An Ontology-Based Method for Extracting and Classifying Domain-Specific Compositional Nominal Compounds

Maria Pia di Buono
TakeLab
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
mariapia.dibuono@fer.hr

Abstract

In this paper, we present our preliminary study on an ontology-based method to extract and classify compositional nominal compounds in specific domains of knowledge. This method is based on the assumption that, applying a conceptual model to represent knowledge domain, it is possible to improve the extraction and classification of lexicon occurrences for that domain in a semi-automatic way. We explore the possibility of extracting and classifying a specific construction type (nominal compounds) spanning a specific domain (Cultural Heritage) and a specific language (Italian).

1 Introduction

In the Cultural Heritage domain, as in many other specific domains of knowledge, phrases and word sequences present recursive formal structures. Some of these structures may form lists of compounds that have specific meanings only when used with reference to that domain and, for this reason, constitute the terminology of that domain. This means that if such compounds present a certain degree of polysemy, it will be possible to disambiguate usages according to the different specific domains they belong to. Thus, taking into account a specific domain of knowledge, compounds become unambiguous and clear terms, useful for conceptualizations, which contribute to outline formalizations. In this sense, we can assert that domain-specific compounds present two levels of representation, which are separated but interlinked: a conceptual-semantic level, pertaining to the knowledge domain and its ontology, and a syntactic-semantic level, pertaining to sentence and word production. We adopt the expression atomic linguistic units (ALUs) to indicate any kind of lemmatizable terminological compound words which, even being very often semantically compositional, can be lemmatized due to their particular non-ambiguous informational content (Vietri and Monteleone, 2013). In this paper, we explore the possibility of extracting and classifying a compositional ALU type (nominal compounds) spanning a specific domain (Cultural Heritage) and a specific language (Italian).

This paper is organized as follows, section 2 describes the background and related work. Our approach is detailed in section 3. The description of testing and results is given in section 4. Section 5 concludes the paper and points to future work.

2 Related Work

The task of dealing with ALUs attracts the interest of several researches, due to the issue of consistently recognizing those groups of words, able to carry a different semantic expressiveness and charge than single words. Thus, the prediction of these linguistic phenomena in natural language processing (NLP) applications has been investigated by several scholars from different point of views. Due to the success for simple words, a growing attention concerns the application of distributional approaches for coping with compositional compounds (McCarthy et al., 2003; Reddy et al., 2011; Salehi et al., 2014). Likewise, different scholars aim at achieving distributed representations of word meanings using word embeddings for various purposes (Mikolov et al., 2013; Patel and Bhattacharyya, 2015).

We will see that, being both ontology and dictionary based, our identification and classification of ALUs is founded on a systematic and exhaustive formalization of natural language.

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings, pages 83–88, Vancouver, BC, August 4, 2017. ©2017 Association for Computational Linguistics

3 Methodology

Our method lays its foundations in the Lexicon-Grammar (LG) framework (Gross, 1984, 1986). LG considers lexicon as a group of terminal values, in a formal grammar of natural languages, which have to be associated to ordered sequences on the basis of independent combinatory behaviours and rules. Thus, lexicon is not separable from syntax, namely every lexical element, occurring in a sentence context, holds a grammatical function which combines with grammatical functions of other constituents. Combinatory behaviours are driven by co-occurrency and restriction-selection rules. Furthermore, we deal with compositional ALUs also according to semantic expansion mechanisms, firstly attested by Harris (1946). These mechanisms are useful to fully account for compositional ALUs, or better for free word groups. Due to the fact that this kind of phenomena may have some possibility to be automatically and successfully parsed by means of regular expressions[1].

In our research, we focus on a specific construction type, which may be described and retrieved as ALUs: nominal compounds which present a restricted semantic expansion. It means that such ALUs are formed by a head phrase, generally fixed or semi-fixed, followed by variable elements which belong to specific grammatical categories. These variable elements are characterized by a selection restriction, which is determined by the head phrase, which functions as a predicate, and by the semantic provisions which they represent. We define such ALUs as semi-open nominal compounds, namely word sequences formed by one or more (semi)fixed elements and a restricted selection of co-occurring elements. As in the example it follows:

> (*palmetta+semipalmetta+rosetta*) + Adjective + Preposition + DNUM (*petali* + *lobi* + *foglie*).

In the previous ALU, we can recognize a restricted head *palmetta+semipalmetta+rosetta* followed by an adjective and a preposition, and a numeral, characterized by a high variability, and a restricted selection of nouns, i.e., *petali+lobi+foglie*.

