Unsupervised Monolingual and Bilingual Word-Sense Disambiguation of Medical Documents using UMLS

Dominic Widdows, Stanley Peters, Scott Cederberg, Chiu-Ki Chan
Stanford University, California
{dwiddows,peters,cederber,ckchan}@csli.stanford.edu

Diana Steffen
Consultants for Language Technology, Saarbrücken, Germany
steffen@clt-st.de

Paul Buitelaar
DFKI, Saarbrücken, Germany
paulb@dfki.de

Abstract

This paper describes techniques for unsupervised word sense disambiguation of English and German medical documents using UMLS. We present both monolingual techniques which rely only on the structure of UMLS, and bilingual techniques which also rely on the availability of parallel corpora. The best results are obtained using relations between terms given by UMLS, a method which achieves 74% precision, 66% coverage for English and 79% precision, 73% coverage for German on evaluation corpora and over 83% coverage over the whole corpus. The success of this technique for German shows that a lexical resource giving relations between concepts used to index an English document collection can be used for high quality disambiguation in another language.

1 Introduction

This paper reports on experiments in monolingual and multilingual word sense disambiguation (WSD) in the medical domain using the Unified Medical Language System (UMLS). The work described was carried out as part of the MUCHMORE project [1] for multilingual organisation and retrieval of medical information, for which WSD is particularly important.

The importance of WSD to multilingual applications stems from the simple fact that meanings represented by a single word in one language may be represented by multiple words in other languages. The English word *drug* when referring to medically therapeutic drugs would be translated as *medikamente*, while it would be rendered as *drogen* when referring to a recreationally taken narcotic substance of the kind that many governments prohibit by law.

The ability to disambiguate is therefore essential to the task of machine translation — when translating from English to Spanish or from English to German we would need to make the distinctions mentioned above and other similar ones. Even short of the task of full translation, WSD is crucial to applications such as cross-lingual information retrieval (CLIR), since search terms entered in the language used for querying must be appropriately rendered in the language used for retrieval. WSD has become a well-established subfield of natural language processing with its own evaluation standards and SENSEVAL competitions (Kilgarriff and Rosenzweig, 2000).

Methods for WSD can effectively be divided into those that require manually annotated training data (supervised methods) and those that do not (unsupervised methods) (Ide and Véronis, 1998). In general, supervised methods are less scalable than unsupervised methods because they rely on training data which may be costly and unrealistic to produce, and even then might be available for only a few ambiguous terms. The goal of our work on disambiguation in the MUCHMORE project is to enable the correct semantic annotation of entire document collections with all terms which are potentially relevant for organisation, retrieval and summarisation of information. Therefore a decision was taken early on in the project that we should focus on unsupervised methods, which have the potential to be scaled up enough to meet our needs.

This paper is arranged as follows. In Section 2 we describe the lexical resource (UMLS) and the corpora we used for our experiments. We then describe and evaluate three different methods for disambiguation. The bilingual method (Section 3) takes ad-

[1] http://muchmore.dfki.de

vantage of our having a translated corpus, because knowing the translation of an ambiguous word can be enough to determine its sense. The collocational method (Section 4) uses the occurence of a term in a recognised fixed expression to determine its meaning. UMLS relation based methods (Section 5) use relations between terms in UMLS to determine which sense is being used in a particular instance. Other techniques used in the MUCHMORE project include domain-specific sense selection (Buitelaar and Sacaleanu, 2001), used to select senses appropriate to the medical domain from a general lexical resource, and instance-based learning, a machine-learning technique that has been adapted for word-sense disambiguation (Widdows et al., 2003).

2 Language resources used in these experiments

2.1 Lexical Resource — UMLS

The Unified Medical Language System (UMLS) is a resource that contains linguistic, terminological and semantic information in the medical domain.[2] It is organised in three parts: Specialist Lexicon, MetaThesaurus and Semantic Network. The MetaThesaurus contains concepts from more than 60 standardised 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 concept in the UMLS MetaThesaurus. We consider the possible 'senses' of a term to be the set of CUI's which list this term as a possible realisation. For example, UMLS contains the term *trauma* as a possible realisation of the following two concepts:

> C0043251 Injuries and Wounds: Wounds and Injuries: trauma: traumatic disorders: Traumatic injury:
>
> C0021501 Physical Trauma: Trauma (Physical): trauma:

Each of these CUI's is a possible sense of the term *trauma*. The term *trauma* is therefore noted as ambiguous, since it can be used to express more than one UMLS concept. The purpose of disambiguation is to find out which of these possible senses is actually being used in each particular context where there term *trauma* is used.

