Corpus based Enrichment of GermaNet Verb Frames

Manuela Kunze and Dietmar Rösner

Otto-von-Guericke-Universität Magdeburg
Institut für Wissens- und Sprachverarbeitung
P.O. Box 4120, D–39016 Magdeburg,
Germany
makunze, roesner@iws.cs.uni-magdeburg.de

Abstract

Lexical semantic resources, like WordNet, are often used in real applications of natural language document processing. For example, we integrated GermaNet in our document suite XDOC. In addition to hypernymy and synonymy relations, we want to exploit GermaNet verb frames for our analysis. In this paper, we outline an approach for the domain related enrichment of GermaNet verb frames by corpus based syntactic and co-occurrence data analyses of real documents.

1. Introduction

Lexical resources, like WordNet (Fellbaum, 1998), GermaNet (Hamp and Feldweg, 1997; Kunze, 2001), or EuroWordNet, will be more than ever applied as resources in applications, like Text Mining and Data Mining. Often it is necessary to extend or to adapt the resources for the application (see also (Vossen, 2001; Navigli and Velardi, 2002)).

In (Kunze and Rösner, 2003) we presented the integration of the lexical resource GermaNet into our document processing system XDOC. In XDOC, GermaNet is used as linguistic (lexical) resource for tasks like

- semantic tagging of tokens,
- case frame analysis, and
- semantic interpretation of syntactic structures (SIsS).

One problem for the integration of GermaNet resources was the usage of GermaNet's verb frames. The problem there was that the information encoded in GermaNet verb frames is not sufficient (i.e., not detailed enough) for usage within the case frame analysis of XDOC.

Case frame analysis needs detailed information about the syntactic form (e.g., which preposition and which case of the PP), the semantic category of the filler of a relation, and which thematic role is described by the relation (e.g., agent, location, etc.) for a complement of a verb.

GermaNet's verb frames have two deficits with respect to usage in XDOC:

1. for verbs, the information given is incomplete (e.g., preposition, semantic category, and thematic role are missing) and
2. for nouns, no frame information is available.

The information missing can be classified into several types: lexical (preposition), syntactic (case of noun phrase in a preposition phrase[1]) and semantic (category of the filler and the thematic role of the relation described). The creation or manual adaption of GermaNet's resources is time-consuming. Related works included the automatic building of subcategorisation lexicons for German verbs (Wauschkuhn, 1999; Schulte im Walde, 2002) and the automatic identification of thematic roles (see (Gildea and Hockenmaier, 2003); (Gildea and Jurafsky, 2002)) by exploitation of a syntactically parsed corpus as input data. The method described in this paper extracts the necessary syntactic information and information about the required semantic categories for possible complements of a verb. This approach uses a corpus annotated with chunks (noun phrases and prepositional phrases) and GermaNet's verb frames information.

This paper is organized as follows: first we give a short description about the evaluation environment. After this, we present the methodology by discussing an example. This is followed by the presentation and discussion of results from our experiments.

2. Evaluation Environment

2.1. Evaluation Corpus

For our work we used a corpus of medical documents in German (forensic autopsy protocols) with more than 1 million running word forms. The autopsy protocols have a strictly defined content and layout. They are separated into different document parts, e.g. findings, background, discussion, death causes, etc. Each document part has its own characteristics (sub-language).

The analyses with XDOC are concentrated on the sections of *findings*, *background* and *discussion*. The *findings* section contains a high ratio of nouns and adjectives and syntactic (sentence) structures are mostly short. This section describes the medical findings in an everyday vocabulary without domain specific (medical) terms. A standard distribution of all word classes and regular syntactic structures occurs in the *background* and *discussion* sections. The *background* section describes, for example, the details of a traffic accident, while the section *discussion* contains a combination of the results of the *finding* section and the facts reported in the *background* section. Both document parts contains a high and multifaceted number of named entities (NE). For example, each forensic autopsy protocol has a registration number (e.g., G 123/45), which is often referred to in the document. Furthermore, other NEs like

[1] For example, the preposition *in* can required a NP with the case *accusative* or *dative*.

date specifications, names of locations (e.g., streets, cities), or names of persons etc., occur in the texts.

The analysis described in this paper is only concentrated on the document parts *background* and *discussion*. These parts were chosen, because both document parts contain regular syntactic structures of German and a minor ratio of domain specific terms. The tokens seem to belong much more likely to everyday language than to a sub-language of a specific domain.

2.2. Tools and Resources

For the adaptation of GermaNet verb frames, a syntactically annotated corpus is required. For this and other preprocessing steps, the document suite XDOC[2] is used. In particular, following preprocessing steps of XDOC are used:

- sentence splitter,
- POS tagger,
- syntactic parser.

