Treebank Transformations for Performance Optimizations of a PCFG-based Tagger

Erhard W. Hinrichs and Julia S. Trushkina
Seminar für Sprachwissenschaft, University of Tübingen, Germany
{eh,jul}@sfs.uni-tuebingen.de

In computational linguistics statistical models have become very popular in recent years. One of the most common tasks to which these models have been successfully applied is the assignment of morpho-syntactic categories to words, i.e., morpho-syntactic tagging. For languages with rich inflectional morphology such as German incorporation of case, gender, person and number information into a tagset is of crucial importance. However, it has a drawback of significant expansion of the tagset, which affects tagging accuracy. Moreover, some essential features that supply necessary evidence for correct annotation of inflectional morphology, such as clause type, word order and subject-verb agreement, are of a non-local nature. (Hinrichs and Trushkina, 2003) demonstrate that this non-locality of important relations makes n-gram models insufficient for the given task. A suitable alternative is provided by probabilistic context-free grammars (PCFGs) which take more context into account and can therefore capture long-distance dependencies.

A weakness of PCFGs is the strong independence assumption about the distribution of words and phrases which is inherent in the context-freeness of the underlying grammar formalism. As (Johnson, 1998) has shown, one of the ways to weaken these independence assumptions is by encoding more information in node labels, which can have a dramatic effect on the performance of a PCFG parser.

The purpose of this paper is to investigate the influence of various treebank transformations on the performance of a PCFG tagger. We will show that introduction of pertinent linguistic evidence such as case, number, function and finiteness on node labels and systematic transformations in tree structure lead to a significant improvement of the results.

It is important to note that applying PCFGs to part-of-speech tagging constitutes a rather different task from the ordinary use of PCFGs in probabilistic parsing. For PCFG parsing a sequence of part-of-speech tags is assumed as input, and the nature of the problem is how to project the correct syntactic structure over this given input. By contrast, when PCFGs are used for tagging, the sequence of tags is unknown, and

the task is to find the model of rule probabilities that best predicts the correct part-of-speech sequence. What both tasks have in common is the need to extract those regularities from the given input that are most predictive for the structure to be found. In the case of PCFG parsing, this includes but is not limited to the identification of lexical head-head relationships. In the case of PCFG tagging it is of primary importance to identify those regularities of syntactic structure that are highly predictive for the assignment of the correct tag sequence.

We performed a series of experiments with the PCFG-parser LoPar (Schmid, 2000) trained on the same portion of data with different tree representations and node label encodings. LoPar provides a tagging mode that outputs the sequence of part-of-speech tags with highest estimated frequency, i.e. the product of the inside and outside probabilities of a tag divided by the overall probability of the parse forest. The parser was trained on 51288 manually annotated tokens from the taz newspaper portion of the TüBa-D treebank and tested on 8949 tokens from the same corpus. Additional 5854 lexical tokens served as the development set. A back-up lexicon containing the set of possible morphological analyses for each unknown word was provided for the parser.

A baseline accuracy of 77.65% was achieved when the parser was trained on the data that had original structure inherent to the treebank. Further improvements on the accuracy of the model can be attained by treebank transformations which make the syntactic structure of the treebank more transparent for tracing morpho-syntactic information. The optimal performance was gained by eliminating information contained in the original treebank and by percolating relevant morphological and functional information between lexical and phrasal nodes in combination with binarization of tree structure.

Elimination of information: Topological fields such as `VF`, `MF`, `NF` (see (Hinrichs et al., 2000) for details) are present in the original treebank as an intermediate layer of syntactic structure. Elimination of all topological field information except `C` and `VC` proved advantageous since retension of only these two fields turned out to be sufficient for determining the syntactic macrostucture across different clause types.

Percolation of relevant information: Due to the context-freeness of the underlying grammar formalism, regularities in syntactic structures should whenever possible be encoded as mother-daughter relations. In the simplest case regularities among sister constituents are best encoded by percolating this information onto the mother node. In the case of more non-local dependencies information needs to be percolated to the first common ancestor. Percolation of the following morphological and functional information between lexical and phrasal nodes proved particularly significant:

1. Morphological information (case, number and gender) on phrasal categories that constitute an NP (nouns, adjectives, determiners etc.) and partial morphological information (number) on verbal categories.

