Multiword Expression Identification with Recurring Tree Fragments and Association Measures

PARSEME-5 – WG3

Federico Sangati
Fondazione Bruno Kessler (FBK)
Trento, Italy
sangati@fbk.eu

Andreas van Cranenburgh
Huygens ING, Royal Netherlands Academy of Arts & Sciences; ILLC, Univ. of Amsterdam.
andreas.van.cranenburgh@huygens.knaw.nl

Abstract[1]

In the current work, we present a novel approach for the identification of multiword expressions (MWEs). The methodology extracts a large set of recurring syntactic fragments from a given treebank using a Tree-Kernel method (Collins and Duffy, 2002; Sangati et al., 2010). Differently from previous studies, the expressions underlying these fragments are arbitrarily long and can include intervening gaps. We are using three different treebanks for extracting MWEs across three languages: the French Treebank (Abeillé et al., 2003), the Dutch LASSY Small treebank (Noord, 2009), and sample of the Annotated English Gigaword treebank[2]. See table 1 for statistics on treebank sizes and number of fragments, and figure 1 for a comparison of the MWE annotations in the treebanks.

Treebank	Trees	Total Frags	Selected Frags
French	13K	274K	86K
Dutch	52K	536K	193K
English	500K	4.3M	2.8M

Table 1: Treebank size and number of fragments extracted and employed in the experiments. The last column reports the number of fragments after filtering out all those which do not contain at least a content word and a non-punctuation word.

In the initial study we use recurring fragments to identify MWEs as a parsing task (in a supervised manner) as proposed by Green et al. (2011). We use the Double-DOP (2DOP) model (Sangati and Zuidema, 2011), as implemented in the `disco-dop` parser (van Cranenburgh and Bod, 2013). Here we obtain a small but significant improvement over previous results (see table 2).

Parser	F1	EX	MWE-F1
FRENCH			
Green et al. (2013): DP-TSG	76.9	16.0	71.3
Green et al. (2013): Stanford	79.0	17.6	70.5
disco-dop, 2DOP	**79.3**	**19.9**	**71.9**
DUTCH			
disco-dop, PCFG baseline	63.9	21.8	50.4
disco-dop, 2DOP	**77.0**	**35.2**	**75.3**

Table 2: Performance of the parsing models on the French and Dutch treebanks, with respect to parsing results (F1 score and exact match) and the MWE-F1 score, for sentences ≤ 40 words.

In the second study we define an unsupervised method for MWEs identification using both the set of recurring syntactic fragments and various association measures (AMs). We define a new AM (Log Inside Ratio), which specifies the probability that a PTSG grammar generates a given fragment in a single step with respect to the total probability of generating it in any possible way, i.e., by combining smaller fragments together.

We show how this newly defined measure obtains competitive results when compared against other classical association measures: Pointwise Mutual Information (PMI) and Log-Likelihood Ratio (LLR). See table 3 for the results details.

Treebank	PMI	LLR	LIR
French	33.0	32.3	45.8
Dutch	49.4	46.6	50.5

Table 3: F1 scores for the top 1/5 candidates of each bin as ranked by the three AMs evaluated against MWEs in extracted recurring fragments.

Acknowledgement We greatly acknowledge the PARSEME IC1207 COST Action for supporting this work. We also thank the three anonymous reviewers for the very useful comments.

[1] For the full version of this paper please see Sangati and van Cranenburgh (2015).

[2] `http://catalog.ldc.upenn.edu/LDC2012T21`

(a) French treebank

lit.: *on the hand*, "going on."

(b) Dutch Lassy treebank

(c) Annotated English Gigaword

Figure 1: A comparison of treebanks and their MWE annotation. (a) French treebank; flat MWE annotation. (c) Dutch Lassy treebank; flat MWE annotation. (b) Annotated English Gigaword; no MWE annotation.

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.

Collins, Michael and Nigel Duffy. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 263–270, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

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.

Green, Spence, Marie-Catherine de Marneffe, and Christopher D. Manning. Parsing models for identifying multiword expressions. *Comput. Linguist.*, 39(1):195–227, March 2013.

Noord, Gertjan Van. Huge parsed corpora in lassy. In *TLT7*, Groningen, Netherlands, 2009.

Sangati, Federico and Andreas van Cranenburgh. Multiword Expression Identification with Recurring Tree Fragments and Association Measures. In *Proceedings of the Workshop on Multiword Expressions: MWE 2015 (NAACL)*, 2015.

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).

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.