mSDA, can construct highly accurate NLI models (85%+) by combining large source datasets with relatively small target datasets. Contrary to

Test Corpus	ICLE	Lang-8	ACL
SRC-only	76.7	54.5	49.5
TGT-only	96.0	74.8	84.9
mSDA-big	98.7	88.2	86.5

Table 4: Accuracies on reduced feature set

expectations, we do not see a clear division between learner and scientific datasets: rather, the less well controlled Lang-8 behaves much more like the ACL dataset than like ICLE, which in turns clusters together with the TOEFL-11 dataset, the other "classical" learner corpus. This explains the good generalization results found by Bykh and Meurers (2012) but indicates that they may be restricted to "classical" learner corpora.

A difference between Lang-8 and ACL, however, is that Lang-8 still profits significantly from domain adaptation while ACL does not. There is a numerical increase, though, so the low number of documents in the ACL dataset may be responsible.

4.3 Topic vs. L1 Transfer at the Feature Level

To better understand the models, we inspect the most highly weighted n-gram features in the NLI models. As expected, in all models we find language and country names which directly indicate the authors' L1 topically (*"I am from China"*), as opposed to language transfer. To test the importance of these features, we construct a stop word list including the relevant language and country names for each language (e.g. *Italian, Italy* for IT), including *Hong, Kong* for ZH. We use this list to filter out all features that include these stop words.

The results for the reduced feature set are shown in Table 4. They do not differ substantially from our previous experiments. Thus, simple country and language mentions do not seem to have a huge impact on NLI. While this does not exclude the possibility of topic effects among less prominent features, many of the features acquired from the learner corpora that received the most weight are actually interpretable in terms of language transfer, thus exposing writing habits that point towards the author's L1. For example, the FR and ES models include misspellings of loanwords (*"exemple"*, *"advertissements"*, *"necesary"*, *"diferent"*) while DE authors are influenced by German punctuation rules for embedded clauses ("*, that*", "*, because*"). We also see lexical transfer expressed as the overuse of words that are more frequent in the L1 ("*concern*" for FR). What is notable in the ICLE corpus are L1-specific register differences that were found to correlate with topics by Brooke and Hirst (2011): JP writers prefer a colloquial style ("*I think*", "*need to*") while FR writers adopt a more formal style ("*may*", "*the contrary*", "*certainly*").

The situation is quite different in the ACL corpus. While we still find mentions of languages (*"of Chinese", "the German"*), many n-grams reflect scientific jargon and preferred research *topics*. For example, TR researchers write about morphology (*"suffixes", "inflectional", "morphological"*) and ES authors discuss Machine Learning (*"stored", "trained", "the system"*). For some languages, the features appear to be a mixture of specific topics and language transfer: for IT, we find *"category", "implement", "availability", "we obtain", "results in", "accounts for"*. Are these indicative of empirical methodology, or merely results of the (over)use of particular collocations? While we cannot answer this at the moment, we believe that the ACL corpus can thus be considered to have an idiosyncratic form of topic bias – but one that is very different from the learner corpora, which explains the difficulty of generalizing to ACL.

5 Conclusion

This study investigated the generalizability of NLI models across learner corpora and a novel corpus of scientific ACL documents. We found that generalizability is directly tied to corpus properties: well-controlled learner corpora (TOEFL-11, ICLE) generalize well to one another (Bykh and Meurers, 2012). Together with the minor effect on performance of removing topic-related features, we conclude that topic bias *within a similar text type* does not greatly affect generalization.

At the same time, "classical" learner corpora do not generalize well to less-controlled learner corpora (Lang-8) or scientific corpora (ACL). Lang-8 and ACL show comparable performance, which seems surprising given the small size of the ACL corpus and its quite different nature. Our analysis shows that the ACL corpus exhibits an idiosyncratic topic bias: scientists from different countries work on different topics, which is reflected in the models. As a result, the improvements that Lang-8 can derive from domain adaptation techniques carry over to the ACL corpus only to a limited extent. Nevertheless, the use of mSDA can halve the amount of ACL data necessary for the same performance, which is a promising result regarding the generalization to other low-resource domains.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Marin Chodorow. 2013. TOEFL11: A Corpus of Non-Native English. *ETS Research Report Series*, 2013(2):1–15.

Julian Brooke and Graeme Hirst. 2011. Native Language Detection with 'Cheap' Learner Corpora. In *Proceedings of the 2011 Conference on Learner Corpus Research*, Louvain-la-Neuve, Belgium.

Julian Brooke and Graeme Hirst. 2013. Using Other Learner Corpora in the 2013 Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–196, Atlanta, Georgia.

Serhiy Bykh and Detmar Meurers. 2012. Native Language Identification Using Recurring N-Grams – Investigating Abstraction and Domain Dependence. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 425–440, Mumbai, India.

Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin, and Soeren Sonnenburg. 2008. Liblinear: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Minmin Chen, Zhixiang (Eddie) Xu, and Kilian Q. Weinberger. 2012. Marginalized Denoising Autoencoders for Domain Adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of the 28th International Conference on Machine Learning*, volume 27, pages 97–110.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English, Version 2*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Ken Hyland. 2009. *Academic Discourse*. Continuum, London.

Anna Maria Lamkiewicz. 2014. Automatische Erkennung der Muttersprache von L2-Englisch-Autoren. Magisterarbeit, Institut für Computerlinguistik, Neuphilologische Fakultät, Ruprecht-Karls-Universität Heidelberg.

Terence Odlin. 1989. *Language Transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2585–2602, Mumbai, India.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4).

Frank Yates. 1984. Tests of Significance for 2x2 Contigency Tables. *Journal of the Royal Statistical Society Series A*, 147 (3):426–463.

Sentiment Polarity Classification with Low-level Discourse-based Features

Evgeny A. Stepanov, Giuseppe Riccardi
Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Trento, TN, Italy
{evgeny.stepanov,giuseppe.riccardi}@unitn.it

Abstract

English. The poor state-of-the-art performances of discourse parsers prevent their application to downstream tasks. However, discourse parsing sub-tasks such as the detection of connectives and their sense classification have achieved satisfactory level of performance. In this paper we investigate the relevance of discourse connective features for tasks such as sentiment polarity classification. In the literature, discourse connectives are usually considered as modifiers of a polarity of a sentence or a word. In this paper we present experiments on using automatically extracted connectives and their senses as low-level features and as an approximation of a discourse structure for polarity classification of reviews. We demonstrate that, despite insignificant contributions to bag-of-words, the discourse-only models perform significantly above chance level.