[2] UMLS is freely available under license from the United States National Library of Medicine, http://www.nlm.nih.gov/research/umls/

CUI's in UMLS are also interlinked to each other by a number of relations. These include:

- 'Broader term' which is similar to the hypernymy relation in WordNet (Fellbaum, 1998). In general, x is a 'broader term' for y if every y is also a (kind of) x.
- More generally, 'related terms' are listed, where possible relationships include 'is like', 'is clinically associated with'.
- Cooccurring concepts, which are pairs of concepts which are linked in some information source. In particular, two concepts are regarded as cooccurring if they have both been used to manually index the same document in MEDLINE. We will refer to such pairs of concepts as *coindexing* concepts.
- Collocations and multiword expressions. For example, the term *liver transplant* is included separately in UMLS, as well as both the terms *liver* and *transplant*. This information can sometimes be used for disambiguation.

2.2 The Springer Corpus of Medical Abstracts

The experiments and implementations of WSD described in this paper were all carried out on a parallel corpus of English-German medical scientific abstracts obtained from the Springer Link web site.[3] 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.). The corpus was automatically marked up with morphosyntactic and semantic information, as described by Špela Vintar et al. (2002). In brief, whenever a token is encountered in the corpus that is listed as a term in UMLS, the document is annotated with the CUI under which that term is listed. Ambiguity is introduced by this markup process because the lexical resources often list a particular term as a possible realisation of more than one concept or CUI, as with the *trauma* example above, in which case the document is annotated with all of these possible CUI's.

The number of tokens of UMLS terms included by this annotation process is given in Table 1. The table shows how many tokens were found by the annotation process, listed according to how many possible senses each of these tokens was assigned in UMLS (so that the number of ambiguous tokens is the number

[3] `http://link.springer.de/`

Number of Senses	1	2	3	4
Before Disambiguation				
English	223441	31940	3079	56
German	124369	7996	0	0
After Disambiguation				
English	252668	5299	568	5
German	131302	1065	0	0

Table 1: The number of tokens of terms that have 1, 2, 3 and 4 possible senses in the Springer corpus

of tokens with more than one possible sense). The greater number of concepts found in the English corpus reflects the fact that UMLS has greater coverage for English than for German, and secondly that there are many small terms in English which are expressed by single words which would be expressed by larger compound terms in German (for example *knee* + *joint* = *kniegelenk*). Table 1 also shows how many tokens of UMLS concepts were in the annotated corpus *after* we applied the disambiguation process described in Section 5, which proved to be our most successful method. As can be seen, our disambiguation methods resolved some 83% of the ambiguities in the English corpus and 87% of the ambiguities in the German corpus (we refer to this proportion as the 'Coverage' of the method). However, this only measures the number of disambiguation decisions that were made: in order to determine how many of these decisions were correct, evaluation corpora were needed.

2.3 Evaluation Corpora

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 German as well as English text in the medical domain, we had to develop our own evaluation corpora in order to test our disambiguation methods.

Because in the MUCHMORE project we developed an extensive format for linguistic and semantic annotation (Špela Vintar et al., 2002) that includes annotation with UMLS concepts, we could automatically generate lists of all ambiguous UMLS types (English and German) along with their token frequencies in the corpus. Using these lists we selected a set of 70 frequent types for English (token frequencies at least 28, 41 types having token frequencies over 100). For German, we only selected 24 ambiguous types (token frequencies at least 11, 7 types having token frequencies over 100) because there are fewer ambiguous terms in the German annotation (see Table 1). We automatically selected instances to be annotated using a random selection of occurrences if the token frequency was higher than 100, and using all occurrences if the token frequency was lower than 100. The level of ambiguity for these UMLS terms is mostly limited to only 2 senses; only 7 English terms have 3 senses.

