Incremental Morphosyntactic Disambiguation of Nouns in German-Language Law Texts*

Kyoko Sugisaki
University of Zurich
Institute of Computational Linguistics
Binzmühlestrasse 14
8050 Zürich, Switzerland
sugisaki@cl.uzh.ch

Stefan Höfler
University of Zurich
Institute of Computational Linguistics
Binzmühlestrasse 14
8050 Zürich, Switzerland
hoefler@cl.uzh.ch

Abstract

Morphosyntactic disambiguation is a crucial pre-processing step for the recognition of grammatical functions in morphologically rich languages like German and heavily nominalized domains like law texts. This paper explores how far linguistically motivated hard rules can contribute to morphosyntactic disambiguation. It introduces an incremental system that is capable of reducing the rate of morphosyntactically ambiguous nouns in sentences from German-language law texts from 91.12% to 32.31%. The evaluation indicates that disambiguation rules based on feature unification within complex sub-clausal structures such as noun phrase coordinations and participle phrases have the most impact on the reduction of morphosyntactic ambiguity.

1 Introduction

Existing German full parsers (Sennrich et al., 2009; Foth et al., 2004) aim explicitly or implicitly at parsing a broad coverage of text types. Building a domain-specific parser from scratch is considered to be time-consuming and of limited value, since it is by definition not aimed at a broad usage outside of its domain. However, domain adaptation approaches such as self-training (McClosky and Charniak, 2008; Sagae, 2010) seem reasonable thanks to their scalability and portability.

In any new domain, the initial lack of a large annotated corpus remains, however, a big challenge for domain-specific parsing. Under these circumstances, a rule-based preprocessing approach appears fairly reasonable and promising, considering how accurate rule-based approaches have performed, especially in the field of POS tagging (Schneider and Volk, 1998; Voutilainen, 1995; Brill, 1992).

In this paper, we report on the development of a rule-based system for the morphosyntactic disambiguation of nouns as a preprocessing component of a supertagger for law texts. Suppertagging is an "almost parsing" approach in the sense that the supertags represent rich syntactic information such as valence, voice and grammatical functions (Foth et al., 2010; Harper and Wang, 2010; Nasr and Rambow, 2004) and a parser needs then "only combine the individual supertaggs" (Bangalore and Joshi, 1999).

We argue that the morphosyntactic disambiguation of nouns is a crucial step for the recognition of grammatical functions in a morphologically rich language, German. Especially, for the legislative domain, morphosyntactic disambiguation is a challenging task, since nouns in law texts are used intensively due to the frequency of coordination structures, light verb constructions and appositions (cf. Hansen-Schirra and Neumann, 2004; Nussbaumer, 2009).

The paper is organized as follows. In the next section, we describe the general architecture of our supertagger. In section 3, we present the three major components of the morphosyntactic disambiguation of nouns. In section 4, we evaluate the performance of our system and discuss the rate of the reduction

*The project is funded under SNSF grant 134701.

of morphosyntactic ambiguity for these three components.

2 Overview: Supertagger

Our supertagger has been specifically developed for detecting style guide violations in Swiss legislative drafts written in German (Höfler and Sugisaki, 2012). Our supertagger particularly aims at tagging core syntactic structures such as topological fields and grammatical functions (GF). It consists of a pipeline with the following components:

1. Sentence segmentation and tokenization
2. Morphological analysis
3. Topological field recognition
4. Morphosyntactic disambiguation
5. Grammatical function recognition

Sentence segmentation and tokenization (component 1) are carried out as described in Höfler and Sugisaki (2012).

For the morphological analysis (component 2), our system employs Gertwol, a classical two-level rule-based morphological analyser that provides fine-grained morphosyntactic features (Haapalainen and Majorin, 1994). However, Gertwol does not return any analysis if it cannot find the root of a word in its lexicon. In these cases, our system resorts to the analysis of a statistical decision-tree-based POS-tagger, TreeTagger (Schmid, 1995). TreeTagger has proven to be robust and its performance with regard to unknown words is relatively high (Volk and Schneider, 1998).

