From Natural Language to Ontology Population in the Cultural Heritage Domain.

A Computational Linguistics-based approach.

Maria Pia di Buono, Mario Monteleone
University of Salerno
Address
Email: mdibuono@unisa.it, mmonteleone@unisa.it

Abstract

This paper presents an on-going Natural Language Processing (NLP) research based on Lexicon-Grammar (LG) and aimed at improving knowledge management of Cultural Heritage (CH) domain. We intend to demonstrate how our language formalization technique can be applied for both processing and populating a domain ontology. We also use NLP techniques for text extraction and mining to fill information gaps and improve access to cultural resources. The Linguistic Resources (LRs, i.e. electronic dictionaries) we built can be used in the structuring of effective Knowledge Management Systems (KMSs).

In order to apply to Parts of Speech (POS) the classes and properties defined by the Conseil Interational des Musees (CIDOC) Conceptual Reference Model (CRM), we use Finite State Transducers/Automata (FSTs/FSA) and their variables built in the form of graphs. FSTs/FSA are also used for analysing corpora in order to retrieve recursive sentence structures, in which combinatorial and semantic constraints identify properties and denote relationship. Besides, FSTs/FSA are also used to match our electronic dictionary entries (ALUs, or Atomic Linguistic Units) to RDF subject, object and predicate (SKOS Core Vocabulary).

This matching of linguistic data to RDF and their translation into SPARQL/SERQL path expressions allows the use ALUs to process natural-language queries.

Keywords: Natural Language Processing, Ontology Population, Cultural Heritage

1. Introduction

The Cultural Heritage domain brings critical challenges as far as the application of NLP and ontology population techniques are concerned, mainly because it embraces a wide range of content variable by type and properties, semantically interlinked whit other domains thanks to Semantic Expansion (SE). Actually, by definition CH copes with several different aspects, which concern tangible and intangible heritage, therefore necessarily generating several kinds of descriptive data and metadata. On such premises, NLP techniques for text extraction and mining can be used to fill information gaps and improve access to cultural resources.

This paper presents an on-going Natural Language Processing (NLP) research based on Lexicon-Grammar (LG) and aimed at improving knowledge management of Cultural Heritage (CH) domain. We intend to demonstrate how our language formalization technique can be applied for both processing and populating a domain ontology.

We think that a coherent and consistent language formal description is crucial and indispensable to achieve a correct semantic representation of whatsoever knowledge domain. Our idea also springs from Bachimont (2000) who states that "defining an ontology for knowledge representation tasks means defining, for a given domain and a given problem, the functional and relational signature of a formal language and its associated semantics". Therefore, this study focuses on a different approach to content analysis and information retrieval, essentially based on language formal description, and which takes into account the fact that CH ALUs may also be interlinked with, or refer to, other knowledge domains, characterized by the simultaneous presence of free text fields, data and metadata.

2. Related Works

Prevalent approaches to ontology population are based on extraction tool-kits, also used for ALU identification in order to recognise instances of concepts or instances of relations between concepts.

Petasis *et al.* (2011) provide a comparative analysis of these systems.

Artequakt (Kim *et al.*, 2002) performs recognition and analysis using GATE[1], a toolkit for text engineering.

SOBA (Drozdzynski *et al.*, 2004) employs a rule-based information extraction system.

Instead, some systems extract patterns provided by users (e.g. KnowItAll, which applies domain-indipendent lexico-syntactic patterns [Etzioni *et al.*, 2004], based on Hearts [1992] patterns).

Navigli *et al.* (2006) present a system which use manual extraction patterns in order to populate the CIDOC CRM ontology.

These systems employ machine learning techniques, with statistically-based term identification and pattern

1 https://gate.ac.uk/

extraction. The LEILA (Suchanek *et al.,* 2006), which also applies linguistic knowledge, uses k-Nearest-Neighbor-classifiers and Support Vector Machines.

