

“NIBA-TAG” – A tool for analyzing and preparing German texts

G. Fliedl, G. Weber

University of Klagenfurt – Department for Business Informatics and Application Systems(IWAS); Austria

Abstract

NIBA-TAG is a kind of multilevel natural language tagger with rich functionality. It functions as a word-stemmer, a morphological parser and a normal POS-Tagger, which uses syntactic and semantic features for contextually influenced word-tagging.

Each rule is based on a ranking-mechanism which is currently related to the levels “fact”, “proposal” and “guess”. One of the postprocessing-units analyzes the ranking-structure and can change a “proposal” to a “fact”, if enough rules made an identical proposal for a word.

The default output is XML, where the level of precision can be specified. So one could generate a XML-file only including the guesses, or a file with all attributes relevant for the status of a proposal.

1 Introduction

The automated analysis of language is important for many tasks in computer science. A mass of information exists in unstructured texts. For analyzing all different sorts of unstructured text we developed our tagging-tool. It is a very efficient instrument for filtering out the STRUCTURE OF CONTENT.

In the field of textual analysis no satisfactory results are known up to now, because methods have always been focussed on mainly statistical methods (Kupiec [1]) or to a reduced linguistic functionality. So we concentrated on the development of a general tool for multilevel tagging. The system was implemented in Perl and Prolog; the technical features are: 15423 sentences (164380 words) of a testcorpus have been tagged in 1450 seconds. Linguistic analysis is done according to the NTMS model, which stands for

‘Natürlichkeitstheoretische Morphosyntax’ (‘Naturalness Theoretical MorphoSyntax’) (Fliedl [4,5,6]) and (Mayerthaler [7]). NTMS is a generative oriented grammar model that uses category-specific phrase structures for the description of grammatical phenomena.

2 Some general aspects of Natural Language Processing (NLP)

NLP should provide a better way for navigation through texts, because linguistic analysis enables a very systematic form of “lemmatizing”, which can be defined as an efficient way of reduction of word-variation. The automation of document summarization includes such a linguistically based “lemmatizing-strategy”. Improving applications in information retrieval and data mining based on such a strategy is a challenging task for NLP. The aim of information extraction is converting unstructured text to structured databases. This includes systems, which accurately identify names, locations, dates and other atomic features of texts.

3 The functionality of the tagger

Deciding whether words are nouns, verbs, adjectives, adverbs etc. – the so called P(art)O(f)S(peech)-tagging - is partly based on dictionary information, partly on corpora information and includes very often sentence context.

POS-tagging has concentrated on English in the past (Brill [2], Marcus [3]). It was assumed, that methods having been developed for English would work for other languages as well. But there have grown some problems during the development of taggers for “foreign” languages. With respect to german one big problem arises from the morphological “richness”. The complexity of words and the ambiguity of word-endings doesn’t always allow a straight interpretation of the input-material.. So many context-rules had to be formulated, which function as disambiguation triggers.

The tagger includes some preprocessing features, a lexicon, a morphological analyzer, a stemmer and a shallow syntax-parser.

3.1 Some preprocessing features

A pure text is normally the only possible input for all sorts analyzing tasks. Prelinguistic steps include filtering out textformatting features, identifying sentence boundaries as expressed by surface punctuation and capitalization cues.

Document filters for HTML, XML , XML and DTDs are already implemented, see Figure 1:

```
<hr><p>Der Mann <b>im Mond</b></p>
<text>
<sentence>
  <spz0 position="0">Der</spz0>
  <n0 position="1">Mann</n0>
  <p0 position="2" Type="loc">im</p0>
  <n0 position="3" lastposition="1">Mond</n0>
</sentence>
</text>
```

Figure 1

Recognizing common expressions is also done during preprocessing, because the identification of their related patterns is not really a linguistic job. The subsequent list is not at all complete, but somehow representative:

- Monetary expressions;
- Spelled-out numerals;
- Numerals in decimal format, percent expressions etc.;
- Times of day in many different formats and variants; lists or sequences of times;
- Date expressions whether fully or partially spelled out and compound expressions giving lists or sequences of dates;
- Internet domain names, URLs, email addresses;
- Alphanumeric words, acronyms and abbreviations;
- Names of cities, in the format "City + Region", "City + Region + Postal Code";
- Names of corporations, easily recognizable from expressions like "co." and "inc.";
- trademarks and product names, similarly recognizable from obvious surface cues;
- Human names of the form (Title) (First name) (Middle name) (Last name);
- Telephone numbers, see Figure 2:

```
<sentence>
  <n0 Type="proper noun">Helen</n0>
  <spz0 Type="tvag2" >wählt</spz0>
  <spz0 >die</spz0>
  <n0>Nummer</n0>
  <n0 Type="phone">+43 676 000</n0>
  <spz0>das</spz0>
  <q0 numeral="zweite" number="2">zweite</q0>
  <n0>Mal</n0>
</sentence>
<sentence>
  <a0 ReferTo="unser">Unsere</a0>
  <n0 Type="propernoun">URL</n0>
  <spz0 Type="iv" Numerus="sg">lautet</spz0>
  <n0 type="url">http://www.foo.bar</n0>
</sentence>
```

Figure 2

3.2 The lexicon

The lexicon has been developed in the subsequent way. Lexicon data have been collected in a relational data base and transformed into class-specific and morphosyntactically relevant replacement rules using VBA-functions. Up to now the following steps have been completed:

- Assigning the stems of simplex verbs to 20 semantically interpreted verb classes;
- Generating verb class specific lexicon entries for the shallow syntax parser;
- Building up a nominal data base in which nouns are specified with regard to primary and domain specific features.

The verb lexicon contains more than 10000 verb stems including the most frequent German simplex verbs, prefix verbs and particle verbs and the respective inflected forms based on the following verbclasses (Figure 3):

Class	Abbreviation	Predicate Argumenst Structure
1	AUX	<Ø>/V _{-fin}
2	eV	[TH]
3	iV	AG/TH[]
4	locV	TH[LOC]
5	possV	GO[TH]
6	psychV	TH[GO/TH]
7	tVag/2	AG[TH]
7,1	tVag/2d,g	AG[GO/BEN]
7,2	tVag/2pp	AG[SO/GO]
7,3	tVag/2sk	AG[THabstr]
7,4	tV/ppsk	AG[GO/SOabstr]
8	tV/3	AG/GO(TH,GO/SO)
8,1	tV/3ti	AG[TH...,GOabstr]
8,2	tV/3tda	AG[THacc,GOabstr]
8,3	tV/3td	AG[THdat,GOabstr]
8,4	tV/3sk,ak	AG[GO,THabstr]
9	sentV	EXP[TH]
10	Vcop	TH[N2/A2]
11	tV/2	-AG/-EXP[TH]
12	reflV	AGi/THi[i,(GO/LOC/TH)]

Figure 3

In Figure 3 the abbreviations are defined as follows:

(1) AUX = auxiliary; (2) eV = ergative verb; (3) iV = intransitive verb; (4) locV = locative verb; (5) possV = possessive verb; (6) psychV = psychological verb; (7) tVag2 = bivalent agentverb; (7,1) tVag/2d,g = bivalent agentverb with dative-object or genitive-object; (7,2) tVag/2pp = bivalent agentverb with prepositional objekt; (7,3) tVag/2sk = bivalent agentverb with sentential object; (7,4) tVag/2pp = bivalent agentverb with sentential prepositional object; (8) tV/3 = trivalent verb; (8,1) tV/3ti = trivalent verb with infinitival complement sentence and thematic identity of the antecedent and the logical subjects of the infinitival group; (8,2) tV/3tdd = trivalent verb with infinitival complement sentence and thematic difference of the antecedent and the logical subject of the infinitival group; (8,3) (tV3tdd) = agentive, trivalent verb with concrete theme in the dative and abstract „GOAL“; (8,4) (tV/3sk,ak) = trivalent verb with infinitival complement and subject control or arbitrary control; (9) sentV = verba sentiendi; (10) Vcop = copula verb; (11) tV/2 = transitive verb, whose subjekt does not carry the AG-Role nor the EXP-role trägt; (12) reflV = reflexive verb.

Our tagging of german verbs is based on this classification. Tags refer to the listed morhosyntactic, semantically motivated verbclasses.

Generally spoken, lexicon entries are classified by features and organized corresponding to application domains. The noun lexicon comprises a list of unmarked, frequent nouns. It is continually updated and enlarged. Nouns have been subcategorized with respect to the parameters [agentive], [animated], [derived], [compound] etc.

Moreover, nouns have been labeled domain specifically and assigned to the domains trade, traffic, education and tourism. Within these selected domains homonyms, synonyms, hypernyms and hyperonyms have been worked out.