The three main components of the system, dealing with topological field recognition (component 3), morphosyntactic disambiguation (component 4) and grammatical function recognition (component 5) respectively, have been implemented in the framework of Constraint Grammar. Constraint Grammar (Karlsson et al., 1995) is a grammar formalism that has been successfully employed for morphological disambiguation in English (Voutilainen, 1993) as well as in morphologically rich languages such as Irish (Uí Dhonnchadha, 2006) and Icelandic (Loftsson, 2008).

In the remainder of this paper, we will focus on component 4 and its strategies for morphosyntactic disambiguation.

3 Morphosyntactic Disambiguation

Morphosyntactic features such as case or number are primary cues for the recognition of grammatical functions in German. Most of these features can be recognized on the basis of inflectional endings. However, not all endings are unique: the morphological paradigm of German exhibits a certain degree of syncretism. In such cases, further hard constraints like argument structures can be exploited. Only afterwards, less secure (soft) cues such as word order, definiteness, animacy and information structure have to be resorted to the decoding of grammatical functions. In order to model the flexibility and inflexibility of the language appropriately, components using these two types of cues (hard vs soft constraints) should be kept distinct in a system. In this paper, we demonstrate how far linguistically motivated hard constraints can reduce morphosyntactic ambiguity before any soft constraints are applied.

Morphological ambiguity is reduced in an incremental way. The system for morphosyntactic disambiguation consists of three steps: (1) local phrase-level feature unification, (2) upper phrase-level feature unification, (3) clause-level feature unification.

Table 1 illustrates the three-step incremental disambiguation of the case feature for the nouns and pronouns of the following example sentence:

(1) Sie berücksichtigt dabei den der Tierhalterin oder dem Tierhalter entstehenden Aufwand und das Wohlergehen der Tiere.[1]

'In doing so, it [the agency] takes into account the expenses arising for the animal owners and the welfare of the animals.'

In what follows, we will briefly explain each of the three disambiguation steps.

Step 1: Local phrase-level feature unification

In German noun phrases, the features case, number and gender of the head nouns are in agreement with those of their dependents. In step 1, the feature

[1] Swiss Animal Protection Ordinance, Art. 10 para. 3.

	Sie	*Tierhalterin*	*Tierhalter*	*Aufwand*	*Wohlgehen*	*Tiere*
Input: Gertwol	NOM NOM AKK AKK	NOM AKK DAT GEN	NOM NOM AKK AKK DAT GEN	NOM AKK DAT	NOM AKK DAT	NOM AKK DAT GEN
Step1: Local phrase-level feature unification	NOM NOM AKK AKK	DAT GEN	DAT	NOM AKK DAT	NOM AKK	DAT GEN
Step2: Upper phrase-level feature unification	NOM NOM AKK AKK	DAT	DAT	AKK	AKK	DAT GEN
Step3: Clause-level feature unification	NOM	DAT	DAT	AKK	AKK	DAT GEN

Table 1: Incremental case feature disambiguation of nouns in sentence (1).

sets of head nouns and their dependents are therefore compared, and features that cannot be unified are discarded.

For the token *Tierhalter* in sentence (1), Gertwol provides the following possible morphosyntactic analyses:[2]

(2) "Tierhalter"
"Tier#halt~er" "S MASK SG NOM"
"Tier#halt~er" "S MASK SG AKK"
"Tier#halt~er" "S MASK SG DAT"
"Tier#halt~er" "S MASK PL NOM"
"Tier#halt~er" "S MASK PL AKK"
"Tier#halt~er" "S MASK PL GEN"

In contrast, the determiner *dem* preceding *Tierhalter* yields the following six morphosyntactic analyses:[3]

(3) "dem"
"der" "ART DEF SG DAT MASK"
"das" "ART DEF SG DAT NEUTR"
"der" "PRON DEM SG DAT MASK"
"das" "PRON DEM SG DAT NEUTR"
"der" "PRON RELAT SG DAT MASK"
"das" "PRON RELAT SG DAT NEUTR"

In this case, the head noun *Tierhalter* and its dependent *dem* have only one shared feature set, namely ⟨MASK SG DAT⟩. All other feature sets do not unify and are thus removed in step 1. As a result, the head noun *Tierhalter* has only one feature set and has thus been successfully disambiguated (cf. Table 1).

