Automated Acquisition of Multiword Expressions for Robust Deep Parsing

Valia Kordoni
Dept. of English, Humboldt-Universität zu Berlin, Germany
kordonie@anglistik.hu-berlin.de
(for inclusion in the program of WG2)

Abstract

In this presentation, I mainly deal with automated acquisition of Multiword Expressions as a means of enhancing robustness of lexicalised grammars used in robust deep parsing for real-life applications.

Specifically, I begin by taking a closer look at the linguistic properties of MWEs, in particular, their lexical, syntactic, as well as semantic characteristics. The term Multiword Expressions has been used to describe expressions for which the syntactic or semantic properties of the whole expression cannot be derived from its parts (cf., Sag et al., 2002), including a large number of related but distinct phenomena, such as phrasal verbs (e.g., "come along"), nominal compounds (e.g., "frying pan"), institutionalised phrases (e.g., "bread and butter"), and many others. Jackendoff (1997) estimates the number of MWEs in a speaker's lexicon to be comparable to the number of single words.

However, due to their heterogeneous characteristics, MWEs present a tough challenge for both linguistic and computational work (cf., Sag et al., 2002). For instance, some MWEs are fixed, and do not present internal variation, such as "ad hoc", while others allow different degrees of internal variability and modification, such as "spill beans" ("spill several/musical/mountains of beans"). With the observations about the linguistic properties of MWEs at hand, I turn to methods for the automated acquisition of these properties for robust deep parsing. To this effect, I first investigate the hypothesis that MWEs can be detected by the distinct statistical properties of their component words, regardless of their type, comparing various statistical measures, a procedure which leads to extremely interesting conclusions. I then investigate the influence of the size and quality of different corpora, using the BNC and the Web search engines Google and Yahoo. I conclude that, in terms of language usage, web generated corpora are fairly similar to more carefully built corpora, like the BNC, indicating that the lack of control and balance of these corpora are probably compensated by their size.

Then, I show a qualitative evaluation of the results of automatically adding extracted MWEs to existing linguistic resources. To this effect, I first discuss two main approaches commonly employed in NLP for treating MWEs: the words-with-spaces approach which models an MWE as a single lexical entry and it can adequately capture fixed MWEs like "by and large", and compositional approaches which treat MWEs by general and compositional methods of linguistic analysis, being able to capture more syntactically flexible MWEs, like "rock boat", which cannot be satisfactorily captured by a words-with-spaces approach, since this would require lexical entries to be added for all the possible variations of an MWE (e.g., "rock/rocks/rocking this/that/his…boat"). On this basis, I argue that the process of the automatic addition of extracted MWEs to existing linguistic resources improves qualitatively, if a more compositional approach to grammar/lexicon automated extension is adopted.

Finally, I also propose that the methods developed for the acquisition of linguistic knowledge in the case of the English MWEs can be tuned to enhance robustness of lexicalised grammars in parsing languages with richer morphology and freer word order, as is the case of German, and can benefit from gold standard syntactically and semantically annotated corpora, for the (semi-automated)

development of which I am briefly showing a very simple statistical ranking model which significantly improves treebanking efficiency by prompting human annotators to the most relevant linguistic annotation decisions.

References

Ray Jackendoff. 1997. Twistin' the night away. *Language*, 73:534–59.

Carlos Ramisch, Aline Villavicencio and Valia Kordoni (eds.). 2013. *ACM Transactions on Speech and Language Processing (TSLP) - Special issue on multiword expressions: From theory to practice and use.* Part 1, Volume 10, Issue 2, June 2013.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Kristina Toutanova, Christoper D. Manning, Stuart M. Shieber, Dan Flickinger, and Stephan Oepen. 2002. Parse ranking for a rich HPSG grammar. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pages 253–263, Sozopol, Bulgaria.

Gertjan van Noord. 2004. Error mining for wide-coverage grammar engineering. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, Main Volume, pages 446–453, Barcelona, Spain, July.

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*, The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1034-1043, Prague, June 28-30, 2007.

Yi Zhang and Valia Kordoni. 2006. Automated deep lexical acquisition for robust open texts processing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.