3.3 The morphological analyzer and stemmer

Morphological-based categorization of unknown words is one of the key features of our system. Morphological analysis includes identification of wordforms through lemmatization, see Figure 4:

```
Fahr vorsichtig und melde dich.

<sentence>
  <spz0 position="0">Fahr</spz0>
  <a0 position="1" Type="Adjektiv">vorsichtig</a0>
  <conj0 position="2">und</conj0>
  <spz0 position="3" Stem="meld">melde</spz00>
  <n0 position="4" Type="Pronoun">dich</n0>
</sentence>
```

Figure 4

This sort of “reduction” implies accessing a German language dictionary which includes stemming and part of speech information.

336 Data Mining III

One step further morphosyntactically interpreted microstructures are assigned to wordgroups and words including the automatic splitting of German compound nouns; see Figure 5:

```
Alle Anweisungen ergeben an Kundenbetreuer

<sentence>
  <n3>
    <q0>Alle</q0>
    <n0 ReferTo=„Anweisung“>Anweisungen</n0>
  </n3>
  <spz0 Numerus=„pl“ Type=„tvag2pp“>ergehen</spz0>
  <n3>
    <n0 ReferTo=„Kunden,“ <n0 ReferTo= „betreuer“>
      Kundenbetreuer</n0>
  </n3>
</sentence>
```

Figure 5

The words of a given text are compiled into a lexicon which implies automatic categorial, semantic and contextual specification of these words.

3.4 The shallow syntax-parser

Analysis of word groups is done according to the specific X'-mechanism of NTMS (Fliedl [4]).

Identifying the structure of groups of words up to the level of simple clauses is one of the most relevant tasks in the field of linguistically based data mining. Our tool is able to identify prepositional phrases (PPs) and nominal phrases (NPs or N3s). In the following example the N3-node is established through identification of word patterns and context specifications; see Figure 6:

```
<sentence>
  <n3>
    <n3>
      <spz0 position="0">Der</spz0>
      <n0 position="1">Mann</n0>
    </n3>
    <p2>
      <p0 position="2" Type="loc">im</p0>
      <n3>
        <n0 position="3" lastposition="1">Mond</n0>
      </n3>
    </p2>
  </n3>
</sentence>
```

Figure 6

Our tagger also reduces the variety of words by *linking* nouns which only differ by their form (e.g. *man* and *men*) or by trying to link pronouns to their

referencing noun. The identification of word groups is done by our tagger-internal miniparser, using morphosyntactic rules for identification of phrasal units (we prefer rules for pattern identification, because the lexicon-based identification of word groups is a never-ending story).

4 Future work

In the near future we will concentrate on two very challenging tasks: summarizing of texts and automatic indexing. The automation of document summarization is needed for dealing with content economically; indexing we all need for text-filtering.

5 Bibliography

- [1] Kupiec, J.: "Robust part of speech tagging using a hidden markov model". Computer, Speech and Language, 1992.
- [2] Brill, E.: "A simple rule-based part of speech tagger." In: Proceedings of the Third Conference on Applied Natural Language Processing, ACL, 1992.
- [3] Marcus, M.; Santorini, B. and Marcienkiewicz, M.: "Building a large annotated corpus of English: the Penn Treebank." Computational Linguistics, 1993.
- [4] Fliedl, G: „Natürlichkeitstheoretische Morphosyntax, Aspekte der Theorie und Implementierung.“ Habilitationsschrift, Gunter Narr Verlag, Tübingen, 1999.
- [5] Fliedl, G.; Kop, Ch.; Mayerthaler, W.; Mayr, H.C.; Winkler, Ch.: "NTS-based derivation of KCPM perspective determiners." Proceedings of the 3rd International Workshop on Applications of Natural Language to Information Systems, June 26-27, 1997, The Harbour Centre Campus of SFU, Vancouver, Simon Fraser University 215-226, 1997.
- [6] Fliedl, G.; Mayerthaler, W.; Winkler, Ch.: "The NT(M)S Parser: An Efficient Computational linguistic Tool." Proceedings of the 1st International Workshop on Computer Science and Information Technologies, January 18-22, Moscow, 125-128, 1999,.
- [7] Mayerthaler, W.; Fliedl, G.; Winkler, Ch.: „Lexikon der Natürlichkeitstheoretischen Syntax und Morphosyntax.“ Stauffenburg Verlag, Tübingen, 1998.