These methods output their results as XML structures, which are accepted by the subsequent processing steps based on XML structures as their input. For the extraction of relevant information (syntactic structures), XSL transformation is applied (Clark, 1999).

The quality of expected results is strongly dependent on the quality of input data, especially the results of the chart parser. The syntactic parser of XDOC is a bottom-up chart parser, which works with a context free grammar for German (ca. 400 rules). The robust XDOC syntactic parser outputs sentences completely parsed (readings) or only structures partially parsed (coverings). In these coverings, basic structures, like noun phrases or prepositional phrases (frequent elements in GermanNet verb frames) are annotated by XDOC's parser (see Fig. 1).

3. An Outline of the Approach

The basic assumption for the approach is that in a corpus with similar texts (news, expert's report, abstracts, etc.) a frequent verb co-occurs with the same complements. The complements of such a verb often appear in a similar syntactic structure at the same position in a sentence. Further, the fillers have the same semantic category. In the case of the autopsy corpus the number of authors of the documents is small. This results in a high rate of repetition of specific wordings or phrases, because authors have the tendency to use the same phrase for the description of similar facts (author style).

The steps of the procedure are:

- use the verb frames given by GermaNet as simple patterns for the recognition of potential candidates for case frames in the corpus,
- extract information about prepositions used in a complement (element of the case frame candidate), and
- count the occurrences of similar semantic fillers (roles) for an element of the case frame.

The approach is presented by considering the verbs: *verstarb* (to pass away), *kollidieren* (to collide), *befahren* (to cruise), *operieren* (to operate) as examples.

verb	occurrences	frame information from GermaNet
kollidieren	34	NN.Pp
operieren	14	NN.AN or NN.AN.BL
versterben	128	NN.BT
erfolgen	187	NE.AN or NN.PP
befahren	59	NN.AN or NN.AN.AZ or NN.AN.BM
ereignen	29	NE or NE.AR or NN.AR.BT or NN.BL

Table 1: Verbs and their GermaNet verb frame information.

All these verbs have only one sense in GermaNet with the correct meaning for our cases. But multiple verb frames can be assigned to a sense, see for example the verb *befahren* (see also table 1)[3] with 3 verb frames.

GermaNet verb frames characterise the syntactic subcategorisation of a verb. The elements of a verb frame describe different complements of verbs. For example, '*NN*' or '*AN*' stand for a noun phrase in case *nominative* resp. *accusative*, '*PP*' represents a prepositional phrase but without information about the concrete preposition used. Both elements give no information about the semantic category of the role filler and the thematic role. Only elements like, '*BM*' or '*BL*' (stands for an adverbial complement or a prepositional phrase, which indicates a *manner* or *local* complement), give some semantic restriction for the filler.

In the following, we do sketch the approach: The initial basis is a corpus of documents, which are separated into a sequence of sentences.

To complete the case frame analysis of a frequent verb in the corpus, all sentences in which the verb occurs were selected. These sentences are annotated by the POS Tagger of XDOC and are parsed by the syntactic parser of XDOC. For the analysis, only the annotation of NPs and PPs is required. In this case, the grammar of the chart parser was reduced to 15 rules for the annotation of basic structures, like noun phrases and prepositional phrases.

According to the verb frame information of GermaNet, possible candidates are selected from the structures parsed by the usage of XSL transformations. For example, for elements like *NN* and *AN*, noun phrases with the case nominative or accusative resp. are selected. Elements like *PP* [4] are prepositional phrases with non specified case or preposition. Other elements are ambiguous, like the element *BM*. The syntactic realisation of *BM* could be a prepositional phrase or an adverb. In this case, both realizations must be considered during search and analysis.

The GermaNet verb frame of the verb *collide* contains the following information: 'NN.Pp' – a noun phrase in case *nominative* and an *optional* prepositional phrase. Following sentences occur in the corpus:

[2]For a full description of the methods inside XDOC see (Rösner and Kunze, 2002).

[3]A detailed description of the notation of the verb frames is available at: http://www.sfs.nphil.uni-tuebingen.de/lsd/