Step 2: Upper phrase-level feature unification

In the second step, the context window for feature unification is wider than that in the first step: feature agreement within upper-level structures such as participle phrases, coordination structures and prepositional phrases is considered. Pattern-matching methods are employed to identify the phrase boundaries for feature unification.

In the first step, the four possible case features that Gertwol had returned for the noun *Tierhalterin* in sentence (1) have been reduced to two (i.e. DAT and GEN). In the second step, these remaining features are now compared with those of the coordinated noun *Tierhalter* (i.e. DAT). In the process, the genitive feature of *Tierhalterin* is discarded as it does not unify with any feature of *Tierhalter*. This results in the complete disambiguation of *Tierhalterin* (cf. Table 1).

Step 3: Clause-level feature unification

At the third step of morphosyntactic disambiguation, an even wider context window is considered: features at a clausal level are examined, including subject-verb agreement, the voice of predicates and the position of nouns in topological fields. The clause boundaries are determined by means of a

[2] S = noun, MASK = masculine, SG = singular, PL = plural, NOM = nominative, AKK = accusative, DAT = dative, GEN = genitive

[3] ART = article, DEF = definite, PRON = pronoun, DEM = demonstrative, RELAT = relative, NEUTR = neuter

Nr.	**Rule:** Heuristic	Feature(s)
1	**Every clause has a subject:** Select the nominative case of nouns if there are no other GF-candidates in the clause that could be in nominative case.	+ NOM
2	**A clause has only one subject:** Discard the nominative case of nouns if there is another nominative GF-candidate in the clause.	– NOM
3	**A clause has only one accusative object:** Discard the accusative case of nouns if there is another accusative GF-candidate in the clause	– AKK
4	**A clause has only one dative object:** Discard the dative case of nouns if there is another dative GF-candidate in the clause	– DAT
5	**A verb phrase in infinitive form does not have a subject:** Remove the nominative case of nouns that belong to an infinitive VP.	– NOM
6	**Passive sentences, adjective clauses and copula sentences do not contain any accusative nouns:** Remove the accusative case of GF-candidates in passive constructions, pseudo-passive constructions, adjective predicate structures and copula sentences.	– ACC
7	**Subjects agree with the finite verb:** Select the nominative case of nouns, pronouns and relative pronouns if there is no other GF-candidate that agrees with the finite verb of the clause.	+ NOM
8	**There is only one constituent in the vorfeld:** Remove the case features of vorfeld nouns if they are incompatible with the case features of other GF-candidates in the vorfeld.	All case features

Table 2: Hard syntactic rules with their heuristics applied in step 3.

constraint-based topological field recogniser (Sugisaki and Höfler, 2013).

To this aim, we have defined a set of heuristics that are based on hard syntactic constraints (Table 2). In these heuristics, nouns are categorised into two types: (1) GF-candidates (2) non-GF-candidates. Only dependents of predicates and heads of maximally projected noun phrases can be GF-candidates. Nouns dependent on prepositions, in contrast, cannot be GF-candidates. Coordinated nouns are regarded as one GF-candidate, since they constitute a single noun phrase.

For instance, our example (1) contains the main verb *berücksichtigt*. As illustrated in Table 1, only *Sie* and *Tiere* have not been completely disambiguated at the end of step 2. To disambiguate the pronoun *Sie*, its clausal context is now considered. Given that *Aufwand* and *Wohlergehen* are accusative, the pronoun should be nominative, since every clause (except for subject-less passive constructions) must have at least one GF-candidate in nominative case (i.e. subject). This corresponds to heuristic 1 in Table 2.

4 Evaluation

The strategies for morphosyntactic disambiguation presented in the previous section have been evaluated over 118 sentences (2,114 tokens, including 655 nouns and pronouns) that were randomly selected from the the Swiss Legislation Corpus (Höfler and Piotrowski, 2011). Each morphosyntactic analysis returned by the components described in this paper was examined against a manually annotated gold standard for the respective sentences.

As shown in Table 3, the system found 96.30% of the correct analyses (recall), and 67.60% of all analyses returned by the system were correct (precision). It thus achieved an F1-score of 79.43%.[4]

Since we are dealing with pre-processing components, it is more important to achieve good recall than to obtain high precision: wrongly removed correct morphosyntactic analyses (false negatives) cannot be restored, whereas any false positive can still be cast out later, e.g. by means of applying soft constraints. False negatives were caused mostly because appositions and coordination structures were not correctly recognized.

