Evaluation Corpora for Sense Disambiguation in the Medical Domain

Diana Raileanu[§], Paul Buitelaar[§], Spela Vintar[§], Jörg Bay*

[§] DFKI GmbH
Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany
{raileanu, paulb, vintar}@dfki.de

* Zinfo, University of Frankfurt
60590 Frankfurt am Main, Germany
jbay@add.uni-frankfurt.de

Abstract

An important aspect of word sense disambiguation is the evaluation of different methods and parameters. Unfortunately, there is a lack of test sets for evaluation, specifically for languages other than English and even more so for specific domains like medicine. Given that our work focuses on English as well as German text in the medical domain, we had to develop our own evaluation corpora in order to test our disambiguation methods. In this paper we describe the work on developing these corpora, using GermaNet and UMLS as (lexical) semantic resources, next to a description of the annotation tool KiC that we developed for support of the annotation task.

1. Introduction

The wider context of the work described here is the EU/NSF funded project MUCHMORE1 on the development of technologies for concept-based cross-lingual information retrieval, applied to medical information management. One of the research areas that we are focusing on in this project is word sense disambiguation (WSD), which is an important enabling task in concept-based, cross-lingual information access.

Our efforts concentrate on WSD on two levels, a medical and a general one, for the purpose of which we use two different semantic resources. The general one, EuroWordNet (Vossen, 1997), is a multilingual database with WordNets for a large number of European languages. The medical semantic resource we use is UMLS2 (Unified Medical Language System), which also contains information in many languages. However, for the purposes of the MUCHMORE project, we only use the German and English parts from both EuroWordNet and UMLS.

An important aspect of word sense disambiguation is the evaluation of different methods and parameters. Unfortunately, there is a lack of test sets for evaluation, specifically for languages other than English and even more so for specific domains like medicine. Given that our work focuses on English as well as German text in the medical domain, we had to develop our own evaluation corpora in order to test our disambiguation methods.

We decided to construct a set of lexical sample corpora3 to test our WSD methods with EuroWordNet (or rather GermaNet) for German, and with UMLS for both German and English. Lexical samples are taken from a corpus of medical scientific abstracts that has been constructed also within the MUCHMORE project (Vintar et al. 2002).

Given that the size of the German part in EuroWordNet is rather small, we decided to use a more recent, larger version of GermaNet instead. GermaNet is a lexical semantic resource for German (Hamp and Feldweg, 1997) with a structure similar to that of WordNet (Miller, 1995) and EuroWordNet. In parallel we started to develop two evaluation corpora for UMLS[4] (English and German). The lexical sample corpus for GermaNet is finished, while the UMLS corpora are not yet fully annotated, but this will be finished also soon.

The paper describes our work in constructing these evaluation corpora. The next section gives some more detail on the semantic resources used in annotation, followed by a section on the annotation tool KiC that we developed for support of the annotation task. The final selection gives an overview of the medical corpus used, the selection of ambiguous terms, our annotation guidelines and the resulting inter-annotator agreement.

2. Semantic Resources

2.1. GermaNet

GermaNet is a broad-coverage semantic lexicon for German which currently contains some 16.000 words and which models the German base vocabulary. Obviously, particular domains, like the medical domain, are represented only sparsely in this resource. Nevertheless, in order to compare domain-specific and general language use, we were interested to use this resource in tagging medical text. In this way, we are hoping to gain more

[1] http://muchmore.dfki.de
[2] http://umls.nlm.nih.gov
[3] See (Kilgarriff, 1997) for a discussion of lexical sample corpora for the evaluation of sense disambiguation.

[4] Parallel to our work, a WSD evaluation corpus has been constructed on the basis of MEDLINE and UMLS (Weeber et. al 2001). The corpora we describe here is complementary to this, with an emphasis on both English and German, on general vs. medical language use, and on the distinction between different ambiguity classes.

insight in the distribution of domain specific senses vs. more general ones. For example, the German noun *Gewebe* has two senses in GermaNet, of which only the first one applies to the medical domain.