In other words, in such lexical phenomena the fixed or semi-fixed head defines grammatical and semantic types of all variable elements. This phenomenon is mainly observable inside the lexicons of specific knowledge domains, even if it presents features belonging to both common-usage lexicon and terminology. Indeed, such semi-open ALUs are characterized by a variability of non-fixed elements but, at the same time, they are also characterized by a non-ambiguous meaning as a result of the compositional process.

In the following sections, we will show how we can recognize and, subsequently, classify by means of a domain ontology, such linguistic phenomena through a set of finite state automata (FSA), basing our method on co-occurrence likelihood of elements in semi-open ALUs.

3.1 Linguistic and Semantic Features

The high variability of non-fixed elements is related to the possibility of selecting elements from non-restraint sets of lexical items, the grammatical categories of which are predictable thanks to components constituting the head. On a lexical level, such a feature is correlated to the paradigmatic relationship which indicates words belonging to the same part of speech (POS) class. On the other hand, constraints deriving from heads components are associated to the syntagmatic relationship among words, that means to semantic aspects of ALUs. Thus, for example, in the Cultural Heritage domain, we may observe this phenomenon of semi-open ALUs in Coroplastic descriptions, as the following example shows:

- (1) *statua di* (statue of) [NPREP]+N
- (2) **statua di* (statue of) [NPREP]+A

'Statue of' represents the head, which determines the type of the element which comes afterwards, that must be a noun (1), and not an adjective (2). Indeed, if the head is composed by a noun, belonging to a specific semantic category, as *statua* (statue), followed by a preposition, like *di* (of), the element which comes afterwards must belong to noun POS. Similarly, the head works as a constraint for the type of noun selected, which means that we have a restricted semantic expansion concerning the semantic type of noun. Thus, the semi-

[1] On the contrary, non-compositional ALUs, as the result of a continuous interaction between *langue* and *parole* (De Saussure, 1989), are formed according also to linguistically motivated, manifold, deep-logic contiguous juxtapositions, and they present unpredictable formation routines. For this reason, they cannot be stochastically coped with, and must be necessary lexicalized if they present the even slightest non-compositional link among their components.

open NP 'statue of' may select a proper noun as 'Silenus', or a noun as 'woman', but not a noun as 'table'.

As far as syntactic aspects are concerned, some semi-open ALUs, especially referred to Coroplastic description, are sentence reductions in which a present participle construction is used. For instance,

- (3) *statua raffigurante Sileno* (statue representing Silenus) is a reduction (Gross, 1975; Harris and Gross, 1976) of the sentence:

  (3a) *Questa statua raffigura Sileno* (This statue represents Silenus)

  (3b) [relative] *Questa è una statua che raffigura Sileno* (This is a statue which represents Silenus)

  (3c) [pr. part.] *Questa è una statua raffigurante Sileno* (This is a statue representing Silenus)

These semi-open ALUs, which present sentence reductions, may be retrieved using FSA which recognize specific verb role-sets. Therefore, such an FSA recognizes sentence structures as they follow:

- NP(Head)+VP+NP
- NP(Head)+VP+NP+AP
- NP(Head)+VP+AP+NP
- NP(Head)+PREP+NP
- NP(Head)+PREP+NP+AP
- NP(Head)+PREP+AP+NP.

In the previous sample, the noun phrase (NP) which stands for the head of semi-open ALUs is composed by a group of non-restricted nouns related to Coroplastics. It means that in such a group we insert nouns as statue, bust, figure and so on.

As for semantics, we also observe the presence of semi-open ALUs in which the head does not occur in the first position. For example, the open series *frammenti di* (*terracotta*+*anfora*+*laterizi*+N) (fragments of (clay+amphora+bricks+N)), places the heads at the end of the compounds, being *frammenti* used to explicit the notion 'N0 is a part of N1'.

On the basis of our theoretical premises, and applying these selection restriction rules, we may identify syntactic and semantic sets of lexical elements which may co-occur in specific semi-open ALUs. Such recursive formal structures allow the development of non-deterministic FSA, suitable to recognize all the elements of a specific open list (di Buono et al., 2013).