Italiano. *Lo stato dell'arte degli analizzatori automatici del discorso impediscono la loro adozione nei contesti applicativi. Tuttavia, i sotto-processi automatici di analisi del discorso quali l'denticazione dei connettivi e la classicazione della loro etichetta semantica hanno comunque raggiunto un livello di prestazioni soddisfacente. In questo documento indaghiamo la rilevanza dei connettivi del discorso per i task come la classificazione della polarità dei sentimenti. In letteratura i connettivi del discorso sono comunemente considerati come modificatori della polarità di una frase o di una parola. In questo documento presentiamo alcuni esperimenti sull'estrazione automatica di connettivi, e relativi significati, e del loro utilizzo come caratteristiche di basso livello e come approssimazione della struttura di un discorso al fine di permettere la classificazione della polarità nelle recensioni. I connettivi del discorso assieme ai modelli bag-of-words permettono di ottenere risultati allo stato dell'arte e molto al di sopra dei modelli di base.*

1 Introduction

Discourse analysis has applications in many Natural Language Processing tasks; Webber et al. (2011) and Taboada and Mann (2006) among others list opinion mining, summarization, information extraction, essay scoring, etc. Availability of large discourse annotated resources such as Penn Discourse Treebank (PDTB) (Prasad et al., 2008a) and Rhetorical Structure Theory - Discourse Treebank (RST-DT) (Carlson et al., 2002) made it possible to develop statistical discourse parsers (e.g. (Marcu, 2000; Lin et al., 2014; Ghosh et al., 2011; Stepanov and Riccardi, 2013)). However, independent of the theory (RST or PDTB) the problem of end-to-end discourse parsing is far from being solved; thus, downstream application of these parsers yields mixed results.

In this paper we focus on PDTB approach to discourse parsing, which can be roughly partitioned into detection of discourse relations, extractions of their argument spans and sense classification. In CoNLL 2015 Shared Task on Shallow Discourse Parsing (Xue et al., 2015) the best system (Wang and Lan, 2015) achieved F_1 of 24 on the end-to-end parsing on a blind test set using strict evaluation that required exact match of all the spans and labels. Having such low end-to-end performances makes it difficult to apply PDTB-style discourse parsing to other NLP tasks. However, if we consider discourse parsing tasks individually, detection of discourse connectives and their classi-

Class	Type	Sub-Type
Comparison	*Contrast*	–
	Concession	–
Contingency	*Cause*	Reason Result
	Condition	–
Expansion	*Conjunction*	–
	Instantiation	–
	Restatement	–
	Alternative	– Chosen Alternative
	Exception	–
Temporal	*Synchronous*	–
	Asynchronous	Precedence Succession

Table 1: Simplified PDTB discourse relation sense hierarchy from CoNLL 2015 Shared Task.

fication into senses achieve high results: ≈ 90 for discourse connective detection and similarly ≈ 90 for connective sense classification (Stepanov et al., 2015). Thus, the output of these tasks could be used in other NLP applications.

Discourse connectives are essentially function words and phrases. Function word frequencies is a popular feature in NLP tasks such as authorship detection (Kestemont, 2014), and it has also been applied to sentiment polarity classification (Abbasi et al., 2008). Resolving connective usage and sense ambiguities (Section 2), they are potentially able to provide more refined features than simple function word counts. On the other hand, grouping connectives with respect to their senses yields more coarse features. In this paper we explore the utility of these features for sentiment polarity classification of movie reviews (Pang and Lee, 2004).

2 Discourse Connectives and Their Senses

In PDTB discourse relations are annotated using 3-level hierarchy of senses. The top level (level 1) senses are the most general: **Expansion**: one clause elaborates on the information given in another (e.g. 'and', 'in addition'); **Comparison**: there is a comparison or contrast between two clauses (e.g. 'but'); **Contingency**: there is a causal relationship between clauses (e.g. 'because'); and **Temporal**: two clauses are connected time-wise (e.g. 'before').

A relation signaled by a discourse connective is an *explicit* discourse relation. *Implicit* discourse relations between text segments (usually sentences), on the other hand, are inferred. The two classes are almost equally represented (53% vs. 47%). While detection of senses of *implicit* discourse relations is a hard problem (Lin et al., 2009; Xue et al., 2015); presence of a discourse connective in a sentence is sufficient for detection and classification of *explicit* discourse relations.

There are two levels of ambiguity present for a connective (Pitler and Nenkova, 2009): (1) it might be used to connect discourse units, or coordinate smaller constituents (e.g. 'and'); (2) some connectives might have different senses depending on usage (e.g. 'since' might signal causation or temporal relation). AddDiscourse tool was developed by (Pitler and Nenkova, 2009) to resolve these ambiguities. While using just connectives the 4-way sense classification accuracy of the tool is 93.67%, incorporating syntactic features raises performance to 94.15%; which is as good as the inter-annotator agreement on the same data (PDTB corpus - 94.00% (Prasad et al., 2008b)). Classification of discourse connectives into full depth of sense hierarchy also has an acceptable level of performance: 89.68% on PDTB development set of CoNLL 2015 Shared Task (Stepanov et al., 2015). For the Shared Task some senses were merged, and partial senses were disallowed (Xue et al., 2015); as a result, there are only 14 senses listed in Table 1. We classify discourse connectives identified by the addDiscourse tool further into this simplified hierarchy of senses.

3 Methodology

We test the utility of discourse connectives and their senses on sentiment polarity classification task. We follow the supervised machine approach and use SVM^{light} (Joachims, 1999) classifier with default parameter settings. A document is represented as a boolean vector of features (i.e. presence) and discourse-based features are added through vector fusion. Through out experiments 10-fold cross-validation is used, and results are reported as average accuracy, which is equivalent to micro- precision, recall, and F_1 for a binary classification where both classes are of interest.

3.1 Data Set

For the experiments we use the polarity dataset (v. 2.0) of (Pang and Lee, 2004), also known as Movie Reviews Data Set. The Data Set consists of 1,000 negative and 1,000 positive reviews extracted from the Internet Movie Database (IMDb).

3.2 Baseline Results

Using the 10-fold cross-validation split of (Pang and Lee, 2004), SVM unigram model achieves 86.25% average accuracy. Unlike the original paper, data set is used *as is*: no additional preprocessing such as frequency cut-off or prefixing the tokens following 'not', 'isn't', etc. till the first punctuation with 'NOT_' (Das and Chen, 2001) was used (same as (Stepanov and Riccardi, 2011)).

3.3 Representation of Discourse Connectives as Features

Function words were already used as features for polarity classification in (Abbasi et al., 2008), and the authors report that function words 'no' and 'if' tend to occur more frequently in negative reviews. Thus we experiment considering presence of connectives and their raw and normalized frequencies. Discourse connectives contain multi-word expressions (e.g. 'in_addition', 'on_the_other_hand', etc.), long-distance connective pairs (e.g. 'if_then', 'either_or'), and open class words (e.g. adverbs 'finally', 'similarly', etc.); and they are all treated as a single token.

Under these settings, we explore both the refinement and the generalization scenarios. In the refinement scenario discourse connective surface forms are appended with automatic Class (most general sense) or Sense decisions. and in the generalization scenario Class and Sense decisions replace the connective surface string. Consequently, we have 5 conditions, ordered from general to specific:

- **Class:** Class of a connective (one of 'Expansion', 'Comparison', 'Contingency', or 'Temporal');
- **Sense:** Sense of a connective from Table 1 (e.g. 'Temporal.Synchronous');
- **Surface:** Connective tokens (e.g. 'as');
- **Surface/Class:** Surface and Class tuple of a connective (e.g. 'as-Temporal' or 'as-Contingency');
- **Surface/Sense:** Surface and Sense tuple of a connective (e.g. 'as-Temporal.Synchronous' or 'as-Contingency.Cause.Reason');

In the following sections we evaluate these representations in isolation and fused into bag-of-words vectors.

Feature	B	R	N
BL: Chance	51.05		
Class	52.60	59.77	58.55
Sense	56.00	59.15	59.25
Surface	61.65	63.40	63.00
Surface/Class	61.35	**64.00**	63.15
Surface/Sense	61.20	63.65	62.70

Table 2: 10-fold cross-validation average accuracies for discourse connective as stand-alone features in comparison to the chance level baseline (*BL: Chance*). Results are reported for presence (B), raw (R) and normalized frequencies (N).

