Abs-207

Peter Uhrig and Thomas Proisl (Universität Erlangen-Nürnberg)

A fast and user-friendly interface for large treebanks

Although parsed corpora have been around for a long time, many are either comparatively small (e.g. ICE-GB) and/or come without an easy-to-use interface (e.g. Penn Treebank). The unintuitive nature of powerful corpus query tools such as TGrep2 has led researchers to create applications such as SearchTree (Nygaard/Bondi Johannessen 2004) or Treebank Search (Ghodke/Bird 2010). Due to their online interfaces, these tools provide considerable improvements in usability. However, they rely on Penn Treebank style phrase structure trees, which makes queries cumbersome for users who are used to more traditional labels such as “subject” or “direct object”. We claim that a graphical interface (without a query language) with relatively simple labels of that kind is needed to lower the inhibitions of less technically minded linguists to use treebanks in their research. The most widespread and probably the most intuitive system of such labels is currently provided by Stanford Dependencies (SD, de Marneffe/Manning 2008). A big advantage of the SD framework is that it is built to read Penn Treebank style phrase structure trees. That means that an SD representation can be obtained from dependency parsers trained on converted treebanks as well as from the output of many currently available phrase structure parsers (see Cer et al. 2010 for a comparison of speed and accuracy).

The dependencies do not necessarily form a tree structure, so they have to be represented more generally as a directed (possibly cyclic) graph. This is necessary to allow for instance for an object to depend on two coordinated verbs, and similar structures. Existing tree structure storage mechanisms thus cannot be used in our system. An overview of a high-performance system based on standard software and specifically designed for this purpose will be given in the paper.

In addition to finding concordance lines, the software offers the possibility to perform a collostructional analysis (Stefanowitsch/Gries 2003) of the collexeme variety (Stefanowitsch/Gries 2009:941). Thus it is possible to determine the association strength between (optionally partially lexically filled) syntactic structure in the form of a dependency graph and word forms or lemmata. For instance, we can look for the lemma “give” with a direct and indirect object node dependent on it where we only specify part-of-speech for the nodes, namely PRP (personal pronoun) for the indirect object and NNS (noun in plural) for the direct object. The direct object in turn has a dependent of the type determiner which is lexically filled by “the”. The collexeme searched for is the direct object plural noun. In the BNC, the most strongly associated form (though not the most frequent one) for this structure is “creeps” and while there are frequent literal examples (details, keys, …) the list contains alternatives to “creeps” in a similar meaning, such as willies, shivers, horrors, jitters sorted by association measure.

The paper will include a live demonstration.

References:

Cer, D. / M.-C. de Marneffe / D. Jurafsky / C. Manning (2010): “Parsing to Stanford Dependencies: Trade-offs between speed and accuracy”, in: Proceedings of the 7th Conference on International Language Resources and Evaluation, Valletta, Malta (LREC 2010).

de Marneffe, M. / C. Manning (2008): “The Stanford typed dependencies representation”, in: COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation.

Ghodke, S. / S. Bird (2010): “Fast Query for Large Treebanks”, in: Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association of

Computational Linguistics, Los Angeles, USA, 267–275.

Nygaard, L. / J. Bondi Johannessen (2004): “SearchTree – A User-Friendly Treebank Search Interface”, in: Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004), 183–189.

Stefanowitsch, A. / S. Gries (2003): “Collostructions: Investigating the interaction of words and constructions”, in: International Journal of Corpus Linguistics, 8/2: 209–243.

Stefanowitsch, A. / S. Gries (2009): “Corpora and grammar”, in: A. Lüdeling / M. Kytö (eds.): Corpus Linguistics: An International Handbook. Berlin/New York: Walter de Gruyter, 933–952.