3.2 Ontology-based Extraction and Classification

In order to recognize and extract this kind of semi-open ALUs, we develop a set of FSA, which takes advantage from the semantic information stored in electronic dictionaries developed by means of NooJ (Silberztein, 2008). NooJ allows to formalize natural language descriptions and to apply them to corpora. NooJ is used by a large community, which developed linguistic modules, including Finite State Automata/Transducers and Electronic Dictionaries, for more than twenty languages. The Italian linguistic resources (LRs) have been built by the Computational Linguistic group of University of Salerno, which started its study of language formalization from 1981 (Elia et al., 1981). Our analysis is based on the Italian module for NooJ (Vietri, 2014), enriched with the LRs for the Archaeological domain (di Buono et al., 2014). The Italian LRs are composed of simple and compound word electronic dictionaries, inflectional, syntactic and morphological grammars, annotated with cross-domain semantic tags (e.g., 'Human' and 'Animal' label). The LRs for the Archaeological domain present a taxonomic tagging, derived from the the Italian Central Institute for the Catalogue and Documentation (ICCD) guidelines, which indicate how to classify an object, and a reference to the CIDOC Conceptual Reference Model (CRM), defined by the Conseil International des Musees (Doerr, 2003) . The CIDOC CRM is an object-oriented semantic model, compatible with RDF, which stands for a domain ontology which may be applied to describe Cultural Heritage items and the relations among them. In this conceptual model, entity classes are described by means of pertaining information about the taxonomic relation among entity classes (i.e., Subclass of), a description of class essential properties (i.e., Scope note), sentences which exemplify natural language representations used to denote an element belonging to the class, and the properties which may co-occur with the given entity class.

Entry	Kylix a labbro risparmiato
POS	N
Int. Str.	NPREPNA
FLX	C610
SYN	lip cup
DOM	RA1SUOCR
CCL	E22

Table 1: Sample of a semantic and taxonomic annotated entry from the Archaeological Italian Electronic Dictionary.

Each entry in the LRs presents taxonomic and ontological information, as it follows (Table 1):

- Its POS, internal structure and inflectional code (+FLX), which recall a local grammar suitable for generating and recognizing inflected forms.
- Its variants (VAR) or synonyms (SYN), if any;
- With reference to the taxonomy, the pertaining knowledge domain (DOM), e.g., RA1SUOCR stands for Archaeological Remains/Tools/Kitchen Utensil;
- Finally, we insert a tag referring to the ontological entities derived from the CIDOC CRM, e.g., E22 refers to Man-Made Object class, that is a subclass of E19 Physical Object.

In order to create the role sets suitable to extract and classify the ALUs, we use these semantic information to apply a series of domain constraints. Thus, we firstly employ information which refer to the domain taxonomic hierarchy. For instance, in the sample (1), our first selection restriction is constrained by the tag value which indicates Sculpture class in the taxonomy. Therefore, we extract all ALUs, labeled with this tag, through a semi-automatic method. Consequently, a manual procedure is employed to identify nouns which fit to the meaningful sentence context.

In compounds containing present participle forms, i.e., sample (3), semantic features can be identified using local grammars built on specific verb classes (semantic predicate sets); in such cases, co-occurrence restrictions can be described in terms of lexical forms and syntactic structures. For these occurrences, we apply the specific semantic set, descriptive predicates, in order to put into evidence elements extracted from specific verbal classes (i.e., 20A and 47B[2]). We also employ grammatical and syntactic constraints referred to tense and number of verb phrase; thus, we select just present 3rd persons singular and plural (sample 3a and 3b) and present participle (sample 3c). Due to the complexity of Coroplastic descriptions, sub-graphs presents many recursive nodes, mainly in noun phrases.

4 Testing and Results

Our methodology has been tested on Italian Cultural Heritage texts. The corpus has been built merging different datasets of catalographic data provided by ICCD[3]. We refer to Archaeological Remains datasets, classified according to the guidelines of ICCD and released as open data[4]. The total amount of the dataset is about 123K records. Each record contains different information, structured according to the Functional Requirements for Authority Data (Patton, 2009). An evaluation of the results produced by our approach is given in Table 2. Our method is evaluated by means of Precision, Recall and F-score results in the extraction of the main entity classes, i.e. Building, Clothing, Furniture, Sculpture, and Tools. For this evaluation we consider some of the higher classes in the taxonomic classification of ICCD. This choice is justified by the compositional structures of ALUs, which are comparable for the subclasses related to the same main class. Anyway, the ontological tags used to classify them are fine-grained, so a distinction between these categories is performed any time. As we can notice, the values present a variability with reference to the different categories. Generally speaking, the cause of mismatching results can be retrieved in the use of too broad terms, which determines ambiguity hard to solve, i.e., *bitronconico*, that can be referred to an *askos*, belonging to the class of Tools, or to a kind of necklace, an element in the class of Clothing. Furthermore, another source of mismatching is related to the presence of references to the inventory number, or other information related to the ICCD classification, merged in the definition field, e.g., description of materials used.

