VP-Explore:
Verbal Phrase semantically explained

Enrico Mensa

Dipartimento di Informatica – Università di Torino, Italy
mensa@di.unito.it

Introduction

One long-standing assumption in the Natural Language Processing (NLP) field is that lexicon provides a rich —amenable to automatic analysis— interface between natural language syntax and semantics [27,4]. Since verb is the most relevant and informative lexical category, verb analysis is acknowledged to be a central and challenging task in the computational linguistics investigation [11,26,24].

The following sentence

John ate a salad at the restaurant, where he met a colleague.

can be analyzed at *syntactic* and *semantic* level. In the former case a parse tree is computed, to illustrate the dependency structure underlying the input sentence (i.e., verb, subject, direct object, etc.). In the latter case, by building on top of the syntactic analysis, we are interested in individuating entities, their meaning and mutual relations: thanks to the semantic analysis, we can argue that John is a person, salad is a food, restaurant is a place. However, further —more general— information is conveyed although not explicitly stated, such as that restaurants allow meeting other people, or that people eat. Since such knowledge is plugged into verbs and their argument structure, resources that encode verb's semantics are required to grasp these pieces of information.

Several attempts aimed at developing this type of resources have been made. For example, VerbNet is grounded on one of the most relevant contributions on verbs and their structure [11]. These investigations categorize verbs based on syntactical features, and identify some main classes, hierarchically organized, at the heart of VerbNet by collecting syntactic and semantic information anchored to WordNet sense IDs [26].[1]

Other proposals rely on *semantic roles*, that express the abstract role of predicate arguments [10,25]: e.g., in the sentence 'Maria saw the lion', Maria is 'agent' and lion 'theme'. The consideration of semantic roles under different perspectives led to FrameNet [2] and PropBank [23]. While the former resource is centered on the more abstract notion of frame [20], the latter one —conceived for

[1] WordNet (WN) is a lexical database for the English language [19]. Concepts are represented as nodes in a large semantic network, where the intervening edges represent semantic relations among concepts (such as hyponymy, hypernymy, etc.).

the semantic role labeling task [24]—, is a verb-oriented resource that basically adopts syntactic annotations.

Existing resources and approaches worked out relevant solutions, but they overall suffer from two chief issues: coverage and semantic grounding. The aforementioned resources do not provide adequate *coverage*, they are *not grounded* on a shared set of meaning identifiers, and they *only employ symbolic knowledge* to enrich verbal information, thereby resulting to be limited as regards as the application domain, and restricted to very narrow tasks. The lack of such resources is even more severe for the Italian language, where far less efforts have been spent than for English. These open problems are the focus of the VP-Explore project: devising such resources would be relevant for NLP at large, and would impact on different contexts of application, ranging from text categorization to automatic generation of metadata, detection of text genre, plagiarism detection, automatic summarization, question answering, etc..

The VP-Explore project

My solution starts from the extraction of *subcategorization frames* —encoding the set of syntactic arguments that are either allowed or required by a given verb— from large, syntactically annotated corpora, and proceeds by subsequently enriching them with semantic information.

Task 1: Building resources annotated with syntactic information and word sense disambiguation. As anticipated, a corpus syntactically annotated with dependency structure will be processed to build our subcategorization frames. The multilingual *corpora* collected in the *WaC* project [3] will be employed to extract information for Italian/English verbs in order to build two resources: one inspired to VerbNet but with broad coverage, and a brand new one for the Italian language. One main requirement is that information collected in the subcategorization frames must also account for semantic elements. That is, all information harvested from the chosen *corpora* will be disambiguated (through a word sense disambiguation (WSD) step [21]), and provided with BabelNet [22] meaning identifiers. The intended senses are indicated in the following examples through the subscripts: e.g., by $\texttt{lion}_1$ we denote the first sense listed in BabelNet, referred to the animal rather than to the Zodiac sign, or the Mac OS X Operating System.

This processing step will extract a KB consisting of a set of subcategorization frames, one for each verb sense encountered in the corpus:

$$\begin{aligned}
&\texttt{eat}_1(\underbrace{\texttt{lion}_1}_{subj},\ \underbrace{\texttt{gazelle}_1}_{obj}, \texttt{...}) \\
&\texttt{see}_1(\texttt{John}_1, \texttt{palace}_4, \texttt{...}) \\
&\texttt{play}_2(\texttt{Maria}_1, \texttt{guitar}_1, \texttt{...}) \\
&\vdots
\end{aligned}$$

The first line in the example reports that, in one sentence of the corpus, ‘lion’ has been found to be the subject of an event ‘eat’, having ‘gazelle’ as direct object.

Task 2: Building resources annotated with semantic information. The obtained resources will be enriched with semantic information, by providing subcategorization frames with supersense tags.[2] Such task will be performed by clustering all the records extracted for each verb into one single record for each *sense* of usage of the given verb (the subscript denotes the *sense* identified by the BabelNet synset ID, as mentioned above). This step involves listing all the legitimate supersenses corresponding to each syntactic role, also considering its relative frequency:

$$\texttt{eat}_1\texttt{(}\underbrace{\texttt{\{[person,0.4],[animal,0.3],...\}}}_{subj}\texttt{,}\underbrace{\texttt{\{[food, 0.9],...\}}}_{obj}\texttt{,...)}$$

$$\texttt{see}_1\texttt{(\{[person,0.8],...\},\{[artifact,0.4],...\},...)}$$

$$\vdots$$

Then, the extracted features will be further abstracted by lifting the supersenses to the corresponding semantic roles, that is pieces of information such as ‘agent’, ‘patient’, ‘theme’, ‘location’, *etc.*. The set of semantic roles proposed by [25] will be adopted initially, and possibly enriched with further sources.