3. Methodology

3.1. Lexicon-Grammar Framework

Our NLP activities fall inside Lexicon-Grammar (LG) theoretical and practical framework, which is one of the most consistent methods for natural language formalization, automatic textual analysis and parsing. Its main goal is to describe all mechanisms of word combinations closely related to concrete lexical units and sentence creation, and to give an exhaustive description of lexical and syntactic structures of natural language. LG was set up by the French linguist Maurice Gross during the '60s and subsequently applied to Italian by Annibale Elia, Emilio D'Agostino and Maurizio Martinelli[2]. Its theoretical approach is prevalently based on Zelig Sabbetai Harris' Operator-Argument Grammar (1976), which assumes that each human language is a self-organizing system, and that the syntactic and semantic properties of a given word may be calculated on the basis of the relationships this word has with all other co-occurring words inside given sentence contexts. The study of simple or nuclear sentences is achieved analyzing the rules of co-occurrence and selection restriction, i.e. distributional and transformational rules based on predicate syntactic-semantic properties[3].

3.2. Resources and Tools

As already stated, LG assumes that a coherent natural language formal description is crucial for developing NLP applications. The NLP approach it follows plans the structuring of exhaustive and descriptively taxonomic LRs (i.e. electronic dictionaries, syntactic matrix tables and local grammars). Thanks to this specific formal characteristics, such LRs have proven to be useful also in the development and implementation of effective Knowledge Management Systems (KMSs).

In LG, linguistic formalization is based on an accurate observation of linguistic phenomena, and on an appropriate linguistic data recording of all lexicon and lexical entry combinatory behaviors, encompassing syntax and, also, lexicon. It differs from the best known among current linguistic theories, i.e. Chomsky's deep grammar and its various offspring, which are strictly formalist and syntax-based. Also, LG uses electronic dictionaries to describe words morphological and grammatical features. These dictionaries are mainly based on the concepts of «meaning unit», «lexical unit», «atomic linguistic unit» and «word group», this last one also including Multi-Words Units (MWUs). Today, most frequentist or probabilistic textual analysis methods which apply statistical rules may collapse on MWUs analysis, due for instance to the low frequency of these lexical items in specific texts. Also, statistically-based parsers may not appropriately recognize even higly-frequent MWUs as single meaning units, consequently losing pieces of information. On the contrary we will see that being dictionary-based, LG identification and retrieval of MWUs is founded on a systematic and exhaustive formalization of natural language.

Our MWU/ALU treatment consists in their recognition and classification by means of formal, morph-grammatical information and terminological tags used to label entries of LG electronic dictionaries. Such dictionaries are used as linguistic engines to automatically read and parse texts, therefore also to recognize and locate MWUs/ALUs inside texts. At the same time, in order to achieve NLP applications such as Information Retrieval (IR) and/or Machine Translation, we use morph-syntactic information (co-occurrence and selection restriction) to build local grammars. This is due to the fact that local grammars mostly work as a specific tool to cope with special phenomena of language in applications which make use of natural language. More appropriately, local grammars design is based on the syntactic description which encompasses transformational rules and distributional behaviours. To specify, we build local grammars in form of finite-state transducer and finite-state automata (see [2.4] for more specifications).

3.3. LG Description of Multi-Word Units

The expression "multi-word unit" is a fairly recent on stressing that LG has always referred to such constructions as to compound words. Actually, today the terms "collocations", "multi-words", "multiword expressions" and "multiword units" are often used in literature to indicate "strings of words having a unique overall meaning". These terms seem rather ambiguous and less effective than "compound ALUs" to distinguish between free word groups (i.e. compositional non-terminological free word formations) and all other kind of word formations (going from compound terminological words to proverbs). Besides, according to LG, only the items of the second group are to be lemmatized in electronic dictionaries. Therefore in this paper we will adopt the expressions "compound ALUs" to refer to any kind of lemmatizable "strings of words having a unique overall meaning", including terminological compound words (hence also CH ones), which even being very often semantically compositional, can be lemmatized due to their particular non-ambiguous information content. (Vietri, Monteleone, *in print*).