[2] Classes refer to Italian Lexicon-Grammar Tables, available at http://dsc.unisa.it/composti/tavole/combo/tavole.asp.

[3] http://www.catalogo.beniculturali.it/opendata/?q=dataset.

[4] http://www.catalogo.beniculturali.it/opendata/?q=dataset.

Class	Precision	Recall	F-score
Building	0.87	0.77	0.81
Clothing	0.88	0.72	0.79
Furniture	0.79	0.66	0.72
Sculpture	0.89	0.68	0.77
Tools	0.85	0.75	0.80

Table 2: Evaluation for the main classes.

5 Conclusion and Future Work

In this paper, we have presented our preliminary study on an ontology-based strategy to extract and classify compositional nominal compounds in the Cultural Heritage domain for Italian. This FSA-based system allows to retrieve a very large amount of expressions and ALUs among those present in the analysed corpus. The highly productive formal structures are the following ones:

- Noun(Head) + Preposition + Noun + Preposition +Noun, i.e., *fibula ad arco a coste* (ribbed-arch fibula), in which the fixed component is represented by *fibula* (fibula);

- Noun(Head) + Preposition + Noun + Adjective, i.e., *anello a capi ritorti* (twisted-heads ring), the head is represented by *anello* (ring);

- Noun(Head)+ Preposition + Noun + Adjective + Adjective, i.e., *punta a foglia larga ovale* (oval broadleaf point).

The main hypothesis, leading the development of our system, is that the precision and the recall of extraction and classification systems for compositional compounds can be improved by representing linguistic and semantic features in a more consistent way. We consider the average results quite satisfying, nevertheless we are already planning to enrich our research outcomes with many other improvements in order to solve ambiguity and classification issues.

Acknowledgments

This work has been funded by the Unity Through Knowledge Fund of the Croatian Science Foundation, under the grant 19/15: EVEnt Retrieval Based on semantically Enriched Structures for Interactive user Tasks (EVERBEST).

References

Ferdinand De Saussure. 1989. *Cours de linguistique générale: Édition critique*, volume 1. Otto Harrassowitz Verlag.

Maria Pia di Buono, Mario Monteleone, and Annibale Elia. 2014. Terminology and knowledge representation. italian linguistic resources for the archaeological domain. In *Workshop on Lexical and Grammatical Resources for Language Processing*. page 24.

Maria Pia di Buono, Mario Monteleone, Federica Marano, and Johanna Monti. 2013. Knowledge management and cultural heritage repositories: Cross-lingual information retrieval strategies. In *Digital Heritage International Congress (DigitalHeritage), 2013*. IEEE, volume 2, pages 295–302.

Martin Doerr. 2003. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine* 24(3):75.

Annibale Elia, Maurizio Martinelli, and Emilio D'Agostino. 1981. *Lessico e strutture sintattiche: introduzione alla sintassi del verbo italiano*. Liguori.

Maurice Gross. 1975. *Méthodes en syntaxe*. Hermann.

Maurice Gross. 1984. Lexicon-grammar and the syntactic analysis of french. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 275–282.

Maurice Gross. 1986. Lexicon-grammar: the representation of compound words. In *Proceedings of the 11th coference on Computational linguistics*. Association for Computational Linguistics, pages 1–6.

Zellig S. Harris. 1946. From morpheme to utterance. *Language* 22(3):161–183.

Zellig S. Harris and Maurice Gross. 1976. *Notes de Cours de Syntaxe: Traduit de l'anglais par Maurice Gross*. Seuil.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*. Association for Computational Linguistics, pages 73–80.

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

Dhirendra Singh Sudha Bhingardive Kevin Patel and Pushpak Bhattacharyya. 2015. Detection of multiword expressions for hindi language using word embeddings and wordnet-based features .

Glenn E. Patton. 2009. *Functional requirements for authority data: A conceptual model*, volume 34. Walter de Gruyter.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *IJCNLP*. pages 210–218.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *EACL*. pages 472–481.

Max Silberztein. 2008. Nooj v2 manual.

Simona Vietri. 2014. The italian module for nooj. In *Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it*.

Simona Vietri and Mario Monteleone. 2013. The english nooj dictionary. In *Proceedings of NooJ 2013 International Conference, June*. pages 3–5.