Additionally, we have evaluated the impact of the morphosyntactic disambiguation components on the reduction of ambiguity of nouns in law texts. We

[4] Recall has been calculated as the number of correct morphosyntactic analyses found by the *system* relative to the total number of morphosyntactic analyses present in the *gold standard*. The precision has been measured as the number of *correct* morphosyntactic analyses found by the system relative to the *total* number of morphosyntactic analyses returned by our system.

	1 analysis per token	more than 1 analysis per token
Input after preprocessing	148 (8.87%)	1,520 (91.12%)
Step 1: Local phrase-level feature unification	387 (23.20%)	1,281 (76.79%)
Step 2: Upper phrase-level feature unification	917 (54.97%)	751 (45.02%)
Step 3: Clause-level feature unification	1,129 (67.68%)	539 (32.31%)

Table 3: Number of unambiguous and ambiguous analyses after each disambiguation step.

based our evaluation method on work by Hinrichs and Trushkina (2004) in order to be able to compare our results. Hinrichs and Trushkina developed a rule-based morphosyntactic disambiguation system and tested the impact of each of its components on sentences extracted from a newspaper corpus (5,752 tokens). In comparison, the impact of each morphosyntactic disambiguation component described in this paper has been assessed on 239 sentences that were randomly selected from the Swiss Legislative Corpus (4,789 tokens).

Before disambiguation, these sentences contained an average of 4.20 morphosyntactic analyses per noun. After step 1 (local phrase-level feature unification), this number was reduced to 2.72. After step 2 (upper phrase-level feature unification), the number went further down to 1.86. Finally, an average of 1.60 readings per noun had been reached after step 3 (clause-level feature unification).

The described components thus reduced a major part of the ambiguous morphosyntactic analyses. Table 3 illustrates how many morphosyntactic analyses were reduced at each step. Before disambiguation, a majority of the nouns (91.12%) had more than one morphosyntactic analyses.[5] After step 1, the number of morphosyntactically ambiguous nouns was reduced to 79.79%; after step 2, this number fell to 45.02%; after step 3, only 32.31% of the nouns were ambiguous. These results show that the morphosyntactic disambiguation rules for maximally projected noun phrases (e.g. coordination structures) and prepositional phrases applied in step 2 of the system had the most impact on disambiguation.

To estimate the relevance of morphosyntactic disambiguation for parsing, we evaluated the disambiguation of case features in GF-candidates on the system's output data. As documented in Table 4, a majority of GF-candidates (56.49%) had a unique case feature after the completion of step 3, 33.20% of the GF-candidates had two case features and only very few had three or four case features (6,17% and 4.11%, respectively). An analysis of the ambiguous nouns revealed that most GF-candidates with two case features were not completely disambiguated because two arguments in the same clause were both morphologically underspecified. Most GF-candidates with three or four cases could not be disambiguated by the system because they were bare nouns or appositions.

	Tokens	%
1 case feature per token	439	56.49
2 case features per token	258	33.20
3 case features per token	48	6.17
4 case features per token	32	4.11
Total of GF-candidates	777	100

Table 4: Case ambiguity of GF-candidates after all morphological disambiguation steps.

The results of the current system are lower than those of the system presented by Hinrichs and Trushkina (2004), which yielded an unique morphosyntactic analysis for 77.08% of the nouns. One possible explanation for this fact is that Hinrichs and Trushkina (2004) used a different morphological analyzer, XRCE. However, the most plausible explanation may be found in the differences in the use and frequency of nouns in newspaper articles and law texts.

5 Conclusion

In this paper, we have shown that linguistically motivated hard rules are capable of achieving a substantial reduction of morphosyntactic ambiguity in nouns. In the evaluated sentences extracted from a corpus of German-language law texts, the rate of morphosyntactically ambiguous nouns could thus be reduced from 91.12% to 32.31%.

[5] Duplicates contained in the output of Gertwol were counted as one.

If employed as a pre-processing component, such a rule-based system for morphosyntactic disambiguation may thus be able to make a significant contribution to improving the quality of parsing, especially for morphologically rich languages like German and heavily nominalized domains like the domain of law texts. For the future, we are planning to extend our system by soft constraints for the further reduction of morphosyntactic ambiguity.

