Automatic Lexical Acquisition for German Based on Morphological Paradigms

Diploma Thesis Proposal

Peter Adolphs
peter.adolphs@student.hu-berlin.de
Humboldt-Universität zu Berlin
Institut für Informatik

13th June 2006

Abstract

The general aim of my diploma thesis is to develop a (semi-)automatic method for the acquisition of a German inflectional lexicon from raw texts. In particular, I want to explore whether inflectional stems can be deduced from word-form occurences that fit into known morphological paradigm classes.

1 Introduction

Many modules and applications in Computational Linguistics rely on an automatic analysis of word-forms that allows for the disregarding of inflectional markers and for reducing a word-form to a more abstract unit of vocabulary. This lexical unit is called a 'lexeme'[1]. It comprises all word-forms that share a common meaning and may vary only in their form and applicability in specific syntactic contexts. Lexemes are identified by their 'lemma' – a unique name for the lexeme within the language's vocubulary[2].

Properties of word-forms often cannot be deduced from the word-form alone but have to be stored in or derived from a lexicon. There are several reasons for this: various features are highly lexicalised, inflectional markers are often ambigious, and some inflectional paradigms are simply irregular. It is, for example, not possible in German to predict without further knowledge whether an umlaut in a word-form is an (additional) inflectional marker or whether it belongs to the inflectional

[1] It is also called a "word" in everyday speech. However, the word "word" is notoriously ambiguous and is therefore usually avoided as a scientific term.

[2] Other usages of the term 'lemma' exist. Some authors use the term to refer to a 'lexeme' as defined above, and use other terms as 'lemma name' or 'headword' to refer to the identifier for this lexical unit.

stem (cf. *Plänen*, ACC PL[3] of PLAN 'plan' vs. *Flächen*, ACC PL of FLÄCHE 'area'). Furthermore, it is in general not possible to determine the part-of-speech of word-forms or other highly lexicalised morpho-syntactic properties (as gender for nouns) without using a lexicon. Therefore, lexicon-free methods as most word-form conflation methods in Information Retrieval[4] do not lead to satisfactory results for linguistic stemming or lemmatisation tasks.

The manual creation and maintenance of a lexicon is very effort intensive and time consuming. Even existing inflectional lexicons will permanently encounter unknown word-forms when dealing with unconstrained texts[5]. The purpose of my diploma thesis is to develop a (semi-)automatic method for the acquisition of a German inflectional lexicon from raw texts. Instead of a fully manual discovery and lexical encoding of word-forms on the one hand, or a fully automatic induction of morphological units and their combination rules on the other hand, I propose to formalise morphological knowledge about closed-class items (closed part-of-speech classes and common inflectional markers) and to use this knowledge to infer which open-class items do exist. In particular, I want to explore whether inflectional stems and inflectional classes can be deduced from word-form occurences that fit into known morphological paradigm classes.

2 Task

The goal of my diploma thesis is to develop a method for a (semi-)automatic acquisition of lemmas of open part-of-speech classes, where a lemma is a tuple (*inflectional stem*, *inflectional class*). If fully implemented, this tuple allows for the recognition and generation of the complete inflectional paradigm identified by this lemma, including morpho-syntactic properties for each inflected word-form such as:

- person, number, tense, modality, and finiteness for verbs,
- number, case, and gender for nouns,
- number, case, gender, degree, and declension type for adjectives.

The resulting inflectional analysers / generators can be used as components of a lemmatiser; however, it is not intended to provide a method for the required contextual disambiguation. Neither is it intended to model word-formation.

[3]The following abbreviations are used: NOM = nominative, GEN = genitive, DAT = dative, ACC = accusative, SG = singular, PL = plural, MASC = masculine, FEM = feminine, NEU = neuter.

[4]These techniques are commonly, but from a linguistic point of view misleadingly, called 'stemming'.

[5]Due to the generative capacity of German morphology, many of these word-forms can be explained and analysed on the basis of simpler units and word-formation rules. Therefore, an inflectional lexicon alone – no matter how large it is – will not suffice if high coverage of the language's vocabulary should be achieved. It is also crucial to provide a computational model of productive morphological word-formation processes. Such a complete computational morphology for German is out of the scope of my proposed work. However, once such an inflectional lexicon is available, it could be used to learn word-formation patterns, as suggested by Gaussier (1999).