In our CH electronic dictionary, each entry is morph-grammatically and formally described, and is also given an ontological identification, consisting in tags which send

2 See Elia, Martinelli, D'Agostino (1978), on which basis we built our Italian LRs.

3 As we will see, LG co-occurrence and selection-restriction rules may be also described by means of RDF graphs.

back to the knowledge domain(s) within which entries are commonly used (i.e. in which they have terminological non-ambiguous meanings).

freccia di balestra	N+NPN+FLX=C45+DOM=RA1SUOARAL
freccia foliata	N+NA+FLX=C556+DOM=RA1SUOIL
freccia triangolare	N+NA+FLX=C569c+DOM=RA1SUOIL
fregio con coronamento	N+NPN+FLX=C12+DOM=RA1EDEAES
fregio dorico	N+NA+FLX=C523+DOM=RA1EDEAES
fuseruola biconica	N+NA+FLX=C547+DOM=RA1SUOCF
fuseruola biconvessa	N+NA+FLX=C542+DOM=RA1SUOCF
fusto a spirale	N+NPN+FLX=C7+DOM=RA1EDEAES

Table 1: Extract from Italian Electronic Dictionary of the Archaeological Domain.

For instance, the compound word *fregio dorico* («Doric frieze») is labeled with the tag «+DOM=RA1EDEAES», which stands for «Archaeological Artefacts – Building – Architectural Elements – Structural Elements».
For each entry, a formal and morphological description is also given with:

- the internal structure of each compound. So, in the compound word *fregio dorico* the tag «NA» indicates that the given compound is formed by a Noun, followed by an Adjective. At the same time, in the compound word *fregio con coronamento* the tag «NPN», indicates that the given compound is formed by a Noun followed by a Preposition, followed by a Noun;
- the inflectional class. So, the tag «+FLX=C523» indicates the gender and the number of the compound *fregio dorico*, together with its plural form. The inflectional class refers to a local grammar, so, the tag indicates that *fregio dorico* is masculine singular, does not have any feminine correspondent form, and its plural form is *fregi dorici*.

Currently we have developed the Italian electronic dictionary for the Archaeological Domain, which is composed by about 11000 compound words.
In order to develop these electronic dictionaries, we used the Thesauri and Guidelines of the Italian Central Institute for the Catalogue and Documentation (ICCD)[4]. These Thesauri are controlled vocabularies intended to be used by cataloguers and other professionals concerned with information management in the field of Archaeology. They include terms, descriptions and other information needful to objects cataloguing. For each dictionary we have developed a taxonomy, therefore all entries have a terminological and domain label usable for ontologies population.
The use of domain label subset tags is also previewed for those domain sectors which include specific sub-sectors. This is the case with Archaeological Artefacts, for which a generic tag «RA1» is used, while more explicit tags are used for Object Type, Subject, Primary Material, Method of Manufacture, Object Description.

3.4. Finite State Automata/Finite State Transducers (FSA/FSTs)

An FST is a graph which represents a set of text sequences and which associates each recognized sequence to specific analysis result, also considering their semantics. Text sequences are described in the input part of the FST; the corresponding results are described in the output part of the FST. Conversely, an FSA is a special type of finite-state transducer which doesn't produce any result (i.e. it has no output) (Silberztein, 1993). It is typically used to locate morph-syntactic patterns inside corpora, and it extracts matching sequences in order to build indices, concordances, etc. The development of FST/FSA is useful to automatically recognize any kind of text.
Figure 1 shows a finite state automaton composed of a single path with four nodes (from the initial symbol on the left to the end symbol on the right). When the graph is applied to a text, it recognizes all text accounted for by the sequence of nodes and states. Words in angle brackets stand for lemma forms, the shaded boxes represent a sub-graph

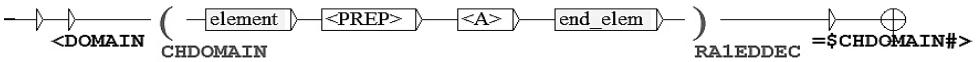

Figure 1: Example of a simple FSA for the recognition of Archaeological artifacts descriptions.