#1 [Gewebe, Körpergewebe] (tissue, body tissue)

#2 [Gewebe, Stoff, Textilstoff] (tissue, cloth, textile)

In GermaNet, like in other WordNets, nouns, verbs and adjectives are organized in synonym sets (synsets[5]), each representing one underlying lexical concept. Synsets are interlinked through relations like antonymy (opposite) and hyperonymy (is-a). For our purposes, we only consider noun concepts and the hyperonymy relation.

2.2. UMLS

In using a general semantic resource like GermaNet or EuroWordNet, we focus on disambiguation between general and domain specific senses. Additionally, however, we also need to disambiguate between several domain specific senses as provided by UMLS.

UMLS is a resource that defines linguistic, terminological and semantic information in the medical domain. It is organized in three parts: Specialist Lexicon, MetaThesaurus and Semantic Network. The MetaThesaurus contains concepts from more than 60 standardized medical thesauri, of which for our purposes we only use the concepts from MeSH (the Medical Subject Headings thesaurus). This decision is based on the fact that MeSH is also available in German.

The semantic information that we use in annotation is the so-called Concept Unique Identifier (CUI) a code that represents a MeSH concept in the MetaThesaurus. We consider the senses of a term to be equal to all the concepts that this term is mapped onto. A term can consist of one or more strings. For example the term *trauma* is mapped onto two MeSH concepts:

#1 C0043251 →

Injuries and Wounds: Wounds and Injuries: trauma: traumatic disorders: Traumatic injury:

#2 C0021501 →

Physical Trauma: Trauma (Physical): trauma:

CUIs in UMLS are also interlinked to each other by a number of relations. Out of these we only consider the "RB" relation (broader term), which is similar to the hyperonymy relation in WordNets.

3. Annotation Tool

To support manual annotation we developed an annotation tool for lexical semantic tagging (KiC) that allows for fast, and consistent manual tagging. KiC is based on the ANNOTATE tool that has been developed in the context of the NEGRA project on syntactic annotation (Plaehn and Brants, 2000). It is implemented in Tcl/Tk and C and uses several mysql databases to store the following information:

- General information about databases and access rights
- Content and structure of the lexical semantic resource
- Content of the medical corpus
- Lexical samples extracted from the medical corpus and their corresponding annotation (one database for every annotator)

Upon starting KiC, the annotator selects a particular corpus and receives a list of words/lemmas to be annotated[6]. After selecting a particular word, the annotator is displayed a list of sentences with this word in KWIC (Key Word In Context) format. At the same time, another display is opened with the senses for this particular word. By selecting one or more of these, the annotator tags every occurrence of the word with the appropriate sense(s). If the lexical semantic resource does not contain an appropriate sense for the corresponding context, the annotator can choose to annotate with *unspec* (unspecified).

To further assist the annotator in distinguishing between senses, he not only has access to the senses themselves but also to the corresponding hierarchies based on the hypernymy relation (in GermaNet) or the broader term relation (in UMLS).

A major problem we had in working with UMLS, in addition to GermaNet and other WordNets, was that KiC had been implemented with the general WordNet structure in mind. UMLS has a completely different structure, which we had to convert into the WordNet format[7].

4. Evaluation Corpora

4.1. Medical Corpus

The evaluation corpora described in this paper have been developed on the basis of a parallel corpus of English-German medical scientific abstracts obtained from the Springer Link web site[8] that has been collected in the context of the MUCHMORE project.

The corpus consists approximately of 1 million tokens for each language. Abstracts are from 41 medical journals, each of which constitutes a relatively homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.).

4.2. Selection of Ambiguous Terms[9]

4.2.1. GermaNet Terms

Selection of ambiguous GermaNet terms to be included in the evaluation corpus proceeds in several steps. First, we calculated relevance values regarding the medical domain for all GermaNet synsets occurring in the medical corpus. These values were determined by an

[5] Every synset is associated with a unique number (offset), which we use in general processing instead of the sysnset itself.

[6] GermaNet is lemma-based whereas MeSH considers only full forms.

[7] The conversion is not systematic in any way, and only meant for the particular purpose described in this paper.