Additionally, our goal is to explore whether senses of explicit discourse relations alone can capture low-level discourse structure; and whether this low-level structure is beneficial for sentiment polarity classification. In order to approximate this, we use bigrams and trigrams of identified Classes and Senses. We introduce beginning and end of document tags to capture document initial and document final explicit relations. In this setting the presence of n-grams is considered, rather than the frequency. The setting is also evaluated in isolation and in fusion with bag-of-words.

4 Experiments and Results

In this section we present sentiment polarity classification experiments using discourse connective features under the settings defined in Section 3: (1) presence and frequencies as stand-alone features, (2) their effect on the bag-of-word model through vector fusion, and (3) effect of n-grams of Class and Senses in stand-alone and fusion settings.

4.1 Discourse Connectives as Stand-Alone Features

Table 2 presents the results of the experiments using discourse connectives as the only features. All the models, except Class presence (**B**), perform significantly above chance-level. Low performance of the Class-only model is expected, since there are only 4 Classes. As expected, the finer the features the better the performance. However, the Surface/Sense setting is lower than its more coarse version Surface/Class for all frequency count settings (statistically not significant). This is caused by Sense-level classifier's inferior performance, that often misses underrepresented senses of connectives.

Feature	B	R	N
BL: BoW	86.25		
Class	86.35	85.25	86.35
Sense	86.15	85.60	86.30
Surface	86.10	84.85	86.35
Surface/Class	85.95	84.90	86.35
Surface/Sense	85.85	84.90	86.35

Table 3: 10-fold cross-validation average accuracies for fusion of discourse connective features with bag-of-words (baseline: *BL: BoW*). Results are reported for presence (B), raw (R) and normalized frequencies (N).

The raw frequency counts perform better for all the representations, followed by normalized frequency counts. The boolean feature vector representation has the lowest performances. In the next Section we fuse these feature vectors with the boolean bag-of-words representation.

4.2 Fusion of Bag-of-Words and Discourse Connective Features

From the previous set of experiments we have observed that discourse connective only models perform above the chance level (even though much below the bag-of-words baseline). In order to investigate the effect of newly proposed discourse features, we fuse them with bag-of-words vectors (i.e. the baseline). The results of fusion are reported in Table 3.

From the results in Table 3 we can observe that the effect of feature fusion overall is insignificant. Raw frequency vectors generally have a negative effect on the performance. For boolean frequency vectors (i.e. presence), the more coarse features (Class and Sense) slightly improve the performance. For the normalized frequency count vectors, on the other hand, both more coarse and more refined features contribute to the performance. However, none of the improvements is statistically significant.

4.3 N-grams of Discourse Connective Senses

The results of experiments using n-grams of Class and Senses of connectives is reported in Table 4. The general observation is that increasing n-gram size has a positive effect on performance when discourse features are used stand-alone, and they are significantly above chance (except Class unigrams). The fusion of n-gram features and bag-of-word representation is also beneficial.

Feature	1	2	3
BL: Chance	51.05		
Class	52.60	58.35	59.55
Sense	56.00	57.40	58.80
BL: BoW	86.25		
BoW + Class	86.35	**86.85**	86.65
BoW + Sense	86.15	86.20	86.65

Table 4: 10-fold cross-validation average accuracies for discourse connective class and sense 1-3 grams and their fusion with bag-of-words. Only presence (boolean) of an n-gram is considered.

The best performing combination is fusion of bigrams of Classes and bag-of-words that achieves accuracy of 86.85. However, the improvement is statistically insignificant. But the fact that performances improve over the fusion of bag-of-words and frequency-based discourse connective vectors indicates that n-grams of explicit discourse relations are able to capture structures relevant for the sentiment polarity classification.

5 Conclusions

We have described experiments on using low-level discourse-based features for sentiment polarity classification. The general observations are (1) discourse connectives in isolation generally significantly outperform the chance baseline; and (2) using even the most general top-level senses provides performance gains. This is particularly notable due to the fact that discourse connective detection and relation sense classification do not generalize well across domains (Prasad et al., 2011).

Discourse connectives signal *explicit* discourse relations, which are only 53% of all discourse relations in PDTB. *Implicit* discourse relations (47%), which have the same senses, are much harder to deal with. Given the state of the art on *implicit* relation sense classification, detection and application of all the discourse relations is not yet possible. However, as indicated by the experiments on using n-grams of relation senses, even approximations can contribute.

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.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26(3):12:1–12:34, June.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank (RST-DT) LDC2002T07.

Sanjiv Das and Mike Chen. 2001. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001)*.

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2011. Shallow discourse parsing with conditional random fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*.

Thorsten Joachims. 1999. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

Mike Kestemont. 2014. Function words in authorship attribution: From black magic to theory? In *The 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL*, pages 59–66, Gothenburg, Sweden. ACL.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151 – 184.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26:395–448.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP Conference*, pages 13–16.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008a. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008b. The penn discourse treebank 2.0. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind K. Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12:188.

Evgeny A. Stepanov and Giuseppe Riccardi. 2011. Detecting general opinions from customer surveys. In *IEEE ICDM Workshops (ICDMW) - Sentiment Elicitation from Natural Text for Information Retrieval and Extraction Workshop (SENTIRE)*, pages 115–122, Vancouver, BC, December. IEEE.

Evgeny A. Stepanov and Giuseppe Riccardi. 2013. Comparative evaluation of argument extraction algorithms in discourse relation parsing. In *The 13th International Conference on Parsing Technologies (IWPT 2013)*, pages 36–44, Nara, Japan, November.

Evgeny A. Stepanov, Giuseppe Riccardi, and Ali Orkan Bayer. 2015. The UniTN discourse parser in CoNLL 2015 shared task: Token-level sequence labeling with argument-specific models. In *The 19th SIGNLL Conference on Computational Natural Language Learning (CoNLL) - Shared Task*, pages 25–31, Beijing, China, July. ACL.

Maite Taboada and William C. Mann. 2006. Applications of rhetorical structure theory. *Discourse Studies*, 8(4):567–88.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the The SIGNLL Conference on Computational Natural Language Learning*, Beijing, China, July. ACL.

Bonnie L. Webber, Markus Egg, and Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering*, pages 437–490.

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 *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.

Analyzing and annotating for sentiment analysis the socio-political debate on #labuonascuola

Marco Stranisci[1-2], Cristina Bosco[1], Viviana Patti[1], Delia Irazú Hernández Farías[1-3]
[1]Dipartimento di Informatica, Università di Torino
[2]Cooperativa weLaika, Torino
[3]Universitat Politecnica de Valencia
{bosco,patti}@di.unito.it, marco.stranisci@welaika.com,
dhernandez1@dsic.upv.es

Abstract

English. The paper describes a research about the socio-political debate on the reform of the education sector in Italy. It includes the development of an Italian dataset for sentiment analysis from two different comparable sources: Twitter and the online institutional platform implemented for supporting the debate. We describe the collection methodology, which is based on theoretical hypotheses about the communicative behavior of actors in the debate, the annotation scheme and the results of its application to the collected dataset. Finally, a comparative analysis of data is presented.