Correct senses of the English tokens in context were chosen by three medical experts, two native speakers of German and one of English. The German evaluation corpus was annotated by the two German speakers. Interannotator agreement for individual terms ranged from very low to very high, with an average of 65% for German and 51% for English (where all three annotators agreed). The reasons for this low score are still under investigation. In some cases, the UMLS definitions were insufficient to give a clear distinction between concepts, especially when the concepts came from different original thesauri. This allowed the decision of whether a particular definition gave a meaningful 'sense' to be more or less subjective. Approximately half of the disagreements between annotators occured with terms where interannotator agreement was less than 10%, which is evidence that a significant amount of the disagreement between annotators was on the *type* level rather than the *token* level. In other cases, it is possible that there was insufficient contextual information provided for annotators to agree. If one of the annotators was unable to choose any of the senses and declared an instance to be 'unspecified', this also counted against interannotator agreement. Whatever is responsible, our interannotator agreement fell far short of the 88%-100% achieved in SENSEVAL (Kilgarriff and Rosenzweig, 2000, §7), and until this problem is solved or better datasets are found, this poor agreement casts doubt on the generality of the results obtained in this paper.

A 'gold standard' was produced for the German UMLS evaluation corpus and used to evaluate the disambiguation of German UMLS concepts. The English experiments were evaluated on those tokens for which the annotators agreed. More details and discussion of the annotation process is available in the project report (Widdows et al., 2003).

In the rest of this paper we describe the techniques that used these resources to build systems for word sense disambiguation, and evaluate their level of success.

3 Bilingual Disambiguation

The mapping between word-forms and senses differs across languages, and for this reason the importance of word-sense disambiguation has long been recognised for machine translation. By the same token, pairs of translated documents naturally contain information for disambiguation. For example, if in a particular context the English word *drugs* is translated into French as *drogues* rather than *medicaments*, then the English word *drug* is being used to mean *narcotics* rather than *medicines*. This observation has been used for some years on varying scales. Brown et al. (1991) pioneered the use of statistical WSD for translation, building a translation model from one million sentences in English and French. Using this model to help with translation decisions (such as whether *prendre* should be translated as *take* or *make*), the number of acceptable translations produced by their system increased by 8%. Gale et al. (1992) use parallel translations to obtain training and testing data for word-sense disambiguation. Ide (1999) investigates the information made available by a translation of George Orwell's *Nineteen Eighty-four* into six languages, using this to analyse the related senses of nine ambiguous English words into hierarchical clusters.

These applications have all been case studies of a handful of particularly interesting words. The large scale of the semantic annotation carried out by the MUCHMORE project has made it possible to extend the bilingual disambiguation technique to entire dictionaries and corpora.

To disambiguate an instance of an ambiguous term, we consulted the translation of the abstract in which it appeared. We regarded the translated abstract as disambiguating the ambiguous term if it met the following two criteria:

- Only one of the CUI's was assigned to any term in the translated abstract.
- At least one of the terms to which this CUI was assigned in the translated abstract was unambiguous (i.e. was not also assigned another CUI).

3.1 Results for Bilingual Disambiguation

We attempted both to disambiguate terms in the German abstracts using the corresponding English abstracts, and to disambiguate terms in the English abstracts using the corresponding German ones. In this collection of documents, we were able to disambiguate 1802 occurrences of 63 English terms and 1500 occurrences of 43 German terms. Comparing this with the evaluation corpora gave the results in Table 2.[4]

	Precision	Recall	Coverage
English	81%	18%	22%
German	66%	22%	33%

Table 2: Results for bilingual disambiguation

As can be seen, the recall and coverage of this method is not especially good but the precision (at least for English) is very high. The German results contain roughly the same proportion of correct decisions as the English, but many more incorrect ones as well.

Our disambiguation results break down into three cases:

1. Terms ambiguous in one language that translate as multiple unambiguous terms in the other language; one of the meanings is medical and the other is not.
2. Terms ambiguous in one language that translate as multiple unambiguous terms in the other language; both of the terms are medical.
3. Terms that are ambiguous between two meanings that are difficult to distinguish.