Task 3: Semantic annotation with Conceptual Spaces. The Conceptual Spaces (CS) framework will be adopted to encode the semantics of the enriched subcategorization frames in close conjunction with [14,16,12]. CSs are a particular class of vector representations where knowledge is represented as a set of quality dimensions, each based on a geometrical or topological structure [8], e.g., a *color* can be characterized by 3 dimensions: brightness, saturation and hue. In this framework, that can be considered as intermediate between the symbolic level and the sub-symbolic level [7], concepts correspond to convex regions, and regions with different geometrical properties correspond to different sorts of concepts.

Task 4: Evaluation, dissemination and applications. The final activities in the VP-Explore project will include experimenting on the released resources, disseminating the results in different application settings, and preparing applications to research projects at both the national and EU levels, where close themes [1].

[2] Supersense tags are high-level concepts such as ‘person’, ‘artifact’; a popular set of supersenses includes the 26 top nouns in the WordNet hierarchy [5].

Working with nouns and common-sense: the COVER resource

Verbs are known to be the most polysemous part of speech [19], and the resources produced in the VP-Explore project will be helpful in taming this aspect of language. One main hypothesis of this work is that the overall meaning of a verb is to a good extent determined by the dependents of the verb itself.

To these ends, providing verb dependents with fully fledged semantic description was a precondition. The main outcome of the first year of my PhD course is precisely the development of COVER: COVER is a lexical resource providing common-sense knowledge on nouns, and it is automatically generated by combining the lexicographic precision propert to BabelNet and the common-sense hosted in ConceptNet [9]. The COVER is a vectorial resource that has been designed to be compatible with the constructive underlying the CSs [18]: it provides a conceptual representation for the main senses associated to the $60K$ most common English terms [13]; the COVER has been employed to compute conceptual similarity [17], to perform the conceptual categorization task [15], and in the task of keywords extraction [6].

References

1. European cultural heritage, access and analysis for a richer interpretation of the past. Topic identifier: CULT-COOP-09-2017. `http://goo.gl/iA4Ebs`, 2016.
2. Collin F Baker, Charles J Fillmore, and John B Lowe. The Berkeley Framenet Project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. Association for Computational Linguistics, 1998.
3. Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky Wide Web: a Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.
4. Wei-Te Chen, Claire Bonial, and Martha Palmer. English Light Verb Construction Identification Using Lexical Knowledge. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
5. Massimiliano Ciaramita and Mark Johnson. Supersense Gagging of Unknown Nouns in WordNet. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 168–175. Association for Computational Linguistics, 2003.
6. Davide Colla, Enrico Mensa, and Daniele P. Radicioni. Document Retrieval for Question Answering. (to appear).
7. Peter Gärdenfors. Conceptual spaces: The geometry of thought. a bradford book. *MIT Press*, 3:16, 2000.
8. Peter Gärdenfors. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press, 2014.
9. Catherine Havasi, Robert Speer, and Jason Alonso. ConceptNet: A lexical resource for common sense knowledge. *Recent advances in natural language processing V: selected papers from RANLP*, 309:269, 2007.
10. Dan Jurafsky. *Speech & Language Processing*. Pearson Education, 2000.

11. Beth Levin. *English Verb Classes and Alternations: a Preliminary Investigation.* 1993.
12. Antonio Lieto, Enrico Mensa, and Daniele P. Radicioni. A Resource-Driven Approach for Anchoring Linguistic Resources to Conceptual Spaces. In *XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 December 1, 2016, Proceedings*, volume 10037 of *Lecture Notes in Artificial Intelligence*, pages 435–449. Springer, 2016.
13. Antonio Lieto, Enrico Mensa, and Daniele P. Radicioni. Taming sense sparsity: a common-sense approach. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.*, 2016.
14. Antonio Lieto, Daniele P. Radicioni, and Valentina Rho. A Common-Sense Conceptual Categorization System Integrating Heterogeneous Proxytypes and the Dual Process of Reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 875–881, Buenos Aires, July 2015. AAAI Press.
15. Antonio Lieto, Daniele P. Radicioni, Valentina Rho, and Enrico Mensa. Towards a Unifying Framework for Conceptual Represention and Reasoning in Cognitive Systems. *Intelligenza Artificiale*, 2017 (to appear).
16. Enrico Mensa. Design and Implementation of a Methodology for the Alignment of Semantic Resources and the Automatic Population of Conceptual Spaces. Master's thesis, Università degli Studi di Torino, 2016.
17. Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. Merali at semeval-2017 task 2 subtask 1: a cognitively inspired approach. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 236–240, Vancouver, Canada, August 2017. Association for Computational Linguistics.
18. Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. TTCS$^{\mathcal{E}}$: a Vectorial Resource for Computing Conceptual Similarity. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 96–101, Valencia, Spain, April 2017. Association for Computational Linguistics.
19. George A Miller. WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.
20. Marvin Minsky. A framework for representing knowledge. 1975.
21. Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.
22. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250, 2012.
23. Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput. Linguist.*, 31(1):71–106, March 2005.
24. Martha Palmer, Daniel Gildea, and Nianwen Xue. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103, 2010.
25. Volha Petukhova and Harry Bunt. Lirics semantic role annotation: Design and evaluation of a set of data categories. In *LREC*, 2008.
26. Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon.* PhD thesis, Philadelphia, PA, USA, 2005. AAI3179808.
27. Zhibiao Wu and Martha Palmer. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.