[8] http://link.springer.de/

9 Terms correspond to single nouns in both GermaNet and UMLS, as mostly only *single* word terms are ambiguous.

automatic tf.idf-based procedure that compares relative word frequency between several domains (Buitelaar and Sacaleanu, 2001). Given these relevances, we compiled a list of terms with high relevance, at least 100 occurrences in the medical corpus and with more than one synset in GermaNet. This produced a list of 40 terms, for each of which we then automatically extracted 100 occurrences at random. The following table gives an overview of the level of ambiguity (number of senses) of these selected terms:

Number of Senses	Number of Terms
2	12
3	13
4	9
5	3
6	3

During manual annotation it turned out that several of the selected terms as well as a number of the extracted occurrences were not a good choice for the medical domain and had to be discarded. This finally resulted in a set of 25 relevant terms, which can be effectively used in WSD evaluation in the medical domain.

4.2.2. UMLS Terms

The process of selecting ambiguous UMLS terms was slightly different from that of GermaNet. First of all, a computation of relevance values was not needed, because we may assume that UMLS terms will in general be relevant for the medical domain.

Further, because in the MUCHMORE project we developed an extensive format for linguistic and semantic annotation (Vintar et. al, 2002) that includes also annotation with UMLS concepts, we could automatically generate lists of all ambiguous UMLS terms (English and German) along with their frequencies. Using these lists we selected a set of 59 frequent terms for English (with frequencies over 100). For German, we could only select 28 terms (with frequencies over 15[10]), as the German part of UMLS (or rather MeSH) is rather small. The level of ambiguity for these UMLS terms is mostly limited to only 2 senses.

4.3. Annotation Guidelines

The information that annotators have access to in GermaNet based annotation is: GermaNet senses with corresponding hierarchies, the context of the occurrence[11], and where available the synset definitions ("glosses"). In difficult cases, available additional information could be consulted in GermaNet directly. Annotators were allowed to annotate with more than one sense[12] if they could not decide between the senses. If none of the senses was appropriate in the particular context, they had to tag the occurrence with the label *unspec* (unspecified). Neither one of the two annotators is a medical expert, but because most of the terms expressed still rather commonly known (medical or general) concepts they were able to do the annotation task without much difficulty.

In the case of UMLS, medical experts are involved in the manual annotation. Here, annotators have access to information on variants (including synonyms) of the ambiguous term as available in UMLS and on the next higher concept ("supertype") in the corresponding concept hierarchy. Only one higher level is shown, as the complete hierarchies can reach considerable size without bringing any real benefit. Where available, the annotator can also see the definition for a concept.

4.4. Inter-Annotator Agreement

In order to check the agreement between annotators and also to check to reliability of their judgments, we calculated inter-annotator agreement scores based on the kappa statistic as described in (Siegel & Castellan, 1988; Carletta, 1996)[13]. Out of a total of 2421 occurrences for the following 25 terms, 318 were annotated differently between the two annotators.

Total	Ambiguous GermaNet Term	Agr.	κ
100	Band (tape, strap)	100%	1.00
100	Fall (drop, case, instance)	100%	1.00
100	Gefäss (jar, vessel)	100%	1.00
100	Operation (operation, surgery)	100%	1.00
100	Prüfung (survey, tryout, checkup)	100%	1.00
100	Verletzung (injury, trauma)	100%	1.00
99	Wahl (ballot, choice, option)	100%	1.00
100	Lage (site, status, position, layer)	97%	0.95
83	Gewicht (weight, importance)	96.38%	0.86
100	Sicht (sight, prospect)	94%	0.80
98	Programm (routine, manifesto)	90.81%	0.76
100	Ausfall (outage, loss, failure)	98%	0.74
92	Untersuchung (probe, inquiry)	84.78%	0.69
96	Gebiet (zone, region, field, area)	82.22%	0.65
97	Leistung (service, power, activity)	77.31%	0.64
100	Form (shape, mold, mode, form)	97%	0.55
100	Anlage (predisposition, system)	88%	0.52
81	Bewegung (motion, flow, stir)	74.07%	0.51
100	Stand (status, profession, estate)	84%	0.47
83	Infektion (infection)	78.31%	0.43
100	Übertragung (transmission, transfer)	66%	0.39
100	System (system, scheme, regime)	58%	0.38
95	Raum (space, room, range, cavity)	52.63%	0.36
99	Verbindung (contact, link, tie, bond)	73.73%	0.36
98	Praxis (practice, experience)	74.48%	0.33

Table 1: Agreement and Kappa scores

[10] We automatically created evaluation corpora using a *random selection* of occurrences if the term frequency was higher than 100, and using *all* occurrences if the term frequency was lower than 100.