[4]The GermaNet notation 'PP' means a required prepositional phrase.

```
<COVERING NR="1">
 <XXX>Beide</XXX>
 <V ROOT="befind" FLEX="FIN">befanden</V>
 <REFPRO>sich</REFPRO>
 <PP RULE="PP1" CAS="DAT">
  <PRP CAS="DAT">am</PRP>
  <NP TYPE="FULL" RULE="NP1" CAS="DAT" NUM="PL" GEN="_">
   <ADJ>rechten</ADJ>
   <N SRC="UC1">Fahrbahnrand</N>
  </NP>
 </PP>
 <IP>,</IP>
 <S-KONJ>als</S-KONJ>
 <DETI>ein</DETI>
 <NE>Mazda</NE>
 <NR>323</NR>
 <ADJP RULE="ADJP1">
  <XXX AS="ADV">beide</XXX>
  <XXX AS="ADJ">ueberholte</XXX>
 </ADJP>
 <K-KONJ>und</K-KONJ>
 <ADV>dabei</ADV>
 <PP RULE="PP1" CAS="DAT">
  <PRP CAS="DAT">mit</PRP>
  <NP TYPE="FULL" RULE="NP2" CAS="DAT" NUM="SG" GEN="NTR">
   <DETD>dem</DETD>
   <N SRC="UC1">Radfahrer</N>
  </NP>
 </PP>
 <V ROOT="kollidier" FLEX="FIN">kollidierte</V>
 <IP>.</IP>
</COVERING>
```

Figure 1: A syntactically parsed sentence with NPs and PPs chunks.

- *Der erste Hänger kollidierte vermutlich mit der vorderen rechten Seite mit einem ... Haus.*
- *... sein LKW kollidierte mit dem PKW.*
- *Der Pkw ... kollidierte mit 3 Begrenzungsstäben.*
- *Der ... Pkw Peugeot hingegen kollidierte frontal mit dem Pkw Renault*
- *Nachfolgend kollidierten 3 Pkw mit dem VW Golf.*

The first assignment ('NN') is easy to handle, all noun phrases with the case *nominative* are selected. The following instances can be assigned to the 'NN' element of the verb frame given the sentences above:

- *NN: der erste Hänger, sein LKW, der Pkw, der Pkw Peugeot, 3 Pkw*

Semantic classification uses information available in GermaNet. In the first step, GermaNet top level information is used as a shallow classification. To improve the classification, the hypernymy tree information of GermaNet is exploited.

For the verb *collide*, following two types of fillers for the 'NN' element are encountered in the corpus. The first type is a person referenced by pronouns (11), registration numbers (6), like *G 1234/11*, or by a person name (1). And the second type of the filler is a vehicle (16). Both types describe traffic participants (road users).

In the next step, the occurrences of prepositions in these sentences are counted. The verb *collide* occurs in 34 sentences within the corpus. The frequent prepositions used in these sentences are presented in table 2.

preposition	ratio	semantic
mit (dat)	54	temporal, instrument, modal, causal
auf (dat, acc)	23	local, temporal, modal, causal
aus (dat)	17	local, causal
am (dat)	17	modal, temporal
nach (dat)	15	local, temporal, final, modal
in (dat, acc)	13	local, temporal, modal
von (dat)	11	local, temporal, modal

Table 2: High frequent prepositions co-occurred with the verb *collide*.

Table 2 shows, that different prepositions are possible as the indicator for the prepositional phrase of the verb frame. For the selection of the correct preposition, following assumptions are used. The PPs co-occurring with instances of the verb *kollidieren* (collide) are evaluated. Most prepositions allow for different semantic interpretation (cf. 2). Disambiguation is only possible when taking the classification of the embedded NP into account. Only PPs that are not referring to local or temporal circumstances are counted, because temporal or local adjuncts can co-occur with most verbs.[5] From the remaining PPs the preposition with the highest frequency is taken as candidate for the derived case frame.

Furthermore, this approach can be enhanced, when the information about the distance between verb and potential prepositional phrase complement is exploited. In a sentence, it is possible that the same preposition can occur more than once in a PP. Coordination is one of example:

- *Nach Angaben der anwesenden Kliniker soll er **mit** einem PKW von der Fahrbahn abgekommen sein und dort **mit** feststehenden Gegenständen kollidiert sein.*

For examples like this, the approach described can be enhanced in the following way: Only clause structures instead of whole sentences are explored, because more than one verb occur in a sentence. The clauses in these sentences are splitted by commas or by the conjunction '*und*'(and). In addition, a heuristic is used: Only PPs situated next to the verb, before or after the verb (scope of a verb), are analysed. Further work will be the refinement of this simple heurisics.

For the verb *collide* the following prepositions were encountered as results: 27 times 'mit', 5 times 'nach', twice the prepositions 'beim', 'am', 'als, and once the preposition 'in'.

The filler of the 'Pp' with the preposition 'mit' can be assigned to the semantic category 'solid object'.

- *Pp: mit einem Pkw, mit einem Baum, mit dem Mercedes, mit der Mittelleitplanke, mit einem Verkehrsschild*

In sum, the approach results in the following extended verb frame of the verb *collide*:

[5] The treament is different if the GermaNet patterns contain explicit elements about locale or temporal information, like *BT* or *BL*.