References

Bangalore, S. and Joshi, A. K. (1999). Supertagging: an approach to almost parsing. *Computational Linguistics*, 25(2).

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 112–116, Tronto, Italy. Association for Computational Linguistics.

Foth, K., By, T., and Menzel, W. (2010). Guiding a constraint dependency parser with supertags. In Bangalore, S. and Joshi, A. K., editors, *Supertagging: Using Complex Lexical Descriptions in Natural Language Processing*. MIT Press, Cambridge and London.

Foth, K. A., Daum, M., and Menzel, W. (2004). Parsing unrestricted german text with defeasible constraints. In *CSLP'04: Proceedings of the First international conference on Constraint Solving and Language Processing*. Springer-Verlag.

Haapalainen, M. and Majorin, A. (1994). GERTWOL: ein System zur automatischen Wortformerkennung deutscher Wörter. Technical report, Lingsoft, Inc.

Hansen-Schirra, S. and Neumann, S. (2004). Linguistische Verständlichmachung in der juristischen Realität. In Lerch, K. D., editor, *Die Sprache des Rechts: Recht verstehen: Verständlichkeit, Missverständlichkeit und Unverständlichkeit von Recht*, volume 1. Walter de Gruyter, Berlin.

Harper, M. P. and Wang, W. (2010). Constraint dependency grammars: Superarvs, language modeling, and parsing. In Bangalore, S. and Joshi, A. K., editors, *Supertagging: Using Complex Lexical Descriptions in Natural Language Processing*. MIT Press, Cambridge and London.

Hinrichs, E. W. and Trushkina, J. S. (2004). Forging agreement: Morphological disambiguation of noun phrases. *Research on Language and Computation*, 2(4):621–648.

Höfler, S. and Piotrowski, M. (2011). Building corpora for the philological study of Swiss legal texts. *Journal for Language Technology and Computational Linguistics (JLCL)*, 26(2):77–89.

Höfler, S. and Sugisaki, K. (2012). From drafting guideline to error detection: Automating style checking for legislative texts. In *EACL 2012: Proceedings of the Second Workshop on Computational Linguistics and Writing*, pages 9–18. Association for Computational Linguistics.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A., editors (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin/New York.

Loftsson, H. (2008). Tagging icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31:47–72.

McClosky, D. and Charniak, E. (2008). Self-Training for biomedical parsing. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (HLT-Short '08)*, pages 101–104. Association for Computational Linguistics.

Nasr, A. and Rambow, O. (2004). Supertagging and full parsing. In *Proceedings of the 7th International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG+7)*, pages 56–63, Vancouver, British Columbia, Canada.

Nussbaumer, M. (2009). Rhetorisch-stilistische Eigenschaften der Sprache des Rechtswesens. In *Rhetorik und Stilistik / Rhetoric and Stylistics*, Handbooks of Linguistics and Communication Science, pages 2132–2150. Mouton de Gruyter, New York/Berlin.

Sagae, K. (2010). Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP 2010)*, pages 37–44. Association for Computational Linguistics.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop.*, Dublin, Ireland.

Schneider, G. and Volk, M. (1998). Adding manual constraints and lexical look-up to a Brill-tagger for German. In *Proceedings of the ESSLLI-98 Workshop on Recent Advances in Corpus Annotation.*

Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A new hybrid dependency parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124.

Sugisaki, K. and Höfler, S. (2013). Verbal morphosyntactic disambiguation through topological field recognition in german-language law texts. In *Third International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2013)*, Berlin, Germany. (to appear).

Uí Dhonnchadha, E. (2006). *Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar*. PhD thesis, Dublin City University.

Volk, M. and Schneider, G. (1998). Comparing a statistical and a rule-based tagger for German. In *Proceeding of the 4th Conference on Natural Language Processing (KONVENS98)*, pages 125–137.

Voutilainen, A. (1993). NPtool, a detector of English noun phrases. In *Proceeding of Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 48–57, Wagenigen, Netherlands.

Voutilainen, A. (1995). A syntax-based part-of-speech analyser. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics (EACL '95)*, pages 157–164, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.