Italiano. *L'articolo descrive un progetto di ricerca sul dibattito socio-politico sulla riforma della scuola in Italia, che include lo sviluppo di un dataset per la sentiment analysis della lingua italiana estratto da due differenti fonti tra loro confrontabili: Twitter e la piattaforma istituzionale online implementata per supportare il dibattito. Viene evidenziata la metodologia utilizzata per la raccolta dei dati, basata su ipotesi teoriche circa le modalità di comunicazione in atto nel dibattito. Si descrive lo schema di annotazione, la sua applicazione ai dati raccolti, per concludere con un'analisi comparativa.*

1 Introduction

The widespread diffusion of social media in the last years led to a significant growth of interest in the field of opinion and sentiment analysis of user generated contents (Bing, 2012; Cambria et al., 2013). The first applications of these techniques were focusing on the users' reviews for commercial products and services (e.g. books, shoes, hotels and restaurants), but they quickly extended their scope to other interesting topics, like politics. Applications of sentiment analysis to politics can be mainly investigated under two perspectives: on one hand, many works focus on the possibility of predicting the election results through the analysis of the sentiment conveyed by data extracted from social media (Ceron et al., 2014; Tumasjan et al., 2011; Sang and Bos, 2012; Wang et al., 2012); on the other hand, the power of social media as "a trigger that can lead to administrative, political and societal changes" (Maynard and Funk, 2011) is also an interesting subject to investigate (Lai et al., 2015). This paper mainly focuses on the last perspective. Our aim is indeed the creation of a manual annotated corpus for sentiment analysis to investigate the dynamics of communication between politics and civil society as structured in Twitter and social media. We focused from the beginning our attention mainly on Twitter because of the relevance explicitly given to this media in the communication dynamics of the current government. In order to describe and model this communicative behavior of the government, we assume the theoretical framework known in literature as *framing*, which consists in making especially salient in communication some selected aspect of a perceived reality (Entman, 1993).

The data selected to create the corpus have been chosen by analyzing in Twitter and other contexts the diffusion of three hashtags, i.e. #labuonascuola, #italicum, #jobsact. In particular, we focus on #labuonascuola (*the good school*), which was coined to communicate the school reform proposed by the actual government.

A side effect of our work is the development of a new lexical resource for sentiment analysis in Italian, a currently under-resourced language. Among the existing resources let us mention Senti-TUT

(Bosco et al., 2013), which has been exploited together with the TWITA corpus (Basile and Nissim, 2013) for building the training and testing datasets in the SENTIment POLarity Classification shared task (Basile et al., 2014) recently proposed during the last edition of the evaluation campaign for Italian NLP tools and resources (Attardi et al., 2015). The Sentipolc's dataset includes tweets collected during the alternation between Berlusconi and Monti on the chair of Prime Minister of the Italian government. The current proposal aims at expanding the available Italian Twitter data annotated with sentiment labels on the topic of politics, and it is compatible with the existing datasets w.r.t. the annotation scheme and other features.

The paper is organized as follows. The next section describes the dataset mainly focusing on collection. In the section 3 we describe the annotation applied to the collected data and the annotation process. Section 4 concludes the paper with a discussion of the analysis applied to the dataset.

2 Data collection: criteria and subcorpora

In this section, we describe the methodology applied in collection, which depends on some assumption about the dynamics of the debate, and the features of the resulting dataset, which is organized in two different subcorpora: the Twitter dataset (TW-BS) and the dataset including the textual comments extracted from the online consultation about the reform (WEB-BS).

In order to describe the communicative behavior of the government, we assume, as a theoretical hypothesis, that the communication strategy acting in the debate can be usefully modeled by exploiting *frames*. In political communication, this cognitive strategy led to impose a narration to opponents (Conoscenti, 2011).
Following this hypothesis we can see that the Prime Minister and his staff coined two categories of frames by hashtagging in order to impose a narration to the public opinion: the first one aimed at legitimating the new born government and its novelty in the political arena (#lavoltabuona; #passodopopasso); the other one in order to create a general agreement on some proposal (#labuonascuola, #italicum, #jobsact). Each of these hashtags could be considered as an indicator of a frame created for elaborating a storytelling on the three most important reforms proposed by the government respectively on school, job and elections.
The observation of Twitter in this perspective led us to focus on messages featured by the presence of the three keywords #labuonascuola, #jobsact, #italicum, and posted from February 22th, 2014 (establishment of the new government) to December 31st, 2014. First, we collected all Italian tweets in this time slot (218,938,438 posts), then we filtered out them using the three hashtags. With 28,363 occurrences #labuonascuola, even if attested later than the others, is featured by the higher frequency, which occur respectively 27,320 (#jobsact) and 3,974 (#italicum) times. This prevalence is due not only to the general interest for the topic, but in particular to the activation by the government of an online consultation on school reform through the website `https://labuonascuola.gov.it`.

The first corpus we collected, WEB-BS henceforth, includes therefore texts from this online consultation[1]. We collected 4,129 messages composed by short texts posted in the consultation platform. All contents were manually tagged by authors with one among the 53 sub-topics labels made available, and organized by themselves in four categories: 'what I liked' (642), 'what I didn't like'(892), 'what is missing' (675) and 'new integration' (1,920). So, the label which conveys a positive opinion represented the 15.55% of the total. Otherwise, the negative label has been used the 21,60% of times. This manual classification in sub-topic and polarity categories of the messages, makes the WEB-BS dataset especially interesting, since the explicit tagging applied by the users can be in principle compared with the results of some automatic sentiment or topic detection engine. Moreover, let us observe that even if the WEB-BS corpus shares linguistic features with the corpus extracted from Twitter described below, it represents a different global context (Sperber and Wilson, 1986) (Yus, 2001) that orients, at the pragmatic level (Bazzanella, 2010), users in the expression of their opinions.

The second corpus we collected is composed of texts from Twitter focused on the debate on school (TW-BS henceforth), selected by filtering Twitter data exploiting the previously cited "fram-

[1]Users could participate to the consultation in different ways: as single users, filling out a survey, or as a group taking part to a debate about a particular topic or aspect of the reform.

ing" hashtags. We focused our attention on tweets posted from September 3rd, 2014 (when the consultation was launched by the government with a press conference) to November 15th, 2014. In addition to #labuonascuola, we used also keywords like 'la buona scuola', 'buona scuola', 'riforma scuola', 'riforma istruzione'. The resulting dataset is composed of 35,148 tweets, which was first reduced to 11,818 after removing retweets, and then to 8,594 after a manual revision devoted to further deletion of duplicates and partial duplicates. A quantitative analysis of the collected data shows us that 4,244 users contributed to the debate on Twitter. Among them, only 1,238 (29,2%) posted at least 2 messages and produced 5,588 tweets, 65% of the total. If we consider the hashtags' occurrences, #labuonascuola appears 5,346 times, while its parodic reprise is very infrequent: 108 total occurrences for three hashtags #lacattivascuola - *#thebadschool*, #lascuolaingiusta - *#theunfairschool*, and #labuonasola - *#thegoodswindle*.

3 Annotation and disagreement analysis

The annotation process involved 8 people with different background and skills, three males and five women. The task was marking each post with a polarity and one or more topic according to the set of tags described below.