- The filler of 'NN' element can be either a person (e.g., NE, pronoun) or a vehicle (e.g., a 'regular' noun, like *PKW*).
- The 'Pp' element describes an object in the syntactic form of a PP with the preposition 'mit' and the case *dative*. In this case, the semantic category of the filler is the category *solid object*.

For the verb *befahren* there exist three verb frames in GermaNet, each consisting of the elements *NN* and *AN*. For these elements, the approach described above delivers following details:

- *NN*: The results contain here again instances of traffic participants. The first subcategory describes persons, in form of NEs (registration number or name of the person), pronouns, or with the noun 'driver' in phrases like *driver of the 'car'*. The second subcategory is presented through vehicles, like NEs (*PKW VW Lupo*) or as nouns, like car, tramway, motor vessel etc.
- *AN*: All possible candidates (e.g., street, German freeway, avenue, canal, etc.), which are found in GermaNet, could be assigned to *traffic route*.

The additional elements *AZ*[6] and *BM* were not analysed, because up to now this work was restricted to the enrichment of elements, like noun and prepositional phrases. The extension to other possible elements in a verb frame is part of our future work.

4. Discussion

The results obtained via this approach can support a designer for verb frames. Based on the verb frame information in GermaNet, this method delivers possible candidates of fillers for noun and prepositional phrases . The results contains information about the semantic category and syntactic form (and elements, like prepositions) of case roles fillers.

The results are dependent on a good lexical coverage to get the correct semantic information for a filler and strongly dependent on the correct annotation of syntactic structures. The number of the coverings delivered for a sentence by the parser can be reduced. At first only coverings are allowed, which are in accordance with GermaNet verb frames. Second, only relevant parts (clause) of a complex sentence are extracted.

One problem, which occurs was the correct handling of NEs in the corpus. In addition to date or time information, the approach must cover names of locations (e.g., streets, like *A 9*), names of vehicles (e.g., *Opel Frontera*), names of persons (e.g., *Mr. Miller'*), and registration numbers (e.g., of persons: *G 1345/78*; or licence plate numbers: *ABZ AB-789*).

5. Conclusion

In this paper, we described an approach for the enrichment of GermaNet's verb frames. It is based on co-occurrence data analyses of a corpus of forensic autopsy protocols.

For the approach described above, the document parts *background* and *discussion* from the forensic autopsy protocol were used. These parts were chosen, because in these parts, a minor ratio of domain specific terms occurs. The form and the content of the *background* section are similar to a report in a newspaper.

The approach outlined here is based on structural and syntactical analysis and on the analysis of co-occurrence data. These co-occurrence data were verbs and syntactical structures in the neighborhood of the verbs. The quality of the results are strongly dependent on the results of the syntactic parser and the correct handling of named entities. Both can be enhanced by an improvement of the domain specific resources, like the grammar of the chart parser. Our future work will be to confirm and to evaluate the approach with another corpus, for example the EUROPARL corpus – available at http://www.isi.edu/∼koehn.

6. References

Clark, J., 1999. XSL Transformations (XSLT) Version 1.0. W3c recommendation, World Wide Web Consortium. URL:http//www.w3.org/TR/xslt.

Fellbaum, C., 1998. *WordNet: An Electronic Lexical Database*. Mass.: MIT Press.

Gildea, Daniel and Julia Hockenmaier, 2003. Identifying semantic roles using combinatory categorial grammar. In *2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sapporo, Japan.

Gildea, Daniel and Daniel Jurafsky, 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Hamp, B. and H. Feldweg, 1997. GermaNet – a lexical-semantic Net for German. In P. Vossen et.al. (ed.), *Proc. of ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid.

Kunze, C., 2001. *Lexikalisch-semantische Wortnetze*. Heidelberg; Berlin: Spektrum, Akademischer Verlag, pages 386–393.

Kunze, M. and D. Rösner, 2003. Issues in Exploiting GermaNet as a Resource in Real Applications. In *GermaNet-Workshop: Anwendungen des deutschen Wortnetzes in Theorie und Praxis*. Tübingen, Germany.

Navigli, R. and P. Velardi, 2002. Automatic Adaption of WordNet to Domains. In *Proc. of LREC 2002*. Las Palmas.

Rösner, D. and M. Kunze, 2002. An XML Based Document Suite. In *Coling 2002*. Taipei, Taiwan.

Schulte im Walde, Sabine, 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV. Las Palmas de Gran Canaria, Spain.

Vossen, P., 2001. Extending, Trimming and Fusing WordNet for technical Documents. In *Proc. of NAACL 2001 workshop on WordNet and Other Lexical Resources*. Pittsbourgh.

Wauschkuhn, O., 1999. *Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora*. Ph.D. thesis, Institut für Informatik, Universität Stuttgart.

[6] *zu-infinitive* complement