2. Function labels -OA and -ON (accusative object and subject) on nodes: passing them up to SIMPX node (node of a clause) allows to capture subject-object regularities

	precision	# of unparsed sentences
baseline:	**77.65**	5
topological fields (except for VC and C) deleted	77.58	5
case passed up to NX and NCX	84.23	6
grammatical functions (-ON and -OA) added and passed up to SIMPX + rules binarized	84.98	5
morph. info passed up to NXs and VXFINs	87.62	7
FIN label with number passed up to SIMPX	**88.21**	9
results on test data	**87.69**	11

Figure 1: LoPar PCFG experiments

in a clause: presence of one subject in a clause and prototypicality of clauses with both subject and accusative object.

3. FIN-label of finite verbs: passing it up to SIMPX node allows to trace regularities between finiteness of a clause and morphological cases of NPs in a clause: necessary presence of subject in finite clause, absence of subject in non-finite clause. Moreover, introducing number feature to -FIN and -ON nodes makes it possible to capture subject-verb agreement.

Binarization of SIMPX subtrees: Enriching syntactic categories by morphological and functional information leads to a severe increase in the size of the rule set with an accompanying drop in recall (i.e. an increase in the number of unparsed sentences). This can be counterbalanced by binarization of tree structures, which makes the grammar more flexible by allowing for structures not present in the training data to be created in the tagging phase.

Figure 1 summarizes transformations in tree structure and different node label encodings used in the experiments and the impact of the transformations on the performance of the model on the development set. The best result obtained has an accuracy of 87.69%, which amounts to significant reduction of 55.08% in the error rate of the baseline.

Summing up: the present paper confirms the findings of (Johnson, 1998) that treebank transformations play a crucial role in the application of PCFG models to natural language data and generalizes this result to a novel use of PCFGs for morpho-syntactic tagging of highly inflectional languages with large tagsets. It underscores the importance of careful linguistic fine-tuning of such models.

References

Hinrichs, E. W., J. Bartels, Y. Kawata, V. Kordoni, and H. Telljohann (2000). The Verbmobil treebanks. In *5. Konferenz zur Verarbeitung natürlicher Sprache (KON-*

