ANNOTATE
A Tool for Interactive Corpus Annotation

Oliver Plaehn
Computational Linguistics
Saarland University
66041 Saarbrücken, Germany
plaehn@coli.uni-sb.de

Annotate provides a comprehensive set of commands for efficient creation and manipulation of syntactic structures. The structures consist of trees with possibly crossing branches, labeled terminal nodes (part-of-speech tags and morphological information), labeled edges (grammatical functions), and labeled non-terminal nodes (phrase categories). In addition, so-called secondary links can be drawn between arbitrary nodes, forming directed graphs. These can be used for incorporating structure sharing information, e.g. for multiple PP attachment, to represent ellipses, or to resolve anaphors. Annotate provides immediate graphical feedback on all changes applied to the syntactic structure. Other features of Annotate include a search function, basic tokenization commands (for splitting or merging words, moving sentence boundaries, etc.), postscript output, an undo function, and online help for the different label sets.

Annotate's most notable feature is its interface to an external parser/tagger, which allows for a semi-automatic annotation procedure. First, the tagger determines a part-of-speech tag for each word of the current sentence. Based on the tags' probabilities, the tagger distinguishes reliable and unreliable assignments. Unreliable assignments are highlighted and the annotator is prompted for confirmation or correction. Second, the syntactic structure is built bottom-up. The parser incrementally suggests new phrases, based on the already constructed partial syntactic structure and the part-of-speech tags. If the suggested phrase is correct, the annotator accepts it. Otherwise, she rejects it and may call the parser for a new suggestion. Alternatively, she may decide to insert a new phrase manually. The internal structures used by the parser are updated, and it again suggests a new phrase based on the existing partial structure. In addition to phrase hypotheses, the parser suggests labels for nodes and edges. Reliable and unreliable assignments are distinguished, and the annotator is prompted for confirmation or correction of the unreliable assignments.

This semi-automatic process facilitates a very rapid annotation. A trained annotator needs on average 50 seconds/sentence (approx. 1,300 tokens/hour) for part-of-speech plus structural annotation in the NEGRA corpus. This is the fastest structural annotation reported in the literature.

Neither the parser/tagger nor Annotate is restricted to a particular tagset or annotation scheme. It is easy to train the tagger and parser on a small amount of annotated data and to subsequently apply them to annotating according to the new scheme. Using the newly annotated data for training improves the parser's performance and thereby constitutes a bootstrapping approach.

Annotate has been implemented in Tcl/Tk and C. A freely available relational database management system (MySQL) is used for storage and retrieval of corpora. In addition, a line-oriented text format has been devised for import/export purposes and for use with standard Unix tools and C programs.

To sum up, our annotation tools facilitate comfortable, customizable and fast semi-automatic part-of-speech and structural syntactic annotation. The tools run under Solaris and Linux and are freely available to universities and similar institutions for research purposes[1].

The talk will include a demonstration of Annotate.

[1] cf. http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate