Customizing GermaNet for the Use in Deep Linguistic Processing

Melanie Siegel
LT-Lab, DFKI
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
siegel@dfki.de

Feiyu Xu
LT-Lab, DFKI
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
feiyu@dfki.de

Günter Neumann
LT-Lab, DFKI
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
neumann@dfki.de

Paper ID:

Keywords: Lexical resource customization, GermaNet, HPSG Grammar, German

Contact Author: Feiyu Xu

Under consideration for other conferences (specify)? no

Abstract

In this paper we show an approach to the customization of GermaNet to the German HPSG grammar lexicon developed in the Verbmobil project. GermaNet has a broad coverage of the German base vocabulary and fine grained semantic classification, while the HPSG grammar lexicon is comparatively small und has a coarse-grained semantic classification. In our approach, we have developed a mapping algorithm to relate the synsets in GermaNet with the semantic sorts in HPSG. The evaluation result shows that this approach is useful for the lexical extension of our deep grammar development to cope with real-world text understanding.

Customizing GermaNet for the Use in Deep Linguistic Processing

Abstract

In this paper we show an approach to the customization of GermaNet to the German HPSG grammar lexicon developed in the Verbmobil project. GermaNet has a broad coverage of the German base vocabulary and fine-grained semantic classification, while the HPSG grammar lexicon is comparatively small und has a coarse-grained semantic classification. In our approach, we have developed a mapping algorithm to relate the synsets in GermaNet with the semantic sorts in HPSG. The evaluation result shows that this approach is useful for the lexical extension of our deep grammar development to cope with real-world text understanding.

Introduction

The lexical-semantic information encoded in online ontologies like WordNet (Miller et al., 1993), GermaNet (Hamp and Feldweg, 1997) and EuroWordNets (Vossen, 1998) is very useful for different natural language applications: information extraction, lexical acquisition and intelligent information retrieval. In this paper, we provide an approach, which customizes the GermaNet lexical semantic information to the HPSG lexicon in order to extend the lexicon for the improvement of the deep linguistic processing of real- world text.

In the DFKI project Whiteboard, we aim to integrate different natural language resources to deal with real-world text understanding. One particular goal is the integration of deep NLP (DNLP) and shallow NLP (SNLP). In recent years, a number of efforts have been spent towards the increase of parsing performance with HPSG (Flickinger et al., 2000). Especially the PET parser developed at the CL department at the University of the Saarland has demonstrated that it is now possible to use an HPSG parser for processing of real-world text using large Grammars for German and English. However, one of the bottlenecks with real-text processing is the high amount of very productive domain-specific lexical entities. In order to cope with this problem, one possibility would be to extend the HPSG lexicon. However, this would increase the search space enormously and could degrade the performance of the HPSG parser. Another possibility would be to let a domain-specific shallow component do the main lexical processing and integrate the lexical entities via the HPSG type system. This is actually the approach followed in the Whiteboard project. In our current system the shallow text processor SPPC (Piskorski and Neumann, 2000) is used for lexical processing. Among others SPPC performs morphological processing (including online compounding), POS disambiguation and highly accurate Named Entity (NE) recognition including NE-reference resolution (overall 85% recall and 95.77% precision using MUC-style NE classes). The PET system is then called with the results of the SPPC lexical processor to perform an HPSG analysis (since PET expects a sentence as input, SPPC has been augmented with a very simple, but effective sentence boundary recognizer). The integration of the SPPC and PET system is based on the HPSG type system. For example, in order to make use of the NE results computed by SPPC, the different NE types (person names, localization, company names etc) are mapped to the corresponding HPSG types of the deep HPSG grammar.

In our current system we have applied the German grammar developed in the Verbmobil project (Müller and Kasper 2000), which originally aimed to understand and translate dialogue language, to economic news. The result was that apart from NE's, 78.49% of the missing lexical items are nouns. Due to the integration of SPPC, NE recognition as well as coverage of nouns is now increased. However, SPPC only computes POS and morpho-syntactic information. But for the deep text analysis, a solution for retrieving nouns with their semantic sorts is essential, because the semantic sorts are useful for the semantic construction and for providing semantically based selectional restrictions, which are essential for guiding the search space defined by the HPSG grammar. GermaNet has a huge coverage of German word stems and the words are tagged with the POS information and their semantic classification. Therefore, we did experiments to automatically convert the semantic concepts in GermaNet to the semantic sorts defined in the HPSG lexicon. We have implemented an algorithm that computes the mapping relevance from a semantic concept in GermaNet to a semantic sort in the HPSG lexicon. In addition, we have also developed a *GermaNet2HPSG* tool which can not only be used for the online text analysis by assigning a word to the most adequate HPSG semantic sorts based on its GermaNet concepts, but can also be used for the offline lexicon generation. The GermaNet2HPSG tool is based on the Whiteboard Germa/WordNet ontology inference tool, which supports the search and navigation of the ontology information in Germa/WordNet.

The remainder of the paper is organized as follows: Section 1 will give a brief introduction to the GermaNet. In section 2, we describe our ontology inference tool. The main approach of the customization of GermaNet to the HPSG grammar lexicon and its evaluation is explained in section 3. The implementation of the approach is shown in section 4. In section 5, we conclude with some ideas for further experiments in the near future.

1 GermaNet

Compared to the huge amount of online English linguistic resources, there are not so many large-scale German lexicons like GermaNet which has properly modelled the lexical syntactic and semantic information. Therefore, GermaNet appears to us as a valuable resource to extend our lexicon.

GermaNet is a lexical semantic net for German, developed at the university of Tübingen. It is mainly based on the WordNet framework, containing about 10.652 nouns, 6.904 verbs and 1.657 adjectives. One big advantage of the GermaNet is that the semantic classification of the words is very fine-grained. Like in WordNet, a semantic concept (so-called *synset*) is represented by a group of words. There are 19.213 synsets in GermaNet and in addition 24.920 synonyms in synsets. The synsets are connected through their lexical and conceptual relations. The basic lexical relations are *synonymy*, *antonymy* and *pertains to*, while the conceptual relations are *hyponymy* ('is-a'), *meronymy* ('has-a'), *entailment* and *cause*. The hyponymy relation generates a hierarchical semantic structure of the GermaNet. Compared to WordNet, verbs in GermaNet are annotated additionally with *selectional restrictions*, which are important for the deep natural language processing.

2 Inference Tool

GermaNet itself provides a simple search interface that allows to search for the relations assigned to one word. However, this search interface is still too restricted to be directly usable for the different applications that are explored in the Whiteboard project (e.g., information extraction (Neumann et al. 1997) and grammar and controlled language checking (Bredenkamp et al., 2000)). We decided to build a flexible inference tool, in order to access the lexical content and semantic relations defined between the concepts of a set of words. With the help of such an inference tool, we can easily build new applications which need lexical semantic information.

We have inserted the GermaNet content into a relational database. After this step, we can make use of the search functions provided by the database server. Three different functions have now been implemented in our inference tool:

- Retrieval of relations assigned to one word
- Retrieval of relations between two words
- Flexible navigation in the GermaNet graph starting from a certain node with

search depth and search relationship as arguments

The first search function is actually a reimplementation of the search interface existing in the GermaNet. For example, a query is 'find all synonyms of the German word *Bank*'. For the first sense *bench*, we find the word `Sitzmöbel` (engl. sitting furniture) as its synonym. For the sense corresponding to *financial institution*, its synomyms are `Geldinstitut` (engl. money institution) and `wirtschaftliche Institution` (engl. financial institution).

The second type of functions is to search for and test the relations between two words. This search type provides important information like 'is-a' and 'has-a' relation between words, which supports the coreference resolution between terms in the information extraction application. Let us give a simple example. We would like to know the relationship between the word 'Internet-Service-Provider' and the word 'Firma' (engl. company). Our search tool tells us that the 'Internet-Service-Provider' is a hyponym of the word 'Firma'. It indicates that the first word is a subconcept of the second one.

Furthermore, we have implemented search functions which take the search depth as an optional argument to guide the navigation in GermaNet. With the help of our inference tool, we have worked out our first approach to the customization of GermaNet to the HPSG lexicon outlined below.

3 Customization of GermaNet to the deep grammar

3.1 Motivation

As mentioned above, the main problem of the adaptation of a general deep grammar to a new domain and application of a deep grammar to real-world text is the lexical coverage. Compared to GermaNet, the lexicon of the German HPSG developed in the Verbmobil project is fairly small. For example, it contains only about 3630 nouns. GermaNet has more than 10,000 nouns. Therefore, the integration of the GermaNet lexicon and the deep grammar lexicon is an important solution for lexical extension.

We started our GermaNet customization with nouns, as 78.49% of missing lexical items were nouns, according to our evaluation in the economic news domain. The lexical items of nouns in the HPSG grammar lexicon need not only stem information, but also the semantic sort information. Compared to the semantic classification in GermaNet, the semantic classification of nouns in the Verbmobil grammar is much more coarse-grained. Therefore, our main work is to map the fine-grained synsets in GermaNet to the coarse-grained semantic sorts in the Verbmobil grammar lexicon.

3.2 Basic Idea

The core idea is to first learn a mapping between the two different semantic classifications that is later used to automatically compute the semantic sorts of words that are not contained in the HPSG lexicon using the corresponding GermaNet classification. The training material for the learning process are those words that are annotated with the semantic sorts of the deep grammar and that at the same time have an annotation of GermaNet synsets. We used these words as an annotated training corpus and reasoned about the relations of GermaNet synsets and HPSG semsorts.

3.2 SemDb versus GermaNet

The semantic database (SemDb) (Bos et al. 1996) in the HPSG lexicon was set up in the Verbmobil project used in different modules. The HPSG grammar makes use of the SemDb in order to restrict and disambiguate readings via sortal restrictions on verbal arguments. It contains words and their semantic sorts as well as valence information and sortal restrictions of arguments. The semantic sorts are organized in a hierarchy. The German semantic database contains about 7800 words. Although the hierarchy is quite simple, it turned out to be very useful in the parsing process.

Let us consider the relationships between the semantic sorts and the synsets in more detail. On the one hand, there are 30 different sorts in this hierarchy as opposed to almost 20.000 synsets in the GermaNet ontology. On the other hand, each single word is annotated with one semantic sort in the SemDb and different sets of synsets in GermaNet. For example, examine the word “Kind” (engl. child). The SemDb gives the sort `human`; and GermaNet gives the following two sets of synsets. For the first sense, where ‘Kind’ means *young human*, its hypernyms (synsets which are its superconcepts) are as follows:

```
Kind
  => junger Mensch
   => alternder Mensch
    => Mensch,Person
     => höheres Lebewesen
      => natürliches Lebewesen,
             Organismus
       => Lebewesen, Kreatur
        => Objekt
```

For the second sense, where it means descendant, its hypernyms are:

```
Nachkomme, Kind, Nachfahre,
 Nachkömmling, Sproß, Sprößling
  => Verwandter, verwandter Mensch,
      Familienangehöriger,
      Familienangehörige,
      Angehöriger, Angehörige
   => Mitmensch
    => Mensch, Person,
        Persönlichkeit, Individuum
      => höheres Lebewesen
       => natürliches Lebewesen,
             Organismus
        => Lebewesen, Kreatur
         => Objekt
```

It is thus obvious that there cannot be a direct match from SemDb sorts to GermaNet synsets. We therefore decided to learn the relationships between the semantic sorts and the synsets.

3.3 Training Method

Using the nouns with semantic sort annotations from the SemDb as our training corpus, we developed a mapping algorithm from semantic sorts to synsets:

1) *Retrieve the hypernyms (synsets) in GermaNet of all nouns in the SemDb.*
2) *Count the frequency (f_{ij}) of each GermaNet synset$_i$ for all words in a certain HPSG semsort$_j$.*
3) *Compute the sum (F_i) of the frequencies of each GermaNet synset$_i$ for all HPSG semsorts in the corpus.*

$$F_i = \sum_{j=1}^{|semsorts|} f_{ij}$$

4) *Compute the mapping relevance (R_{ij}) of a GermaNet synset$_i$ to a certain HPSG semsort$_j$ with respect to the whole training data.*

$$R_{ij} = \frac{f_{ij}}{F_i}$$

The training results in a table of SemDb sorts and GermaNet synsets annotated with their mapping relevance; see the following example which shows the mapping from the synset ‘Stelle, Ort, Stätte’ (engl. place, room) and the synset ‘Äußerung’ (engl. uttrance) to the semantic sorts.

Synset	Semantic Sort	Mapping Relevance (%)
Stelle,Ort,Stätte	*Symbol*	*0.60*
Stelle,Ort,Stätte	*geo_location*	*3.01*
Stelle,Ort,Stätte	*Location*	*6.02*
Stelle,Ort,Stätte	*nogeo_location*	*44.58*
Äußerung	*Field*	*2.63*
Äußerung	*abstract*	*15.79*
Äußerung	*info-content*	*21.05*
Äußerung	*communication situation*	*23.68*

3.4 The Annotation of Words with SemDb sorts

Using the mapping table, words not contained in the SemDb can now be annotated with semantic sorts used in the deep grammar. The annotation algorithm works as follows:

1) *Retrieve the hypernyms (synsets) in GermaNet of a word; different senses have different sets of synsets.*
2) *For each sense,*
   i) sum the mapping relevance weights from its GermaNet synsets to semantic sorts.
   ii) Select the best four mappings

The result is an ordered list of semantic sorts with relevance values. A word that has more than one sense in GermaNet will also obtain more than one list of semantic sorts.

3.6 Evaluation

We examined a corpus of 4664 nouns extracted from economic news (Wirtschaftswoche 1992) that were not contained in the SemDb. 2312 of them are known for GermaNet. They obtain 2811 senses according to the GermaNet and were automatically annotated with semantic sorts. The evaluation of the annotation accuracy yields encouraging results:

- In 76.52% of the cases the computed sort with the highest processed probability was correct.
  For example, the word ‘Kanzler’ (engl. cancellor) is annotated with the following semantic sorts:

Semantic Sort	*Relevance*
Human	*941.7099*
Animal	*63.92*
Thing	*40.18*
Object	*25.800003*

  It is clear that the first semantic sort is also the most adequate one.

- In 20.70% of the cases, the correct sort was one of the next three sorts.
  For example, the second semantic sort below is the best annotation for the word ‘Mannschaft’ (engl. team):

Semantic Sort	Relevance
nongeo_location	*73.35*
Institution	*71.01*
Human	*58.01*
Abstract	*29.01*

- In 2.74% of the cases, the first four computed sorts did not contain the correct one.

This means that the accuracy among the first four annotations is 96.52%. However, we need to improve the accuracy of the first reading. One of the reasons for errors is given the size of HPSG lexicon and therefore our mapping table is incomplete: The training corpus was small and only parts of the synsets are considered during the training phase. Therefore, not all synsets can be related to the semantic sorts. During the annotation phase, the specific synsets that are unknown for the mapping table are still ignored. We will consider this issue in the future work.

4 Implementation

The customization tool makes use of the Whiteboard Germa/WordNet inference tool. We call it Germa/WordNet inference tool because it can also be applied to retrieve the lexical semantic information in WordNet. The WordNet content has been inserted into the relational database MySQL too. Both GermaNet and WordNet share the same database design. The two tools are implemented in JAVA with JDBC access to MySQL. The GermaNet2HPSG component has already been integrated to the Whiteboard text processing server. It supports the deep text processing by assigning online the semantic sort to a word based on the GermaNet synsets. The advantage is that we do not need convert the entire GermaNet lexicon to the deep analysis lexicon. It reduces the online lexicon search and provides only the semantic sort when it is needed.

Conclusion and Outlook

We have built a tool to automatically map GermaNet synsets to semantic sorts of the kind used in a deep HPSG grammar. The mapping result is used in a system that integrates deep and shallow processing for retrieving semantic sorts of nouns not contained in the deep lexicon. In order to extend the accuracy of the mapping table, we plan to use the evaluated annotation for the expansion of the training corpus. A next step will be the application to verbs and adjectives.

We are planning to combine the information of the NEGRA treebank (Brants, 2000) with the GermaNet ontology in order to gain information about the valence and sortal restrictions of verbs. In order to extend the grammar coverage we are thinking of refining the HPSG semantic database ontology by using the GermaNet ontology.

References

Bos, J., M. Schiehlen and M. Egg (1996). *Definition of the Abstract Semantic Classes for the Verbmobil Forschungsprototyp 1.0*. Universität des Saarlandes, IBM Heidelberg, Universität Stuttgart. Verbmobil-Report 165.

Brants, Thorsten (2000). *Inter-Annotator Agreement for a German Newspaper Corpus*. In Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece.

Bredenkamp, Andrew, Berthold Crysmann and Mirela Petrea (2000): *Looking for Errors: A declarative formalism for resource-adaptive language checking*, Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece.

Flickinger, D., S. Oepen, H. Uszkoreit and J. Tsuji (Eds.) (2000). *Special Issue on Efficient processing with HPSG: Methods, Systems, Evaluation*. Journal of Natural Language Engineering 6 (2000) 1. Cambridge, UK: Cambridge University Press. (in press)

Hamp, B. and H. Feldweg (1997) *GermaNet - a Lexical-Semantic Net for German*. In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications". Madrid, 1997

Miller, G. A., R. Beckwith, C. Fellbaum. D. Gross and K. Miller (1993). *Five Papers on WordNet*. Technical Report, Cognitive Science Laboratory, Princeton University, 1993.

Müller, S. and W. Kasper (2000). *HPSG Analysis of German.* In “Verbmobil: Foundations of Speech-to-Speech Translation”, W. Wahlster, ed., Springer Verlag, Berlin, 238-253.

Neumann, G., R. Backofen, J. Baur, M. Becker and C. Braun (1997). *An Information Extraction Core System for Real World German Text Processing.* In Proceedings of 5th ANLP, Washington, March, 1997.

Piskorski, J. and G. Neumann (2000). *An Intelligent Text Extraction and Navigation System*. In the Proceedings of RIAO 2000 - Content-Based Multimedia Information Access, Paris, France.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks,* P. Vossen. Ed., Kluwer Academic Publishers, Dordrecht.

Wahlster, W. (Ed.) (2000) *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer Verlag, Berlin.