For what concerns polarity, we assumed the same labels exploited in the Senti-TUT annotation schema: NEG for negative polarity, POS for positive, MIXED for positive and negative polarity both, NONE in the case of neutral polarity. Finally, we annotated irony, whose recognition is a very challenging task for the automatic detection of sentiment because the inferring process goes beyond syntax or semantics (Reyes et al., 2013; Reyes and Rosso, 2014; Maynard and Greenwood, 2014; Ghosh et al., 2015). As in Sentipolc (Basile et al., 2014), we were interested in annotating manually the polarity of the ironic tweets, where the presence of ironic devices can work as an unexpected "polarity reverser" (e.g. one says something "good to mean something "bad). So, we coined two labels: HUM NEG for tagging tweets ironic and negative, and conversely HUM POS for tagging the ones that were both positive and ironic. The set of labels was completed by a tag for marking unintelligible tweets (UN), one for duplicates (RT), and NP for texts about not related topic.

As far as topics are concerned, among the 53 categories used in the WEB-BS corpus, we selected the 13 most frequent, which occur 2,182 times in the consultation website: docenti - *teachers* , valutazione - *evaluation*, formazione - *training*, alternanza scuola/lavoro - *school-work*, investimenti - *investments*, reclutamento - *recruitment*, curricolo - *curriculum*, innovazione - *innovation*, lingue - *languages*, merito - *merit*, presidi - *headmasters*, studenti - *students*, and retribuzione - *remuneration*. Furthermore, we coined two more general labels for tweets addressing a sub-topic not present in categories, and for tweets just indirectly targeted to school reform.

In order to limit biases among annotators and to make well shared the meaning of all the labels to be annotated, we produced a document including guidelines for annotations, several examples of polarity-labeled tweets, three glossaries about the meaning of the topics- and some recurrent terms on the school reform.

The final dataset, manually annotated by two independent human annotators and cleaned from duplicates, not related, and unintelligible tweets, consists of 7,049 posts. 4,813 out of the total amount of annotated tweets, were tagged with the same label by both annotators. This is the current result for TW-BS; the label distribution is shown in 1. The inter-annotator agreement at this stage was $\kappa = 0.492$ (a moderate agreement). A qualitative analysis of disagreement (the 31.8% of the data) shows that the discrepancies very often depend on the presence of irony which has been detected only by one of the annotators even if both the humans performing the task detected the same polarity. This confirms the fact known in literature that irony is perceived in different ways and frequency by humans, as in the following example which showed a disagreement between annotators:

'Ho letto le 136 pagine della riforma della scuola,
finisce che i giovani si diplomano e vanno all'estero.
#labuonascuola'
'I read the 136 pages about the school reform,
it ends with youngs who graduate and go abroad.
#labuonascuola'

The remaining part of disagreeing annotations can be reported mainly as cases where one annotator detected a polarity and the other annotated the post as neutral. In order to extend the dataset, we are planning to apply a third indipendent annotation on the posts with disagreeing annotations.

4 Analysis of corpora

The analysis is centered on two main aspects of the annotation, i.e. polarities and topics, in the perspective of label frequency and relationships between labels and disagreement.

Table 1 shows the frequency of the labels exploited for polarity and a high frequency of the neutral label can be observed in this graphic. When discussing the guidelines for the application of labels, we decided to use the NONE label for marking all the cases where textual features that explicitly refer to a polarized opinion couldn't be detected. A further investigation would be necessary in order to make a distinction between neutral subjectivity (e.g. expressions of hope, without a positive or negative valence) and pure objectivity (Wilson, 2008; Liu, 2010).

For what concerns tweets marked as positive and negative, if we hold together the ironic-polarized tweets with their corresponding labels, we have 924 negatives (37.09% of the total) against 263 positive posts (10.7%). The disparity is amplified when we take into account just ironic tweets. The use of irony for conveying a positive opinion is very rare (18 occurrences only).

label	occurrences
NONE	2,469
NEG	1381
POS	497
HUM NEG	404
HUM POS	18
MIXED	44

Table 1: Labels distribution in TW-BS.

A comparative analysis of polarity distribution in the TW-BS and the WEB-BS corpora has shown further important differences. The distribution of polarity is more balanced in the latter than in the former, where negative polarity prevails, while irony, frequently occurring in the Twitter corpus is almost absent in the other one. This confirms our theoretical hypothesis that the global contexts underlying these datasets are different, but also raises issues about the higher politeness and the cooperativeness applied by users in the consultation with respect to what is expressed in a social media context like Twitter. Furthermore, the nature of ironic posts on Twitter deserves further and deeper investigations, e.g. about the relation between the presence or the absence of ironic tweets and the occurrence of particular events, like the press conference that launched the reform. For what concerns instead the analysis of topics, we observed that, even if the disagreement has not been high (31.4%), the annotators mostly did agree on the generic label BUONA SCUOLA, which occurs 4,071 times with the agreement of two annotators. This is confirmed by the limited exploitation of the more specific labels for the annotation of topic: the total amount of all the specific labels is 1,502. Moreover, it emerges a difference between the topics selected by users in WEB-BS corpus, and the ones annotated in the TW-BS corpus. This difference between contents proposed by the government and the topics spread out from the micro-blogging platform can be observed by looking at the different distribution of the labels in the two contexts. If we consider just the 13 label used both for TW-BS corpus, and WEB-BS corpus, we can notice important differences. For instance, VALUTAZIONE was the mainly used during the debate (15.72%), but attested few times in Twitter (2.52%). Otherwise, the label RECLUTAMENTO, which was used only the 7.56% of the times in the WEB-BS corpus, is the most frequent in the TW-corpus (39.68% of the occurrences).

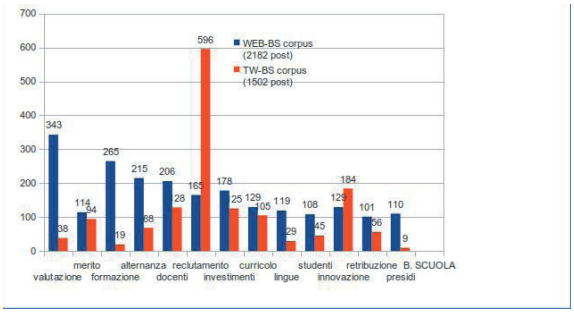


Figure 1: the use of topic labels in WEB-BS corpus, and TW-BS corpus

5 Conclusions

The paper describes a project for the analysis of a socio-political debate in a sentiment analysis perspective. A novel resource is presented by describing the collection and the annotation of the dataset organized in two subcorpora according to the source the texts have been extracted from: one from Twitter and one from the institutional online consultation platform. A first analysis of the resulting dataset is presented, which takes into account also a comparative perspective.

Acknowledgements

The authors thank all the persons who supported the work. We are grateful to our annotators, in particular to Valentina Azer, Marta Benenti, Martina Brignolo, Enrico Grosso and Maurizio Stranisci. This research is supported in part by Fondazione Giovanni Goria e Fordazione CRT (grant Master dei Talenti 2014; Marco Stranisci) and in part by the National Council for Science and Technology, CONACyT Mexico (Grant No. 218109, CVU-369616; D.I. Hernández Farías).

References

G. Attardi, V. Basile, C. Bosco, T. Caselli, F. Dell'Orletta, S. Montemagni, V. Patti, M. Simi, and R. Sprugnoli. 2015. State of the art language technologies for italian: The evalita 2014 perspective. *Journal of Intelligenza Artificiale*, 9(1):43–61.

