MorphoClass — Recognition and Morphological Classification of Unknown Words for German

Preslav Nakov[1]

Abstract. A system for recognition and morphological classification of unknown words for German is described and evaluated. It takes raw text as input and outputs a list of the unknown nouns together with a hypothesis about their possible morphological class and stem. MorphoClass exploits global information (ending-guessing rules, maximum likelihood [1]estimations, word frequency statistics), morphological properties (compounding, inflection, affixes) and external knowledge (lexicons, German grammar information etc.).

1 SYSTEM OVERVIEW

The MorphoClass system accepts raw text as input and produces a list of unknown words together with hypotheses about their *stem* and *morphological class*. We define the *stem* as the common part shared by all inflected forms of the base while the *morphological class* describes both the word gender and the inflexion rules the word follows when changes by case and number. Our morphological classes follow the one used under the DBR-MAT project — a German-Bulgarian-Romanian Machine Translation (see [1]), and given in Bulgarisch-Deutsch Worterbuch (see [2]).

MorphoClass solves the problem as a sequence of subtasks including: unknown words identification, noun identification, inflected forms of the same word recognition and grouping, compounds splitting, morphological analysis, stem proposal for each group of inflected forms, and finally — production of hypothesis about the possible morphological class for each group of words.

2 RELATED WORK

Koskenniemi proposes a language-independent model for both morphological analysis and generation called *two-level morphology* and based on finite-state automata. It is implemented in the *KIMMO* system (see [3]). Finkler and Neumann follow a different approach using *n*-ary tries in the *MORPHIX system* (see [4]). Lorenz developed *Deutsche Malaga-Morphologie* as a system for the automatic word form recognition for German based on *Left-Associative Grammar* (see [5]). Kupiec uses pre-specified suffixes and then learns statistically the POS predictions for unknown word guessing (see [6]). The XEROX tagger comes with a list of built-in ending-guessing rules (see [7]). Brill builds more linguistically motivated rules by means of tagged corpus and a lexicon (see [8]). He does not look at the affixes only but optionally checks their POS class in a lexicon. Mikheev proposes a similar approach that estimates the rule predictions from a raw text (see [9]). Daciuk uses finite state transducers.

3 SYSTEM OVERVIEW

3.1 Unknown word tokens and types identification

MorphoClass is interested in the identification and morphological classification of the nouns with *unknown stems*. The first thing to do is to process the text and to derive a list of the word types. We exploit the German noun property to be always capitalised regardless of its position in the sentence. The capitalisation is discarded when deriving the list but is taken into account since for each word we collect the following three statistics: total frequency, capitalised frequency and start-of-sentence frequency. These are used to determine whether a certain word type could be a (unknown) noun.

3.2 All possible stems generation

We go through the words and generate all the possible stems that could be obtained by reversing all acceptable German inflexions for the word type while taking into account the umlauts and the *ß* alternations. For each word type all acceptable rule inversions are performed. For example for the word *Lehrerinnen* the following stems are generated (by removing *-nen*, *-en*, *-n* and ∅): *Lehrerin*, *Lehrerinn*, *Lehrerinne*, *Lehrerinnen*. We do not impose any limitations when generating a stem except that it must be non-empty. The purpose of the stem generation process is to both identify all the acceptable stems and group the inflected forms of the same word together.

3.3 Stem coverage checking and refinement

We go through the stems and for each one we check whether there exists a morphological class that could generate all the word forms. If at least one is found we accept the current coverage and otherwise we try to refine it in order to make it acceptable. It is possible that a stem is generated by a set of words that it cannot cover together. It is important to say that at this moment we are *not* interested in the question whether this stem is really *correct* but

[1] Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
e-mail: preslav@rocketmail.com

just in whether it is *compatible* with all the word forms it covers taken together.

3.4 Morphological stem analysis

Each stem generated in the previous step is analysed morphologically in order to obtain some additional information that could imply useful constraints on the subsequent analysis. The morphological analysis is based on both lexicon-based and suffix-based morphology. First, for each stem we check whether it is present in our stem lexicon. (We built it using the free lexicon of the Morphy system (see [10])). If so, we reject it since the *unknown word* could **not** have a *known stem*: all words the known stems could generate are already known. Second, we check whether the stem could be a compound by trying to split it in a way that all its parts are found in the lexicon. In case of success we know its morphological class — it is determined by the last word the compound is made of. Third, we try to guess the class looking at the stem ending. We implemented a Mikheev-like ending-guessing rules (see [9]). We selected a confidence level of 90%, considered endings up to 7 characters long that must be preceded by at least 3 characters and whose frequency is at least 10. We trained the model over 8,5 MB of raw text and obtained 1789 rules.

3.5 Word types clusterisation (stem coverage)

After the stem refinements step we are sure that each stem is compatible with the word types it is supposed to cover and that there exists at least one morphological class that could generate them all given the stem. During the next step we obtained some additional information regarding the stems as a result of morphological analysis. We thus have a complex structure, which we can think of as a bi-partitioned graph where the vertices are either stems or word types and each edge links a stem to a word type it is supposed to cover. Our goal is to select some of the stems thus producing stem coverage of the word types. We try to select some stems in a way that:

1 Each word is covered by exactly one stem.
2 The stem covers as much word types as possible.
3 The covered word types set being equal, a stem with more reliable morphological information is selected. We prefer words recognised as compounds, then those analysed using ending-guessing rules and then all the rest.
4 All other being equal, a longer stem is preferred.