4 http://www.iccd.beniculturali.it/index.php?it/240/vocabolari.

(meta-node) which can freely be embedded in more general graphs. Graph embedding allows to reuse sub-graphs in more than one context. At a more theoretical level, it introduces the power of recursion inside grammars.
Sub-graphs may also be used to represent a semantic class and can be encoded in a dictionary with specific semantic features. Electronic dictionaries allow an arbitrary number of semantic features to be represented as tags of lexical entries, and they can also be used in the definition of local grammars.
When the word form is set between angle brackets, the graph locates all the word forms that are in the same equivalence set as the given word form (generally all inflected, derived forms, or spelling variants of a given lexical entry).
Therefore, the graph showed in Figure 1 recognizes the following text strings:

palmetta a cinque petali
(*palmetta+semipalmetta+rosetta*) <any preposition> <any adjective> (*petali+lobi+foglie*)

Also in CH as in many other terminological domains, phrase and sentence structures may present recursive formal structures (see the output of Figure 1). Such structures form what in lexicology are called "open lists", i.e. lists of compound ALUs having the first two or three items in common.
Such feature allows the building of non-deterministic FSA/FSTs, with which it is possible to recognize all the element of a specific open list as the one showed above. Also, as regards declarative sentences, RDF gives the possibility to recognize sentences conveying information of the type "X is an element of Y", which also have recursive structures. All this means that a single FSA/FST can be used to:

1. account for all the items of an open list;
2. account for all declarative sentences of the type "X is a part of Y", in which X and Y are pre-defined classes;
3. allow the matching of POS to RDF triples.

4. Semantic Annotation

The ontology we rely upon is defined by the Conseil International des Musees (CIDOC) Conceptual Reference Model (CRM). CIDOC CRM is composed of 90 classes (which includes sub-classes and super-classes) and 148 unique properties (and sub-properties). This object-oriented semantic model is compatible with the Resource Description Framework (RDF). Therefore, FSA/FSTs are used to identify classes and properties for RDF subjects, objects and predicates to which the Standard Simple Knowledge Organization System (SKOS) concept scheme will be associated. To each instance we add a meaningful relationship with other instances in terms of RDF triple in which the predicate is the descriptor annotated by means of a URI extracted from Dublin Core Metadata Model.
Such a SKOS/RDF concept scheme will be expanded by means of new instances or associative links/relationships, i.e. by adding URIs dealing with concepts and associative relationships among such concepts (see Section 2.3.2 – SKOS Primer). This procedure will grant a coherent semantic expansion useful to ameliorate natural language query effectiveness. Figure 2 gives a sample of FSA/FST variables associated to and applied with an RDF scheme for the following sentence:

Il Partenone (subject) presenta (predicate) colonne doriche e ioniche (object)

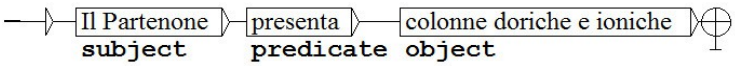

Figure 2: Simple FSA/FST with RDF Graph.

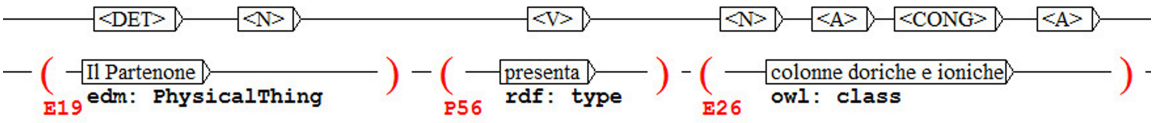

Figure 3: Sample of the use of the FSA variables for identifying classes and property.