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, Georgia. 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'14)*, pages 50–57, Pisa, Italy. Pisa University Press.

Carla Bazzanella. 2010. Contextual constraints in cmc narrative. In Christian Hoffmann, editor, *Narrative Revisited*, pages 19–38. John Benjamins Publishing Company.

Liu Bing. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

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.

E. Cambria, B. Schuller, Y. Xia, and C. Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

Andrea Ceron, Luigi Curini, and Iacus M. Stefano. 2014. *Social Media e Sentiment Analysis: l'evoluzione dei fenomeni sociali attraverso la rete*. Springer.

Michelangelo Conoscenti. 2011. *The Reframer: An Analysis of Barack Obama Political Discourse (2004-2010)*. Bulzoni Editore.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, A. Reyes, and J. Barnden. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proc. Int. Workshop on Semantic Evaluation (SemEval-2015), Co-located with NAACL and *SEM*.

Mirko Lai, Daniela Virone, Cristina Bosco, and Viviana Patti. 2015. Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. In *Proc. of 2015 IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015), Special Track on Emotion and Sentiment in Intelligent Systems and Big Social Data Analysis.*, Paros, France. IEEE. In press.

Bing Liu. 2010. *Sentiment analysis and subjectivity*. Taylor and Francis Group, Boca.

Diana Maynard and Adam Funk. 2011. Automatic detection of political opinions in tweets. In *Extended Semantic Web Conference Workshop*, pages 88–99.

Diana Maynard and Mark 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'14)*, Reykjavik, Iceland, may. ELRA.

Antonio Reyes and Paolo Rosso. 2014. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, 40(3):595–614.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Erik Tjong Kim Sang and Johan Bos. 2012. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: communication and cognition*. Basil Blackwell.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2011. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the ICWSM-11*, pages 178–185, Barcelona, Spain.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 115–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the intensity, polarity, and attitudes of private states*. Ph.D. thesis, University of Pittsburgh.

Francisco Yus. 2001. *Ciberpragmatica : el uso del lenguaje en Internet*. Ariel.

Reference-free and Confidence-independent Binary Quality Estimation for Automatic Speech Recognition

Hamed Zamani[⋆], José G. C. de Souza[†‡], Matteo Negri[‡], Marco Turchi[‡], Daniele Falavigna[‡]
[⋆]School of ECE, College of Engineering, University of Tehran, Iran
[†]University of Trento, Italy
[‡]HLT research unit, Fondazione Bruno Kessler, Trento, Italy
h.zamani@ut.ac.ir {desouza,negri,turchi,falavi}@fbk.eu

Abstract

English. We address the problem of assigning binary quality labels to automatically transcribed utterances when neither reference transcripts nor information about the decoding process are accessible. Our quality estimation models are evaluated in a large vocabulary continuous speech recognition setting (the transcription of English TED talks). In this setting, we apply different learning algorithms and strategies and measure performance in two testing conditions characterized by different distributions of "good" and "bad" instances. The positive results of our experiments pave the way towards the use of binary estimators of ASR output quality in a number of application scenarios.

Italiano. *Questo lavoro descrive un approccio che consente di assegnare un valore di qualità "binario" a trascrizioni generate da un sistema di riconoscimento automatico della voce. Il classificatore da noi sviluppato è stato valutato in un'applicazione di riconoscimento di parlato continuo per grandi vocabolari (la trascrizione di "TED talks" in Inglese), confrontando tra di loro diverse strategie.*

1 Introduction

Accurate and cost-effective methods to estimate ASR output quality are becoming a critical need for a variety of applications, such as the large vocabulary continuous speech recognition systems used to transcribe audio recordings from different sources (*e.g.* YouTube videos, TV programs, corporate meetings), or the dialogue systems for human-machine interaction. For obvious efficiency reasons, in some of these application scenarios, ASR output quality cannot be determined by means of standard reference-based methods. Indeed, besides the fact that reference transcripts are not always available, quality indicators should often be computed at run-time to ensure quick response. This motivates research towards alternative **"reference-free"** solutions. To cope with this problem, word-level confidence estimates have been used in the past either to measure how an ASR system is certain about the quality of its hypotheses (Wessel et al., 1998; Evermann and Woodland, 2000; Mangu, 2000; Xu et al., 2010, inter alia) or to automatically detect ASR errors (Seigel, 2013; Seigel and Woodland, 2014; Tam et al., 2014). The reliance on confidence information and the emphasis on the word/sub-word level mark the major differences between such prior works and our research, which aims to give an objective assessment of ASR output quality: *i)* at the whole utterance level and *ii)* without the constraint of having access to the system's decoding process. This information, in fact, is not always accessible, as in the case of the increasingly large amount of captioned audio/video recordings that can be found on the Web. This advocates for the development of **"confidence-independent"** quality estimation methods.

These problems have been addressed by Negri et al. (2014), who proposed the task of predicting the word error rate (WER) of an automatically transcribed utterance.[1] Results indicate that even with a relatively small set of *black-box* features (*i.e.* agnostic about systems' internal decoding strategies), the predictions closely approximate the true WER scores calculated over reference transcripts. Experiments, however, are limited to a regression problem and further developments either disregard its natural extension to binary classification (Jalalvand et al., 2015), or address it without the same exhaustiveness of this work (C. de Souza et al., 2015).

The automatic assignment of explicit good/bad labels has several practical applications. For instance, instead of leaving to the user the burden

[1]This "quality estimation" task presents several similarities with its counterpart in the machine translation field (Specia et al., 2009; Mehdad et al., 2012; Turchi et al., 2014; C. de Souza et al., 2014, *inter alia*).

of interpreting scores in the [0, 1] interval, easily-interpretable binary quality predictions would help in tasks like: *i)* deciding if an utterance in a dialogue application has been correctly recognized, *ii)* deciding if an automatic transcription is good enough for the corresponding audio recording or needs manual revision (*e.g.* in subtitling applications), *iii)* selecting training data for acoustic modelling based on active learning, and *iv)* retrieving audio data with a desired quality for subsequent processing in media monitoring applications.

To support these applications, we extend ASR quality estimation to the binary classification setting and compare different strategies. All of them significantly outperform the trivial approach based on thresholding predicted regression scores (our first contribution). The best solution, a *stacking method* that effectively exploits the complementarity of different models, achieves impressive accuracy results (our second contribution).

2 Methodology

Task Definition. Given a set of *(signal, transcription, WER)* tuples as training instances, our task is to label unseen *(signal, transcription)* test pairs as "good" or "bad" depending on the quality of the transcription. The boundary between "good" and "bad" is defined according to a threshold τ set on the WER of the instances: those with a $WER \leq \tau$ will be considered as positive examples while the others will be considered as negative ones. Different thresholds can be set in order to experiment with testing conditions that reflect a variety of application-oriented needs. At the two extremes, values of τ close to zero emphasize systems' ability to precisely identify high-quality transcriptions (those with $WER \leq \tau$), while values of τ close to one shift the focus to the ability of isolating the very bad ones (those with $WER > \tau$). In both cases, the resulting datasets will likely be rather imbalanced, which is a challenging condition from the learning perspective.

