Treebank Conversion for LTAG Grammar Extraction

Anette Frank
Language Technology Group
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
frank@dfki.de

Treebank Conversion for LTAG Grammar Extraction

Anette Frank, DFKI Saarbrücken, Germany

We present a method for rule-based structure conversion of existing treebanks, which aims at the extraction of linguistically sound, corpus-based grammars. We apply this method to the NEGRA treebank (Skut et al., 1998) to derive an LTAG grammar of German. We describe the methodology and tools for structure conversion and LTAG extraction. The conversion and grammar extraction process imports linguistic knowledge and generalisations that are missing in the original treebank. This supports extraction of a linguistically sound grammar with maximal generalisation and extension to unseen data. On a broader perspective our approach contributes to a better understanding on where corpus linguistics and theoretical syntax can meet and enrich each other.

Treebank conversion for extraction of corpus-based grammars While corpus-linguistic methods extend to many areas studied in theoretical and computational linguistics, the (well motivated) preference for theory-neutral annotations can lead to a gap between corpus-based, statistical approaches and theoretical linguistics, if corpus annotations cannot be mapped, for example, to basic structural assumptions of a particular syntactic framework. Conversion of treebanks towards structural assumptions of specific syntactic theories is intended to bridge this gap.

LTAG grammar extraction is more complex than extraction of (P)CFGs in that the grammar consists of a set of lexicalised *elementary trees*, which encode all arguments of a lexical head as substitution or adjunction nodes, modelling an "extended domain of locality" (Joshi and Schabes, 1997). Since modifiers and recursively embedding structures are represented as *adjunction* trees, they must be factored from flat treebank trees, and rearranged as tree-adjunction structures. Thus, LTAG grammar extraction consists of structure conversion and fragmentation of the restructured corpus trees.

Treebank Conversion is based on a general tree description language (unlike related work in (Xia, 1999)) to allow for flexible and fine-grained definition of declarative conversion rules. This is particularly important in our application of LTAG grammar extraction, given the challenges of German syntax in conjunction with the very flat NEGRA annotations.

We compile the corpus to a constraint language that represents trees in terms of basic predicates for nodes, mother-daughter and precedence relations, and which can be extended to encode grammatical relations, or feature structures (Frank, 2000). Derived description predicates (first/last daughter, transitivity of dominance/ precedence, etc.) allow for concise definition of conversion rules. These consist of a *Rule Id*, a set of *Constraints*, and a set of *Actions*. *Constraints* specify partial configurations by means of tree description predicates. *Actions* specify tree modifications by removing (`-p`), changing, or adding (`+p`) description predicates `p`. Recurrent transformation patterns are pre-defined in generic templates. A rule is recursively applied to *each* partial tree configuration that satisfies the constraints. Conversion rules are stated in a sequence, and apply in a cascade: the output resulting from application of rule $\mathtt{r_i}$ provides the input to the following rule $\mathtt{r_{i+1}}$.

The restructured trees are input to **tree fragmentation rules**. Fragmentation criteria are stated as *Conditions* which refer to categorial and functional annotations, as well as specially induced properties, by importing external linguistic knowledge, where corpus annotations do not provide sufficent distinctions (e.g. transitive vs. modal use of modal verbs, etc). *Actions* are generic templates which cut out auxiliary trees, or cut off subtrees at the specified fragmentation nodes.

Current state Treebank conversion and fragment extraction are not yet completed. They currently comprise 44 conversion and 21 fragmentation rules. From the restructured corpus (10.000 sentences) we extract 113.525 fragments. Out of these, 75.6% are well-typed according to a set of 2155 tree templates (generated from 65 basic tree types).The remaining 24.4% require further restructuring, or are not yet covered by the tree templates.

Rule-based induction of linguistic knowledge The NEGRA corpus provides a highly informative annotation scheme. Functional labels provide general constraints for conversion and fragmentation; refinements are steered by finer categorial distinctions. Yet in many cases annotations are not ideal from a linguistic perspective. We show how conversion rules, by exploiting external linguistic knowledge, can induce missing information from secondary properties encoded in the corpus. We illustrate how to identify German clause types for extraction of linguistically sound LTAG trees. Based on linguistic insight, we further induce missing subjects in VP and SGF coordination structures.

Generalisation and grammar induction Treebank conversion allows for extraction of LTAG grammars with maximal generalisation and maximal coverage on unseen data, in particular by factorising optional constituents. The next step is *grammar induction*. On the basis of *families* of construction-specific tree types we induce unseen tree fragments from constructional occurrences found in the corpus. Morphological generalisations will further extend the coverage of the extracted grammar.

Corpus linguistics meets theoretical linguistics An interesting grey-scale border-line between theoretical and corpus linguistics emerges in this approach to grammar extraction. With continuous extension of conversion and fragmentation rules, the reduced set of non-typed trees moves the border-line between non-classified corpus data and well-typed trees towards a growing, well-defined grammar and a remnant of non-classified corpus trees. At some point this border-line cannot, or only with difficulties, be moved further. The grammar's "complement set", the set of non-typed corpus-trees, could then be considered the target of research in theoretical syntax. At the same time, non-typed tree fragments can live together with well-typed fragments, as regular LTAG grammar components. In this way, corpus-based and theoretical syntax can "meet" in a corpus-derived LTAG grammar.

Topological Field Structures We illustrate the flexibility of our tree conversion method by providing an alternative representation in terms of topological ("field") structures. We selected 13 conversion rules that identify clues for topological structure, and added 8 conversion rules which, based on these clues, transform NEGRA structures to topological field structures. The derived corpus can be used as training material for statistical topological parsing approaches, or evaluation of existing rule-based topological parsers (Neumann et al., 2000).

References

Frank, A. (2000). Automatic F-structure Annotation of Treebank Trees. In Butt, M. and King, T., editors, *Proceedings of the LFG00 Conference*, CSLI Online Publications, Stanford, CA. `http://www-csli.stanford.edu/publications/`.

Joshi, A. and Schabes, Y. (1997). Tree Adjoining Grammars. In Salomma, A. and Rosenberg, G., editors, *Handbook of Formal Languages and Automata*. Springer Verlag, Heidelberg.

Neumann, G., Braun, C., and Piskorski, J. (2000). A Divide-and-Conquer Strategy for Shallow Parsing of German Free Texts. In *Proceedings of ANLP-2000*, pages 239–246, Seattle, Washington.

Skut, W., Brants, T., and Uszkoreit, H. (1998). A linguistically interpreted corpus of german newspaper text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.

Xia, F. (1999). Extracting tree adjoining grammars from bracketed corpora. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99)*, Beijing, China.