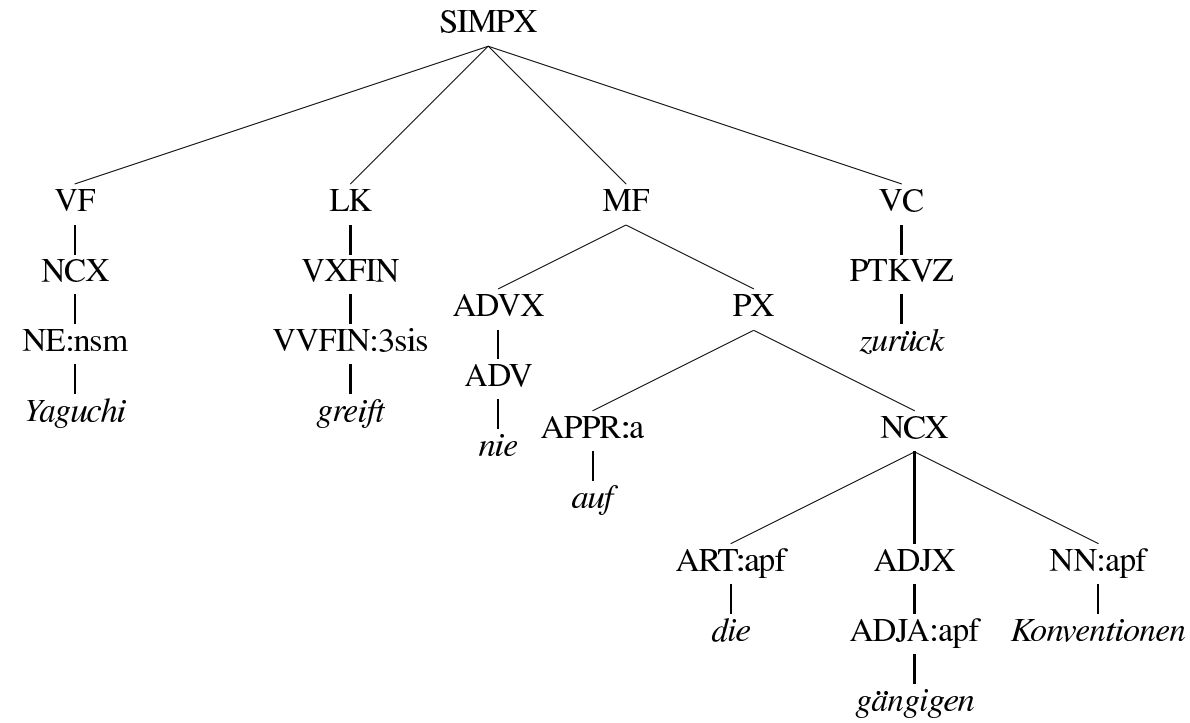

Figure 2: Example of a tree in the original treebank

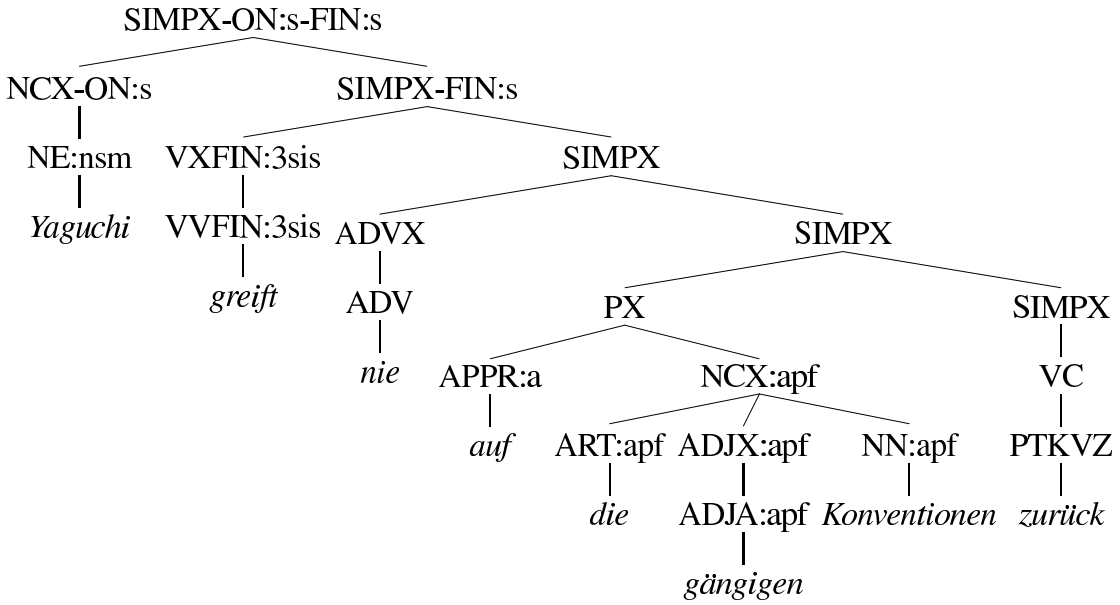

Figure 3: Example of a tree in the transformed treebank

VENS 2000), pp. 107–112. Ilmenau, Germany.

Hinrichs, E. W. and J. Trushkina (2003). N-gram and PCFG models for morpho-syntactic tagging of German. In *Proceedings of Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*. Växjö, Sweden.

Johnson, M. (1998). The effect of alternative tree representations on tree bank grammars. In D. M. W. Powers, ed., *Proceedings of NeMLaP3/CoNLL98*, pp. 39–48. Sydney, Australia.

Schmid, H. (2000). Lopar: Design and implementation. Technical Report 149, IMS Stuttgart. Arbeitspapiere des Sonderforschungsbereiches 340.