Morphological disambiguation in German

Karel Vaculík

German belongs to the group of inflected languages. As such, it has richer morphology than English for instance which has minimal inflection. This makes the process of morphology analysis harder as there are more morphological tags. It follows that morphology disambiguation is also more complex for such languages.

There are more ways how to deal with ambiguity. Researchers from Eberhard Karls Universität Tübingen [1] are using two types of noun phrase disambiguation rules within Xerox Incremental Deep Parsing System (XIP):

1) Ordinary disambiguation rules (ODRs), which eliminate readings for a single lexical node, and
2) Double reduction rules (DRRs), which simultaneously reduce readings of sequence of tokens

These rules are based on left and/or right contexts of the processed token(s). The general format of ODR rule is:

readings_filter = |left_context| selected_readings |right_context|

Simple example can be for example:

det, pron = det |adj*, noun|

This rule applies to tokens which have determiner and pronoun readings. If the token is followed by any number of adjectives and a noun, only determiner reading is retained.

After applying these rules, XIP employs syntactic heuristics on non-disambiguated NPs. These heuristics are also in form of rules.

Another tool with disambiguation is GERTWOL [2] - a system for automatic recognition of German word forms. There are two types of morphology disambiguation in the system [3]:

1) Local disambiguation – it retains only those readings with the fewest suffixes or composition borders. No context is needed. For example, consider "<zugriffsbereite>". Only two readings remain:
   "zug#riff\s|bereit" A POS SG NOM FEM
   "zu|griff\s|bereit" A POS SG NOM FEM
2) Contextual disambiguation – similarly to previous tool, it is carried out by using grammatical and heuristic rules. Grammatical rules consist of four parts: functional area (domain), target, operator and contextual conditions. Heuristic rules are again used for further refinement.

Different approach is proposed in [4]. Using the SMOR [5] morphological analyzer, the input words are first split into morpheme sequences and then analyzed with the probabilistic context-free grammar. Used grammar is quite small and its probabilities are trained on unlabeled data with LoPar parser [6]. It is using the Inside-Outside algorithm which is an instance of the unsupervised EM algorithm. A German model for HMM tagger is presented in

[7]. Using this model, disambiguation of POS is performed. Combination of Brill-based unsupervised tagger and word-case sensitive rules is used for morphological disambiguation in information extraction system SMES [8].

References

[1] Hinrichs, E., Trushkina, J.: Forging Agreement: Morphological Disambiguation of Noun Phrases. In *Proceedings of the First Workshop on Treebanks and Linguistic Theory*. 2002. pp 78—95.

[2] GERTWOL: http://www2.lingsoft.fi/doc/gertwol/intro/overview.html

[3] GERTWOL: http://www2.lingsoft.fi/doc/gercg/NODALIDA-poster.html

[4] Schmid, H.: Disambiguation of Morphological Structure using a PCFG. In *Proceeding HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 2005. pp 515—522

[5] Schmid, H., Fitschen, A., Heid, U.: SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, volume 4, pp 1263—1266, Lisbon, Portugal.

[6] LoPar: http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LoPar.html

[7] Feldweg, H.: Implementation and evaluation of a German HMM for POS disambiguation. In *Proceedings of the EACL SIGDAT Workshop*. 1995.

[8] Neumann, G. et al.: An Information Extraction Core System for Real World German Text Processing. In *proceedings of ANLP-1997*, Washington, DC, pages 209-216.