Approaches. We experiment with two different strategies. The first one, *classification via regression*, represents the easiest way to adapt the method proposed in (Negri et al., 2014). It fits a regression model on the original training instances, applies it to the test data, and finally maps the predicted regression scores into good/bad labels according to τ. The second one is *standard classification*, which partitions the training data into good/bad instances according to τ, trains a binary classifier on such data, and finally applies the learned model on the test set. The two strategies have pros and cons that are worth to consider. On one side, classification via regression directly learns from the WER labels of the training points. In this way, it can effectively model the instances whose WER is far from the threshold τ but, at the same time, it is less effective in classifying the instances with WER values close to τ. Moreover, in case of skewed label distributions, its predictions might be biased towards the average of the training labels. Nevertheless, since such mapping is performed *a posteriori* on the predicted labels, the behaviour of the model can be easily tuned with respect to different user needs by varying the value of τ. On the other side, standard classification learns from binary labels obtained by mapping *a priori* the WER labels into the two classes. This means that the behaviour of the model cannot be tuned with respect to different user needs once the training phase is concluded (to do this, the classifier should be re-trained from scratch). Also, standard classification is subject to biases induced by skewed label distributions, which typically results in predicting the majority class. To cope with this issue, we apply instance weighting (Veropoulos et al., 1999) by assigning to each training instance a weight *w* computed by dividing the total number of training instances by the number of instances belonging to the class of the given utterance.

Since classification via regression and standard classification are potentially complementary strategies, we also investigate the possibility of their joint contribution. To this aim, we experiment with a *stacking method*, or stacked generalization (Wolpert, 1992), which consists in training a meta-classifier on the predictions returned by an ensemble of base classifiers. To do this, training data is divided in two portions. One is used to train the base estimators; the other is used to train the meta-classifier. In the evaluation phase, the base estimators are run on the test set, their predictions are used as the features for the meta-classifier, and its output is returned as the final prediction.

Features. Similar to Negri et al. (2014), we experiment with 68 features that can be categorized into Signal, Hybrid, Textual, and ASR. The first group is extracted by looking at each voice segment as a whole. Hybrid features give a more fine-grained information obtained from knowledge of word time boundaries. Textual features aim to capture the plausibility/fluency of an automatic transcription. Finally, ASR features give information based on the confidence the ASR system has

on its output. Henceforth, we will refer to the first three groups as "*black-box*" features since they are agnostic about the system's internal decoding process. The fourth group, instead, will be referred to as the "*glass-box*" group since they consider information about the inner workings of the ASR system that produced the transcriptions. The glass-box features will be exploited in §3 to train the full-fledged quality estimators used as terms of comparison in the evaluation of our confidence-independent models.

To gather insights about the usefulness of our features, in all our experiments we performed feature selection using Randomized Lasso, or stability selection (Meinshausen and Bhlmann, 2010). Interestingly, the selected black-box features are uniformly distributed in all the groups; this suggests to keep all of them (and possibly add others, which is left for future work) while coping with binary quality estimation for ASR.

Learning Algorithms. Besides comparing the results achieved by different learning strategies, we also investigate the contribution of various widely used algorithms. For *classification via regression* we use Extremely Randomized Trees (XTR (Geurts et al., 2006)) and Support Vector Machines (SVR (Cortes and Vapnik, 1995)) regressors, while for *standard classification* we use Extremely Randomized Trees (XTC), Support Vector Machine (SVC (Mammone et al., 2009)), and Maximum Entropy (MaxEnt (Csiszár, 1996)) classifiers. MaxEnt is also used as the meta-classifier by our *stacking method*. In all experiments, hyper-parameter optimization is performed using randomized search (Bergstra and Bengio, 2012) over 5-fold cross validation over the training data.

3 Experiments

Dataset. We experiment with the ASR data released for the 2012 and 2013 editions of the IWSLT evaluation campaign (Federico et al., 2012; Cettolo et al., 2013) respectively consisting of 11 and 28 English TED talks. The 2012 test set, which has a total speech duration of around 1h45min, contains $1,118$ reference sentences and $18,613$ running words. The 2013 test set has a total duration of around 3h55min, it contains $2,238$ references and $41,545$ running words. In our experiments, we always use $1,118$ utterances for training and $1,120$ utterances for testing. To this aim, the (larger) IWSLT 2013 test set is randomly sampled three times in training and test sets of such dimensions. The use of two datasets is motivated by the objective of measuring variations in the classification performance of our quality estimators under different conditions: *i)* homogeneous training and test data from the same edition of the campaign, and *ii)* heterogeneous training and test data from different editions of the campaign. All utterances have been transcribed with the systems described in (Falavigna et al., 2012; Falavigna et al., 2013).

Evaluation Metric. As mentioned in §2, we need to assess classification performance with potentially imbalanced data distributions. It has been shown that a number of evaluation metrics for binary classification (*e.g.* accuracy, F-measure, etc.) are biased and not suitable for imbalanced data (Powers, 2011; Zamani et al., 2015). For this reason, we use the balanced accuracy (BA – the average of true positive rate and true negative rate), which equally rewards the correct classification on both classes (Brodersen et al., 2010).

Baseline and Terms of Comparison. The simplest baseline to compare with is a system that always predicts the most frequent class in the training data, which would result in a 50% BA score. Furthermore, we assess the potential of our binary quality estimators against two terms of comparison. The first one is an "*oracle*" obtained by selecting the best label among the output of multiple models. Such oracle is an informed selector able to correctly classify each instance if at least one of the models returns the right class. Significant differences between the performance achieved by the single models and the oracle would indicate some degree of complementarity between the different learning strategies/algorithms. Close results obtained with the stacking method would evidence its capability to leverage such complementarity. The second term of comparison is a *full-fledged quality estimator* that exploits glass-box features as a complement to the black-box ones. Performance differences between the black-box models and the full-fledged estimator will give an idea of the potential of each method both in the most interesting, but less favourable condition (*i.e.* when the ASR system used to transcribe the signal is unknown), and in the most favourable condition when confidence information is also accessible.

Results and Discussion. We evaluate our approach in two experimental setups, characterized by different distributions of positive/negative instances. These are obtained by setting the threshold τ to 0.05 and 0.4. In both settings, the minority class contains around 20% of the data in the major-

Table 1: Balanced accuracy (BA) obtained by different methods when $\tau = 0.05$

Train – Test	Features	Classification via regression	Standard classification	Stacking	Oracle
2013 – 2013	BB	SVR: 55.91 ± 3.06	XTC: 66.78 ± 0.18	**76.71 ± 2.23**	85.12 ± 1.69
2013 – 2013	ALL	SVR: 62.76 ± 1.46	SVC: 77.31 ± 1.33	**86.33 ± 1.93**	90.87 ± 1.03
2012 – 2013	BB	XTR: 50.00 ± 0.0	SVC: 62.34 ± 1.97	**75.90 ± 2.54**	85.63 ± 1.11
2012 – 2013	ALL	XTR: 61.72 ± 0.38	MaxEnt: 75.82 ± 0.85	**88.40 ± 0.74**	90.36 ± 0.61

Table 2: Balanced accuracy (BA) obtained by different methods when $\tau = 0.4$

Train – Test	Features	Classification via regression	Standard classification	Stacking	Oracle
2013 – 2013	BB	SVR: 68.49 ± 1.03	SVC: 72.29 ± 0.58	**78.47 ± 3.77**	86.65 ± 0.31
2013 – 2013	ALL	XTR: 76.63 ± 0.54	SVC: 80.43 ± 0.19	**88.06 ± 1.98**	89.11 ± 1.45
2012 – 2013	BB	XTR: 54.67 ± 1.21	SVC: 62.60 ± 2.06	**76.17 ± 2.78**	81.85 ± 1.74
2012 – 2013	ALL	SVR: 69.19 ± 0.62	MaxEnt: 80.02 ± 0.54	**87.34 ± 1.49**	90.80 ± 0.38

ity class. Tables 1 and 2 show the results obtained by: *i)* models trained and evaluated on data either from the same (2013-2013) or different editions of IWSLT (2012-2013), and *ii)* models trained using either ALL the features (*i.e.* glass-box and black-box) or only the black-box ones (BB). For the sake of brevity, only the performance of the best classification algorithms is provided, together with the stacking and oracle results. In order to eyeball the significance of the difference in mean values, for each result we also report the standard deviation.

