Acquiring a Poor Man's Inflectional Lexicon for German

Peter Adolphs[1]

Abstract

Many NLP modules and applications such as morphosyntactic corpus annotation tools require the availability of a module for wide-coverage inflectional analysis. One way to provide such analyses is to look up the word form in an inflectional lexicon. Such a lexicon should list stems and their inflectional classes instead of the full forms for a better maintainability. To my knowledge, there is no such lexicon freely available for German. Furthermore, existing inflectional lexicons need to be expanded, for instance, to encompass domain-specific vocabulary.

The manual creation and maintenance of an inflectional lexicon is a dull and strenuous task. Since large text corpora nowadays are easily available and inflectional systems are in general well understood, it seems feasible to acquire lexical data from raw texts, guided by our knowledge of inflection. Several such methods have been developed in recent years for different languages including Croatian, Russian, French, and Slovak (see references).

I present an acquisition method along these lines for German. The general idea can be roughly summarised as follows: first, generate a set of lexical entry hypotheses for each word-form in the corpus; then, select hypotheses that explain the word-forms found in the corpus "best". To this end, I have turned an existing morphological grammar, cast in finite-state technology (Schmid et al., 2004), into a hypothesiser for lexical entries. Irregular forms are simply listed so that they do not interfere with the regular rules used in the hypothesiser. Running the hypothesiser on a text corpus yields a large number of lexical entry hypotheses. These are then ranked according to their validity with the help of a statistical model that is based on the number of attested and predicted word forms for each hypothesis. First results of the system are promising; e.g., »50% precision and > 75% recall are achieved for verbs.

References

Clément, Lionel; Sagot, Benoît & Lang, Bernhard. 2004. "Morphology based automatic acquisition of large-coverage lexica". In: *Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.

Forsberg, Markus; Hammarström, Harald & Ranta, Aarne. 2006. "Morphological Lexicon Extraction from Raw Text Data". In: *Proceedings of FinTAL - 5th International Conference on Natural Language Processing*. Turku, Finland.

Oliver, Antoni; Castellón, Irene & Màrquez, Lluis. 2003. "Automatic Lexical Acquisition from Raw Corpora: An Application to Russian". In: *Proceedings of the EACL-2003 Workshop on Morphological Processing of Slavic Languages*. Budapest, Hungary.

Oliver, Antoni & Tadi´c, Marko. 2004. "Enlarging the Croatian Morphological

[1] *e-mail*: peter.adolphs@student.hu-berlin.de

Lexicon by Automatic Lexical Acquisition from Raw Corpora". In: *Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC 2004)*, pp. 1259–1262. Lisbon, Portugal.

Sagot, Benoît. 2005. "Automatic acquisition of a Slovak Lexicon from a Raw Corpus". In: Matoušek, Václav; Mautner, Pavel & Pavelka, Tomáš (eds.) *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005.*

Proceedings., volume 3658 of *Lecture Notes in Computer Science*, pp. 156–163. Berlin / Heidelberg: Springer.

Schmid, Helmut; Fitschen, Arne & Heid, Ulrich. 2004. "SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection". In: *Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC 2004)*, pp. 1263–66. Lisbon, Portugal.