According to our approach, electronic dictionaries entries (simple words and MWUs) are the subject and the object of the RDF triple.
In Figure 3 we develop an FSA with variables which apply to the Part of Speech (POS) classes and properties: (i) E19

indicates “Physical Object” class, (ii) P56 stands for “Bears Feature” property, (iii) E26 indicates “Physical Feature” class.
The role pairs Physical Object/name and Physical Feature/type are trigged by the RDF predicate *presenta*.
In Figure 3 we used variables to apply a tag which indicates classes for the compound words; when the FSA recognizes the text string, it applies a text annotation.
Besides in Figure 3 we also indicate specific Parts of Speech (POS) for the first noun phrase *Il Partenone* (DETerminer + Noun), the verb *presenta* (V) and the second noun phrase *colonnedoriche e ioniche* (Noun+Adjective+Conjuntion+Adjective).
By applying the automaton in Figure 3 (built considering the high variability of the lexical class and not of the single form belonging to the class), we can recognize all instances included in E19 and E26 classes, the property of which is P56.

5. Ontology Semi-automatic Population

Our future goal is the development, inside NOOJ, of a module for semi-automatic ontology population. Starting from the entries retrieved and from their specific tags, stored in electronic dictionaries and in FSA/FSTs, such tool will write and fill all fields directly using RDF schema and OWL, automatically generating the strings while correctly coupling ontologies and compound words.
After being tested and debugged, the LRs described so far are actually under final development and completion as part of the NooJ[5] Italian module. The possibility to export the results of NooJ automatic textual analysis using RDF and SKOS, and also the use of Linguistic Linked Open Data (LLOD) URIs to tag electronic dictionary entries are two of the main features by means of which our system of ontology semi-automatic population will be built. This procedure will be structured according to the following steps:

1. NooJ processes a text, parses it, and locates all the terminological ALUs in a given text;
2. subsequently, the ALUs retrieved are conceptually described by means of SKOS schemes and features, as for instance those used in EDM;
3. at the same time, RDF triples are transformed into SKOS tags in which concepts as E19 or P56 are rewritten by means of corresponding “edm PhysicalThing” or “rdf: type”;
4. finally all NooJ output is transformed into full XML, thanks to which users' natural-language queries can be used to retrieve information also in unstructured texts.

5 For more information on NooJ, see www.nooj4nlp.org.

6. Conclusion

Although our methodology relies heavily on a linguistic processing phase and requires robust resources and background knowledge, it allows to perform both object/term and synonym identification and also to recognise relations. Since it is based on a deep analysis and formalization of linguistic phenomena, our approach can also ensures portability to other domains, preserving ontology consistency and entity disambiguation.
NLP routines based on Lexicon-Grammar allow to support the automatic semantic annotation/indexation of textual documents in the field of Cultural Heritage.
Terminological tagging is a central step as regards Information Retrieval, Information Extraction, Information Storage, Machine Translation, ontology development, lexicon-dependent Semantic Web, query-free procedures for knowledge structuring, and also a question answering fostering a better «intelligent agent» interaction between humans and technology.

Note

Maria Pia di Buono is author of sections 1, 2, 3.4, 4 and 5 and Mario Monteleone is author of section 3.1, 3.2, 3.3 and 6.

7. References

Bachimont B. (2000). Engagement sémantique et engagement ontologique: conception et réalisation d’ontologies en ingénierie des connaissances In Charlet, J.; Zacklad, M.; Kassel, G. & Bourigault, D. (Eds.), *Ingénierie des connaissances, évolutions récentes et nouveaux défis.* Paris: Eyrolles.

Crofts, N.; Doerr M.; Gill, T.; Stead, S.; Stiff, M. (2010). *Definition of the CIDOC Conceptual Reference Model*. *ICOM/CIDOC Documentation Standards Group.* CIDOC CRM Special Interest Group. 5.02 ed.

De Bueris, G.; Elia, A. (eds.). (2008). *Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche.* Salerno: Eyrolles.

Drozdzynski, W.; Krieger, H.-U.; Piskorski, J.; Schäfer, U.; Xu, F. (2004) Shallow processing with unification and typed feature structures – foundations and applications. *Künstliche Intelligenz 1*, pp. 17–23.