The analysis of the results yields several findings, relevant from the application-oriented perspective that motivated our research. First, in all the testing conditions our best binary classifiers significantly outperform the majority class baseline (50% BA). Top results with homogeneous data (2013-2013) are up to *78.47%* when only black-box features are available, and *88.40%* when all the features are combined. Not surprisingly, the scores achieved by classifiers trained only with BB features are lower than those achieved by models that can leverage ALL the features. Nevertheless, the positive results achieved by the BB features indicate their potential to cope with the difficult condition in which the inner workings of the ASR system are not known.

As regards the different learning strategies, a visible trend can be observed: standard classification significantly outperforms classification via regression in all cases. This indicates that it substantially benefits from the instance weighting mechanism described in §2, and from the fact that model selection can be performed by maximizing BA (the same metric used to evaluate the system), which cannot be used by the classification via regression strategy. Regarding the algorithmic aspect, the analysis of the results does not lead to definite conclusions. Indeed, none of the tested algorithms seems to consistently prevail across the different testing conditions and, especially for classification via regression, the best score is often not significantly better than the others. Looking at the oracle, its high BA suggests a possible complementarity between the different strategies/algorithms, and large room for improvement over the base estimators. Such complementarity is successfully exploited by the stacking method, which drastically reduces the gap in all cases.

All the learning strategies suffer from evaluation settings where the data distribution is heterogeneous (2012-2013). Although the oracle results do not show large differences when moving from the 2013-2013 to the 2012-2013 setting, almost all the results show consistent performance drops in the latter, more challenging setting. Nonetheless, the BA achieved by the stacking method is rather high, and always above *75%*.

4 Conclusions

We investigated the problem of assigning informative and unambiguous binary quality labels (good/bad) to automatically transcribed utterances. Aiming at an application-oriented approach, we developed a *reference-free* and *confidence-independent* method, which has been evaluated in different settings. Our experiments on English TED talks' transcriptions from the IWSLT campaign show that our best *stacking* models can successfully combine the complementarity of different strategies. With a balanced accuracy ranging from *86.33%* to *88.40%*, the full-fledged classifiers that combine black-box and glass-box (*i.e.* confidence-based) features bring the problem close to its solution. With results in the range *75.90%-78.47%*, our reference-free and confidence-independent models provide a reliable solution to meet the demand of cost-effective methods to estimate the quality of the output of unknown ASR systems.

References

J. Bergstra and Y. Bengio. 2012. Random Search for Hyper-parameter Optimization. *J. Mach. Learn. Res.*, 13(1):281–305.

K. H. Brodersen, C. S. Ong, K. Enno Stephan, and J. M. Buhmann. 2010. The Balanced Accuracy and Its Posterior Distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, ICPR '10, pages 3121–3124, Istanbul, Turkey.

José G. C. de Souza, Marco Turchi, and Matteo Negri. 2014. Machine Translation Quality Estimation Across Domains. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014): Technical Papers*, pages 409–420, Dublin, Ireland, August.

José G. C. de Souza, Hamed Zamani, Matteo Negri, Marco Turchi, and Daniele Falavigna. 2015. Multitask learning for adaptive quality estimation of automatically transcribed utterances. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 714–724, Denver, Colorado.

M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the International Workshop for Spoken Language Translation*, IWSLT '13, Heidelberg, Germany.

C. Cortes and V. Vapnik. 1995. Support-Vector Networks. *Mach. Learn.*, 20(3):273–297.

I. Csiszár. 1996. Maxent, Mathematics, and Information Theory. In *Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods*, pages 35–50, Sante Fe, New Mexico, USA.

G. Evermann and P. C. Woodland. 2000. Large Vocabulary Decoding and Confidence Estimation Using Word Posterior Probabilities. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '00, pages 2366–2369, Istanbul, Turkey.

D. Falavigna, G. Gretter, F. Brugnara, and D. Giuliani. 2012. FBK @ IWSLT 2012 - ASR Track. In *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK.

D. Falavigna, R. Gretter, F. Brugnara, D. Giuliani, and R. Serizel. 2013. FBK@IWSLT 2013 - ASR Tracks. In *Proc. of IWSLT*, Heidelberg, Germany.

M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker. 2012. Overview of the IWSLT 2012 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, Hong Kong, HK.

P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely Randomized Trees. *Mach. Learn.*, 63(1):3–42.

Shahab Jalalvand, Matteo Negri, Falavigna Daniele, and Marco Turchi. 2015. Driving rover with segment-based asr quality estimation. 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)*, pages 1095–1105, Beijing, China.

A. Mammone, M. Turchi, and N. Cristianini. 2009. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3):283–289.

L. Mangu. 2000. *Finding Consensus in Speech Recognition*. Ph.D. thesis, John Hopkins University.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*, pages 171–180.

N. Meinshausen and P. Bhlmann. 2010. Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

M. Negri, M. Turchi, and J. G. C. de Souza. 2014. Quality Estimation for Automatic Speech Recognition. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, Dublin, Ireland.

D. M. W. Powers. 2011. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. *J. Mach. Learn. Tech.*, 2(1):37–63.

M. S. Seigel and P. C. Woodland. 2014. Detecting Deletions in ASR Output. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '14, pages 2321–2325, Florence, Italy.

M. S. Seigel. 2013. *Confidence Estimation for Automatic Speech Recognition Hypotheses*. Ph.D. thesis, Cambridge University.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 28–35, Barcelona, Spain.

Y. C. Tam, Y. Lei, J. Zheng, and W. Wang. 2014. ASR Error Detection Using Recurrent Neural Network Language Model and Complementary ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '14, pages 2331–2335, Florence, Italy.

Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 710–720, Baltimore, Maryland.

K. Veropoulos, C. Campbell, and N. Cristianini. 1999. Controlling the Sensitivity of Support Vector Machines. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, IJCAI '99, pages 55–60, Stockholm, Sweden.

F. Wessel, K. Macherey, and R. Schlüter. 1998. Using Word Posterior Probabilities as Confidence Measures. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '98, pages 225–228, Seattle, Washington.

D. H. Wolpert. 1992. Stacked Generalization. *Neural Networks*, 5(2):241–259.

H. Xu, D. Povey, L. Mangu, and J. Zhu. 2010. An Improved Consensus-Like method for Minimum Bayes Risk Decoding and Lattice Combination. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '10, Dallas, Texas, USA.

Hamed Zamani, Pooya Moradi, and Azadeh Shakery. 2015. Adaptive user engagement evaluation via multi-task learning. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1011–1014, Santiago, Chile.