Identifying Multi-Word Expressions in Large Treebanks with Tree Kernels

Federico Sangati
University of Edinburgh (UK) & FBK, Trento (Italy)
federico.sangati@gmail.com

Andreas van Cranenburgh
Royal Netherlands Academy of Arts & Sciences & Univ. of Amsterdam
andreas.van.cranenburgh@huygens.knaw.nl

Working groups concerned:

WG3: Statistical, Hybrid and Multilingual Processing of MWEs

WG4: Annotating MWEs in Treebanks

Abstract

In many current linguistic theories, language users produce and understand sentences without necessarily decomposing them into just 'words' and 'rules'; rather, multi-word units may function as the elementary building blocks (Goldberg, 1995; Kay and Fillmore, 1997; Stefanowitsch and Gries, 2003). A growing literature is emerging which focuses on "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002) also referred to as multi-word expressions (MWEs) . An important question for computational linguistics is how to identify such building blocks using statistical regularities in large corpora (Zuidema, 2006; Ramisch et al., 2012).

In our work, we investigate ways of automatically detecting MWEs in large treebanks using Tree Substitution Grammars (TSGs) (Bod et al., 2003). In a TSG, the symbolic grammar is a bag of fragments which are arbitrarily large syntactic constructions extracted from a treebank. They can include any number of lexical units, with possible intervening gaps, and are therefore very suitable to represent MWEs ranging from fixed idiomatic cases such as "kick the bucket" to more flexible expressions such as "break X up" or even longer constructions such as "everything you always wanted to know about X but were afraid to ask."

Since extracting all possible fragments from a large treebank is impossible (the number of possible fragments grows exponentially with the size of a tree) it is necessary to work with a restricted set of fragments. Several sampling methods have been proposed (Bod, 2001; Zuidema, 2007; Cohn et al., 2010), but all include some limitations (e.g., use of random sampling methods, restriction in the size of the fragments, number of lexical items).

For extracting the fragments, we choose to employ `FragmentSeeker`[1] (Sangati et al., 2010), which considers only those fragments which occur two or more times in the treebank. This is an ideal constraint if we want to assume that a necessary condition for a fragment to yield a MWE is to *recur multiple times in a representative corpus*. This is also one of the original motivations behind the Data-Oriented Parsing (DOP) framework (Bod, 1992) based on TSGs, in which "idiomaticity is the rule rather than the exception" (Scha, 1990). For instance, if we have seen the MWE "pain in the neck" several times before, we should store the whole fragment for later use.

`FragmentSeeker` is based on an efficient tree-kernel dynamic programming algorithm, which compares every pair of trees of a given treebank and computes the list of fragments which they have in common. This algorithm is an extension of previous work on tree kernels (Collins and Duffy, 2001). While in the original work kernels are used to numerically quantify the similarity between two trees, in the current project the algorithm identifies the actual constructions they share, i.e. the recurring fragments. However, the complexity of the extraction algorithm is quadratic in the size of the treebank. In a recent effort, van Cranenburgh (2014)

[1]The tool is publicly available at `http://http://homepages.inf.ed.ac.uk/fsangati`

developed an improved algorithm[2] for fragment extraction which has linear running time in the size of the treebank (it runs 70 times faster than the original implementation on a 40K sentences corpus). This substantial speedup is due to the incorporation of the Fast Tree Kernel (Moschitti, 2006), and opens up the possibility of handling much larger corpora.
The set of fragments extracted with this tool has proven to be successful for several NLP tasks such as statistical parsing, as in DOP (Sangati and Zuidema, 2011; van Cranenburgh and Bod, 2013), authorship attribution (van Cranenburgh, 2012), and native language detection (Swanson and Charniak, 2012).
However, the extraction tool has so far been used in medium-large corpora (up to 50K sentences) which are not big enough to cover a wide range of MWEs.
For this project, we would like to use the Annotated English Gigaword treebank[3] which contains more than 180 million sentences. Such a size is still prohibitively large even for the fast version of `FragmentSeeker`. But, since our target here are MWEs, we are only interested in lexicalized fragments with at least two lexical items. Such restriction suggests a further optimization that could substantially boost the extraction speed: after indexing sentences by the words they contain, we compare every tree structure only to other structures sharing at least two words.

