Extracting MWE and fixed structures from parsed corpora

Gerold Schneider, University of Zurich

We apply collocation measures followed by manual filtering to automatically extract several categories of multi-word entities from large parsed corpora (Schneider 2008, Lehmann and Schneider 2012). The dependency parser which we use combines a hand-written *competence* grammar with statistical *performance* disambiguation, using a bi-lexical maximum-likelihood model (e.g. Collins 1999) trained on the Penn Treebank.

The categories of multi-word entities we extract are in particular:

- verb-object idioms (such as *kick the bucket*)
- verb-preposition structures (such as *take into consideration*)
- light-verb constructions (such as *take a decision*)

We compare several collocation measures (such as T-score and O/E), re-confirm that parsed data leads to cleaner results (Seretan 2011) and add Yule's K (Yule 1944) as a measure of fixedness, which is another MWE characteristic. We offer these resources to the research community.
As MWE and collocations prevail at all levels, we suggest to use information-theoretic measures like surprisal (Levy and Jaeger 2007) as a measure of fixedness and entrenchment. We discuss that parsing can be seen as a tug-of-war between the idiom principle and the syntax principle (Sinclair 1991). The idiom principle represents the fixedness aspect (e.g. Tomasello 2000), the syntax principle the creative use of syntactic rules which over-generates when parsing unless it is constrained by bi-lexical preferences or other fixedness aspects. Finally we suggest the use of parsers as psycholinguistic models of fixedness and entrenchment vs. creativity (e.g. Keller 2010).

References

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Lehmann, Hans Martin and Gerold Schneider. 2012. A large dependency bank. In *Proceedings of LREC 2012 Workshop on Challenges in the management of large corpora*, pages 23–28.

Keller, Frank. 2010. Cognitively plausible models of human language processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 60–67.

Levy, Roger and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Seretan, Violeta. 2011. *Syntax-Based Collocation Extraction*. Springer, Dordrecht.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. OUP, Oxford.

Tomasello, Michael. 2000. The item based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4:156–163.

Yule, George U. 1944. *The statistical study of literary vocabulary*. Cambridge University Press, Cambridge.