4 EVALUATION

The MorphoClass system has been evaluated over an 85 KB German literature text: *Erzählungen* by Franz Kafka. There were 3510 different word forms found: 862 known nouns, 2155 known non-nouns and *493 unknown nouns*. The evaluation has been performed manually over a quarter of the stems. We considered 120 stems and classified them in the following categories (counts in parentheses):

SET (12) — A *set* of classes has been assigned rather than a single one.
PART (7) — MorphoClass discovered a *correct* class but *not all* the correct classes.
WRONG (18) — MorphoClass assigned a single class but it was *wrong*.
YES (72) — MorphoClass assigned a single class and it was the only correct one.
SKIP (11) — The stem has been skipped. We did so for the proper nouns, incorrect stems etc.

We evaluated the System in terms of *precision* and *coverage.* The *coverage* shows the proportion of the stems whose morphological class has been found, while the *precision* reveals how correct it was. A scaling is performed according to the proportion of possible classes guessed to the total classes count: if a stem belongs to k ($k \geq 2$) classes and MorphoClass found one of them (it finds exactly one) then *precision1* considers it as a failure (will add 0), *precision2* counts it as a partial success (will add *scaled_PART*=1/k) and *precision3* accepts it as a full success (will add 1).

precision1 = YES / (YES + WRONG + PART)
precision2 = (YES + (*scaled_PART*)) / (YES + WRONG + PART)
precision3 = (YES + PART) / (YES + WRONG + PART)
coverage = (YES + WRONG + PART) / (YES + WRONG + PART + SET)

The MorphoClass system performs the morphological analysis using both compound words splitting as well as ending-guessing rules. These are run in a cascade manner: the ending rules are applied *only* if the compound splitting rules failed. Not surprisingly the compound splitting rules gave a high precision: 93.62% (no partial matching: all the rules considered predicted just one class even when more than one splitting was possible) and coverage of 43.12%. These results give an idea of how often the compound nouns occur on German. Another 45.87% of the stems have been covered by the ending-guessing rules. Their precision was much lower: 56% for *precision1* and 70% for *precision3*. This gave us an overall system coverage of 88.99% and precision of 74.23%, 76.08% and 81.44%. (see Table 1)

Table 1. MorphoClass system evaluation results

	Compounds	**Ending-guessing**	**Overall**
coverage	*43.119266%*	*45.871560%*	*88.990826%*
precision1	93.617021%	56.000000%	74.226804%
Precision2	93.617021%	57.470000%	76.082474%
Precision3	*93.617021%*	*70.000000%*	*81.443299%*

5 CONCLUSIONS AND FUTURE WORK

We use very simple rules only, without exploiting any context information and most of the unknown nouns' stems have just one (possibly inflected) noun form. A similar approach could be applied to other inflectional languages and other important open-class POS such as: adjectives, verbs and adverbs. Obviously, this will not be straightforward but most of the steps could be applied with almost no changes.

ACKNOWLEDGEMENTS

I am very grateful to prof. Galia Angelova, prof. Walther von Hahn and Ingo Schröder for the valuable suggestions and discussions. Special thanks to prof. Galia Angelova for the strong support.

REFERENCES

[1] W. von Hahn, G. Angelova. *Combining Terminology, Lexical Semantics and Knowledge Representation in Machine Aided Translation.* In: TKE'96: Terminology and Knowledge Engineering. Proceedings of the Conference "Terminology and Knowledge Engineering", August 1996, Vienna, Austria. pp. 304 – 314.

[2] E. Dietmar and H. Walter. *Bulgarisch-Deutsch Wörterbuch.* VEB Verlag Enzyklopädie Leipzig, 1987

[3] K. Koskenniemi. *Two-level model for morphological analysis.* In IJCAI 1983 pp. 683-685, Karlsruhe, 1983.

[4] W. Finkler, G. Neumann. *MORPHIX. A Fast Realization of a Classification-Based Approach to Morphology.* In: Trost, H. (ed.): 4. Osterreichische Artificial-Intelligence-Tagung. Wiener Workshop - Wissensbasierte Sprachverarbeitung. Proceedings. Berlin etc. pp. 11-19, Springer, 1988.

[5] O. Lorenz. *Automatische Wortformenerkennung für das Deutsche im Rahmen von Malaga. Magisterarbeit.* Friedrich-Alexander-Universität Erlangen-Nürnberg, Abteilung für Computerlinguistik.

[6] J. Kupiec. *Robust part-of-speech tagging using a hidden Markov model.* Computer Speech and Language, 6(3), pp.225-242, 1992.

[7] D. Cutting, J. Kupiec, J. Pedersen, P. Sibun. *A practical part-of-speech tagger*. Proceedings of the Third Conference on Applied Natural Language Processing (ANLP-92), pp. 133-140, 1992.

[8] E. Brill. *Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging.* In Computational Linguistics, 21(4):543-565.

[9] A. Mikheev. *Automatic Rule Induction for Unknown Word Guessing*. In Computational Linguistics vol 23(3), ACL 1997. pp. 405-423.

[10] W. Lezius. *Morphy - German Morphology, Part-of-Speech Tagging and Applications*. In Ulrich Heid; Stefan Evert; Egbert Lehmann and Christian Rohrer, editors, Proceedings of the 9th EURALEX International Congress pp. 619-623 Stuttgart, Germany.