Using Dependency-Annotated Corpora to Improve Collocation Extraction

Proisl, Thomas (FAU Erlangen-Nürnberg)

Uhrig, Peter (FAU Erlangen-Nürnberg)

The extraction of collocations from corpora is of central importance for linguists working on such diverse topics as foreign language lexicography (see for instance Sinclair 1987), language learning (Nesselhauf 2005), linguistic varieties (e.g. Schilk 2006), or psycholinguistic experiments (e.g. Ellis/Frey 2009, Ellis et al. 2009). For all these researchers, it is desirable to have access to "good" lists of collocates for a given word. We shall demonstrate in our presentation that the use of dependency-annotated corpora leads to substantially improved results.

Over the past few years, dependency grammar has gained more and more support in Natural Language Processing (NLP) so that both native dependency parsers as well as systems for the conversion of phrase structures into dependency structures are available today. Since the idea of a dependency analysis is to make explicit the syntactic-semantic relations between the word forms in a sentence, it is perfectly suited to the automatic extraction of collocations and other multi-word units (see for instance Weller/Heid 2010). Dependency annotation allows us to restrict the analysis to only those co-occurring items that are related and thus to exclude all items that are only found in the direct linguistic context by coincidence. On the other hand it can also find long-distance dependencies identified by the parser that are found outside a window of n (typically 5) tokens to either side.

For English, the Stanford Dependencies Model (de Marneffe et al. 2006, de Marneffe/Manning 2008) is a common annotation scheme in the area of Natural Language Processing (see Cer et al. 2010). It comes with a free piece of software[1] that can convert Penn-Treebank-style phrase structure trees (which are still a sort of de-facto standard in NLP) into dependency structures.

In our presentation we will compare our mechanism for the extraction of collocations on the basis of dependency relations to the established window-based approach. The comparison will be based on the parsed version of the British National Corpus (BNC) of the Treebank.info project (Uhrig/Proisl 2011). The results are then evaluated against the Oxford Collocations Dictionary for Students of English (2nd edition). We will show that both precision

[1] http://nlp.stanford.edu/software/lex-parser.shtml

and recall can be improved significantly with our method; thus it can provide both shorter and more accurate lists to researchers.

We will conclude our paper by explaining how researchers can upload their own corpora to the Treebank.info platform and obtain dependency-based collocation lists of their own data. The presentation of the data can be arranged according to the dependency relation that connects the pairs of words in a way similar to the presentation given by the Sketch Engine (Kilgarriff et al. 2004), but based on a full syntactic parse.

References

Cer, Daniel, Marie-Catherine de Marneffe, Daniel Jurafsky & Christopher Manning. 2010. Parsing to Stanford Dependencies: Trade-offs between speed and accuracy. *LREC 2010*, Valletta.

de Marneffe, Marie-Catherine, Bill MacCartney & Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. *LREC 2006*, Genoa.

de Marneffe, Marie-Catherine & Christopher D. Manning. 2008. The Stanford typed dependencies representation. *COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation*. Manchester. 1-8.

Ellis, Nick C. & Eric Frey. 2009. The Psycholinguistic Reality of Collocation and Semantic Prosody (2): Affective Priming. In Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali & Kathleen Wheatley (eds.), *Formulaic Language*, 473-497. Amsterdam: Benjamins.

Ellis, Nick C., Eric Frey & Isaac Jalkanen. 2009. The Psycholinguistic Reality of Collocation and Semantic Prosody (1): Lexical Access. In Ute Römer & Rainer Schulze (eds.), *Exploring the Lexis-Grammar Interface*, 89-114. Amsterdam: Benjamins.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz & David Tugwell. 2004. The Sketch Engine. In *Proceedings Euralex 2004*, 105-116. Lorient, France. http://www.sketchengine.co.uk.

Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus.* Amsterdam: Benjamins.

Oxford Collocations Dictionary for Students of English. 2009. Oxford: OUP.

Schilk, Marco. 2006. Collocations in Indian English – a corpus-based sample analysis. *Anglia* 124(2). 276-316.

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

Uhrig, Peter, Thomas Proisl. 2011. *The Treebank.info project.* Presentation given at ICAME 32, Oslo, 4 June 2011. http://treebank.info.

Weller, Marion & Ulrich Heid. 2010. Extraction of German Multiword Expressions from Parsed Corpora Using Context Features. *LREC 2010*, Valletta. 3195-3201.