[11] The annotator can see the local context - the sentence where the lemma occurs - but also the extended context – one or more sentences before and after.

[12] In fact, no term occurrence was tagged with more than two senses.

[13] Because this method assumes that the annotator can only choose one sense, we ignored in this computation the 79 occurrences that were annotated with two senses. Alternatively, we could have treated each of these sense combinations as an additional sense but this would lead to an explosion of possible senses, thereby artificially lowering inter-annotator agreement significantly.

Agreement scores between annotators seem to be acceptable for most of the terms. However, the resulting kappa statistic scores are not overall satisfying. Carletta (1996) considers that $K > 0.8$ means good reliability, while K between 0.67 and 0.8 allow for tentative conclusions to be drawn. We think the bad scores can be explained by the fact that the kappa statistic algorithm does not take into consideration the difference in distribution of sense probabilities over a domain specific (in this case, a medical) corpus. The probability that all GermaNet senses for a given term are to be found in a particular medical corpus is very small. If such different distribution probabilities would be taken into account by this algorithm, we assume that the kappa scores would rather improve.

5. Conclusions

In this paper we described our work on evaluation corpora that we developed for testing the different WSD methods we explore in the MUCHMORE project on concept-based cross-lingual information retrieval. WSD methods explored in this context use information from aligned parallel corpora, collocations, domain relevance, and context models in instance–based learning.

The resources and tools we used and developed in constructing the evaluation corpora are GermaNet as a general and UMLS as a domain-specific sense inventory, several tools and methods for the semi-automatic selection of relevant ambiguous terms, the MUCHMORE bilingual medical corpus of English-German medical scientific abstracts and the manual annotation tool KiC.

At the moment of writing, the evaluation corpus for ambiguous GermaNet terms is available, whereas the annotation of UMLS terms is still underway.

Acknowledgements

We thank Thorsten Brants for general advice on various aspects of the work. We also thank Oliver Plaehn and Daniel Olejnik for designing and implementing the KiC annotation tool. Finally, we thank Sarah Schmitt and Sebastian Thull for their annotation work. This research has in part been supported by EC/NSF grant *IST-1999-11438* for the MUCHMORE project.

References

Carletta, J.C. *Assessing agreement on classification tasks: the kappa statistic.* In: Computational Linguistics 22(2):249-254, 1996.

Buitelaar, P. and Sacaleanu, B. 2001. *Ranking and Selecting Synsets by Domain Relevance.* In: Proceedings NAACL WordNet Workshop.

Hamp, B. and Feldweg, H. 1997. *GermaNet: a Lexical-Semantic Net for German.* In: Proceedings of the ACL/EACL97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid.

Kilgarriff, A. 1997. *Sample the lexicon.* Technical report ITRI-97-01, University of Brighton.

Miller, G.A. 1995. *WordNet: A Lexical Database for English.* Communications of the ACM 11.

Plaehn P. and Brants Th. 2000. *Annotate -- An Efficient Interactive Annotation Tool* In: Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP, Seattle, WA.

Siegel, S. and N.J. Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, second edition, 1988.

Vintar, S., P. Buitelaar, B. Ripplinger, B. Sacaleanu, D. Raileanu & D. Prescher. *An Efficient and Flexible Format for Linguistic and Semantic Annotation.* In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), May 29-31, Las Palmas, Canary Islands, Spain.

Vossen, P. 1997. *EuroWordNet: a multilingual database for information retrieval.* In: Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.

Weeber M. Mork J. and Aronson A. 2001. *Developing a Test Collection for Biomedical Word Sense Disambiguation.* In: Proceedings AMIA.