We are planning to apply the same extraction algorithm on very large treebanks as a first step to developing methods for automatically identifying MWEs. An initial encouraging result in this respect is the work of Green et al. (2011) where the authors obtained a 36.4% F1 absolute improvement in MWE identification using a TSG parser over an n-gram surface statistics baseline (Ramisch et al., 2010). However, one needs to note that the French Treebank (Abeillé et al., 2003) used in this study, contains explicitly tagged MWEs (as a special phrasal category), and therefore the comparison between syntax-aware (TSG fragments) and surface-based methods (n-grams) is not entirely fair. It will be our quest to investigate if an improvement can be obtained in a more unsupervised fashion.

We believe that this line of research could be potentially beneficial to the following PARSEME working groups:

WG3 Recurring fragments can be used for MWE-informed statistical parsing approach, e.g., the DOP framework.

WG4 Automatically derived MWEs, enriched with their syntactic structures, can be employed to automatically label existing treebank with MWE-informed tags, and can lead to the creation of resources such as MWE lexicons and valence dictionaries.

[2]The tool is publicly available at `https://github.com/andreasvc/disco-dop`
[3]See `http://catalog.ldc.upenn.edu/LDC2012T21`

References

Abeillé, Anne, Lionel Clément, and François Toussenel. *Building a Treebank for French*, volume 20 of *Text, Speech and Language Technology*, pages 165–188. Springer, 2003.

Bod, Rens. A Computational Model of Language Performance: Data Oriented Parsing. In *Proceedings COLING'92 (Nantes, France)*, pages 855–859. Association for Computational Linguistics, Morristown, NJ, 1992.

Bod, Rens. What is the Minimal Set of Fragments that Achieves Maximal Parse Accuracy? In *Proceedings ACL-2001*. Morgan Kaufmann, San Francisco, CA, 2001.

Bod, Rens, Khalil Sima'an, and Remko Scha. *Data-Oriented Parsing*. University of Chicago Press, Chicago, IL, USA, 2003.

Cohn, Trevor, Phil Blunsom, and Sharon Goldwater. Inducing Tree-Substitution Grammars. *Journal of Machine Learning Research*, 11:3053–3096, 2010.

Collins, Michael and Nigel Duffy. Convolution Kernels for Natural Language. In Dietterich, Thomas G., Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*, pages 625–632. MIT Press, 2001.

Goldberg, A.E. *Constructions: A Construction Grammar Approach to Argument Structure*. University Of Chicago Press, 1995.

Green, Spence, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 725–735, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

Kay, Paul and Charles J. Fillmore. Grammatical Constructions and Linguistic Generalizations: the What's X Doing Y? Construction. *Language*, 75: 1–33, 1997.

Moschitti, Alessandro. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In *ECML*, pages 318–329, Berlin, Germany, September 2006. Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Proceedings.

Ramisch, Carlos, Aline Villavicencio, and Christian Boitet. mwetoolkit: a framework for multiword expression identification. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association, 2010.

Ramisch, Carlos, Vitor De Araujo, and Aline Villavicencio. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proceedings of ACL 2012 Student Research Workshop*, ACL '12, pages 1–6, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Sag, IvanA., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In Gelbukh, Alexander, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg, 2002.

Sangati, Federico and Willem Zuidema. Accurate Parsing with Compact Tree-Substitution Grammars: Double-DOP. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 84–95, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

Sangati, Federico, Willem Zuidema, and Rens Bod. Efficiently Extract Recurring Tree Fragments from Large Treebanks. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

Scha, Remko. Taaltheorie en taaltechnologie: competence en performance. In de Kort, Q. A. M. and G. L. J. Leerdam, editors, *Computertoepassingen in de Neerlandistiek*, LVVN-jaarboek, pages 7–22. Landelijke Vereniging van Neerlandici, Almere, 1990. [Language theory and language technology: Competence and Performance] in Dutch.

Stefanowitsch, Anatol and Stephan Th. Gries. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8:209–243, 2003.

Swanson, Benjamin and Eugene Charniak. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,

pages 193–197, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

van Cranenburgh, Andreas. Literary authorship attribution with phrase-structure fragments. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 59–63, Montréal, Canada, June 2012. Association for Computational Linguistics.

van Cranenburgh, Andreas. Linear average time extraction of phrase-structure fragments. In *The 24th Meeting of Computational Linguistics in The Netherlands (CLIN 2014)*, 2014.

van Cranenburgh, Andreas and Rens Bod. Discontinuous parsing with an efficient and accurate dop model. In *Proc. of the 13th International Conference on Parsing Technologies*, 2013.

Zuidema, Willem. What are the productive units of natural language grammar?: a DOP approach to the automatic identification of constructions. In *CoNLL-X '06: Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 29–36, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

Zuidema, Willem. Parsimonious Data-Oriented Parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 551–560, Prague, Czech Republic, June 2007. Association for Computational Linguistics.