3 Proposed Method

The main idea is that if different word-forms of an inflectional paradigm are attested in a corpus, it is possible to infer the corresponding lemma and to associate morpho-syntactic features to each word-form. For example, word-forms like *schläfst*, *schlafend*, *geschlafen* allow the inference that there is a lemma consisting of the verb stem *schlaf-* and the inflectional class that requires umlaut in the second and third person singular present and that forms the past participle with *ge-...-en*. Prior approaches used the same basic line of reasoning to acquire a lexicon of Russian (Oliver et al., 2003), Croatian (Oliver & Tadić, 2004), French (Clément et al., 2004), Slovak (Sagot, 2005), and to build a lemmatiser for German nouns (Perera & Witte, 2005).

In addition to paradigm-based recognition of word-forms, the grammatical context of each word-form might contribute information that helps to disambiguate multiple lemma hypotheses, and provide lexicalised morpho-syntactic properties such as gender for nouns. For instance, a word-form sequence such as *Der Traum* suggests that *Traum* is very likely to be a noun (if also taking into consideration that its first letter is capitalised), and that it is either NOM SG MASC, GEN SG FEM, DAT SG FEM, or GEN PL MASC/FEM/NEU. This idea has been used for the acquisition of lexicalised morpho-syntactic properties in an Italian lexicon (Zanchetta & Baroni, 2006), and for constraining lemma hypotheses in the systems presented by Clément et al. (2004) and Perera & Witte (2005).

The paradigm-based acquisition method relies on the following resources: (i) inflectional rules relating word-forms to their inflectional stem and inflectional class and vice versa, and (ii) a tokenised text corpus. Additionally, (iii) lists of closed part-of-speech classes (prepositions, conjunctions, etc.), and (iv) local grammars based on either categorial or statistical part-of-speech patterns (that is, by using regular expressions, or HMM or decision tree models, respectively) are needed for determining constraints on the morpho-syntactic properties of particular word-forms.

The proposed method works as follows: for each unknown word-form type in the corpus, the inflectional rules are used to build all hypotheses of its lemma and the corresponding inflectional class. Following the ideas of Clément et al. (2004) and Sagot (2005) the lemmas are then ranked by their plausibility where lemma plausibility is correlated with the number of word-form types from its inflectional paradigm that are actually attested in the corpus, as well as their token occurrences. The ranked list of lemmas can be manually validated by a native speaker of the language (Clément et al. (2004) report that they acquired nearly 5000 French verbs "[a]fter only a few hours of manual validation and loop iteration"). Alternatively, it could also be experimented with an automatic approval of the first n lemmas and a reiteration of the whole process.

Unlike the approaches of Oliver et al. (2003), Oliver & Tadić (2004), and Perera & Witte (2005) but like the approaches of Clément et al. (2004), Sagot (2005), and Zanchetta & Baroni (2006), the proposed method ensures that only complete inflectional paradigms of lexemes are represented in the lexicon.

In order to evaluate the quality of the ranking, the proportion of correct lemma hypotheses among the first n ranked hypotheses can be compared for different values of n. Finally, the coverage of the acquired lexicon can be measured by comparing it to a gold-standard corpus with lemma and morpho-syntactic annotation such as the TIGER Treebank (Brants et al., 2002).

References

Brants, Sabine; Dipper, Stefanie; Hansen, Silvia; Lezius, Wolfgang & Smith, George. 2002. "The TIGER Treebank". In: *Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT 2002)*. Sozopol, Bulgaria.
`http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/treeling2002.pdf`

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.

Gaussier, Eric. 1999. "Unsupervised learning of derivational morphology from inflectional lexicons". In: Kehler, Andrew & Stolcke, Andreas (eds.) *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing*. University of Maryland, USA.
`http://www.aclweb.org/anthology/W99-0904.pdf`

Heid, Ulrich. 2000. "Morphologie und Lexikon". In: *Handbuch der Künstlichen Intelligenz*. München, Wien: Oldenbourg, 3rd edn.

Oliver, Antoni; Castellón, Irene & Màrquez, Lluis. 2003. "Use of internet for augmenting coverage in a lexical acquisition system from raw corpora". In: *Proceedings of the International Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages (IESL 2003)*. Borovets, Bulgaria. Held at the International Conference RANLP 2003.

Oliver, Antoni & Tadić, Marko. 2004. "Enlarging the Croatian Morphological 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.

Perera, Praharshana & Witte, René. 2005. "A Self-Learning Context-Aware Lemmatizer for German". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pp. 636–643. Vancouver, Canada.
`http://www.aclweb.org/anthology/H05-1080.pdf`

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

Zanchetta, Eros & Baroni, Marco. 2006. "Morph-it! A free corpus-based morphological resource for the Italian language". In: *Corpus Linguistics 2005*, vol. 1 of *Proceedings from the Corpus Linguistics Conference Series*. Birmingham, UK.
`http://www.corpus.bham.ac.uk/PCLC/`