Elia, A.; Martinelli, M.; D'Agostino, E. (1981). *Lessico e strutture sintattiche. Introduzione alla sintassi del verbo italiano.* Napoli: Liguori Editore.

Elia, A.; Vietri, S.; Postiglione, A.; Monteleone, M.; Marano, F. (2010). Data Mining Modular Software System. In Arabnia H.R., Marsh A., Solo A.M.G, *Proceedings of The 2010 International Conference on Semantic Web & Web Services*, WorldComp 2010 Conference, July 12-15 2010. Las Vegas Nevada USA: CSREA Press, pp. 127-133.

Etzioni, O.; Kok, S.; Soderland, S.; Cagarella, M.; Popescu, A.M.; Weld, D.S.; Downey, D.; Shaker, T.; Yates, A.

(2004) Web-Scale Information Extraction in KnowItAll (Preliminary Results). In *Proceedings of the 13th International World Wide Web Conference* (WWW 2004), New York, pp. 100–110.

Gross, M. (1975). *Méthodes en syntaxe.* Paris: Hermann.

Gruber, T. A. (1993). *Translation Approach to Portable Ontology Specifications*. Knowledge Acquisition, 5(2), pp. 199-220.

Harris, Z. S. (1976). *Notes du cours de syntaxe*, traduction française par Maurice Gross, Paris : Le Seuil.

Harris, Z. S. (1982). *A Grammar of English on Mathematical Principles*. New York: John Wiley and Sons.

Isaac, A.; Summers, E. (eds.) (2009) *SKOS Simple Knowledge Organization System Primer.*

Kim, S.; Alani, H.; Hall, W.; Lewis, P.; Millard, D.; Shadbolt, N.; Weal, M. (2002) Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web. In *Proceedings of Workshop on Semantic Authoring, Annotation & Knowledge Markup* (SAAKM 2002), the 15th European Conference on Artificial Intelligence (ECAI 2002). Lyon, France, pp. 1–6 .

Monteleone, M. (2004). *Lessicografia e dizionari elettronici. Dagli usi linguistici alle basi di dati lessicali.* Napoli: Fiorentino & New Technology.

Navigli, R.; Velardi, P. (2006) Enriching a Formal Ontology with a Thesaurus: an Application in the Cultural Heritage Domain. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* – OLP 2006, Sydney, Australia, pp. 1–9.

Petasis, G.; Karkaletsis, V.; Paliouras, G.; Krithara, A.; Zavitsanos, E. (2011). Ontology Population and Enrichment: State of Art. In G. Paliouras et al. (eds.). *Multimedia Information Extraction*. LNAI 6050. Berlin Heidelberg: Springer-Verlag, pp. 134-166.

Postiglione, A.; Monteleone, M.; Marano, F.; Monti. J.; Napoli, A. (2012). Electronic Dictionaries for Information Retrieval, Automatic Textual Analysis and Semantic-Based Data Mining Software. In *Database, Corpora e Insegnamenti Linguistici.* Bari-Parigi: Schena Editore – Alain Baudry et Cie.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes*. Paris: Masson.

Suchanek, F.M.; Ifrim, G.; Weikum, G. (2006) LEILA: Learning to Extract Information by Linguistic Analysis. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006*, Sydney, Australia, pp. 18–25.

Vietri, S.; Elia, A.; D'Agostino, E. (2004). Lexicon-grammar, Electronic Dictionaries and Local Grammars in Italian. In Laporte, E.; Leclere, C.; Piot, M.; Silbertzein, M.; (eds.). *Syntaxe, Lexique et Lexique Grammaire*. Volume dédié à Maurice Gross, *Lingvisticae Investigationes Supplementa n. 24*. Amsterdam/Philadelphia: John Benjamins, pp. 125-136.

Vietri, S. (2008). *Dizionari elettronici e grammatiche a stati finiti. Metodi di analisi formale della lingua italiana.* Salerno: Plectica.

Vietri, S.; Monteleone, M. (in print). The English NooJ dictionary. In *Proceedings of NooJ 2013 International Conference,* June 3-5 2013, Saarbrücken.