One striking aspect of the results was that relatively few terms were disambiguated to different senses in different occurrences. This phenomenon was particularly extreme in disambiguating the German terms; of the 43 German terms disambiguated, 42 were assigned the same sense every time we were able to disambiguate them. Only one term, *Metastase*, was assigned difference senses; 88 times it was assigned CUI C0027627 ("The spread of cancer from one part of the body to another ...", associated with the English term *Metastasis* and 6 times it was assigned CUI C0036525 "Used with neoplasms to indicate the secondary location to which the neoplastic process has metastasized", corresponding to the English terms *metastastic* and *secondary*). Metastase therefore falls into category 2 from above, although the distinction between the two meanings is relatively subtle.

The first and third categories above account for the vast majority of cases, in which only one meaning is ever selected. It is easy to see why this would

[4]In all of the results presented in this paper, Precision is the proportion of decisions made which were correct according to the evaluation corpora, Recall is the proportion of instances in the evaluation corpora for which a correct decision was made, and Coverage is the proportion of instances in the evaluation corpora for which any decision was made. It follows that

$$\text{Recall} = \text{Precision} \times \text{Coverage}.$$

happen in the first category, and it is what we want to happen. For instance, the German term *Krebse* can refer either to *crabs* (Crustaceans) or to *cancerous growths*; it is not surprising that only the latter meaning turns up in the corpus under consideration and that we can determine this from the unambiguous English translation *cancers*.

In English somewhat more terms were disambiguated multiple ways: eight terms were assigned two different senses across their occurrences. All three types of ambiguity were apparent. For instance, the second type (medical/medical ambiguity) appeared for the term *Aging*, which can refer either to aging people (*Alte Menschen*) or to the process of aging itself (*Altern*); both meanings appeared in our corpus.

In general, the bilingual method correctly find the meanings of approximately one fifth of the ambiguous terms, and makes only a few mistakes for English but many more for German.

4 Collocational disambiguation

By a 'collocation' we mean a fixed expression formed by a group of words occuring together, such as *blood vessel* or *New York*. (For the purposes of this paper we only consider contiguous multiword expressions which are listed in UMLS.) There is a strong and well-known tendency for words to express only one sense in a given collocation. This property of words was first described and quantified by Yarowsky (1993), and has become known generally as the 'One Sense Per Collocation' property. Yarowsky (1995) used the one sense per collocation property as an essential ingredient for an unsupervised Word-Sense Disambiguation algorithm. For example, the collocations *plant life* and *manufacturing plant* are used as 'seed-examples' for the *living thing* and *building* senses of *plant*, and these examples can then be used as high-precision training data to perform more general high-recall disambiguation.

While Yarowsky's algorithm is unsupervised (the algorithm does not need a large collection of annotated training examples), it still needs direct human intervention to recognise which ambiguous terms are amenable to this technique, and to choose appropriate 'seed-collocations' for each sense. Thus the algorithm still requires expert human judgments, which leads to a bottleneck when trying to scale such methods to provide Word-Sense Disambiguation for a whole document collection.

A possible method for widening this bottleneck is to use existing lexical resources to provide seed collocations. The texts of dictionary definitions have been used as a traditional source of information for disambiguation (Lesk, 1986). The richly detailed structure of UMLS provides a special opportunity to combine both of these approaches, because many multiword expressions and collocations are included in UMLS as separate concepts.

For example, the term *pressure* has the following three senses in UMLS, each of which is assigned to a different semantic type (TUI):

Sense of *pressure*	Semantic Type
Physical pressure (C0033095)	Quantitative Concept
Pressure - action (C0460139)	Therapeutic or Preventive Procedure
Baresthesia, sensation of pressure (C0234222)	Organ or Tissue Function

Many other collocations and compounds which include the word *pressure* are also of these semantic types, as summarised in the following table:

Quantitative Concept	mean pressure, bar pressure, population pressure
Therapeutic Procedure	orthostatic pressure, acupressure
Organ or Tissue Function	arterial pressure, lung pressure, intraocular pressure

This leads to the hypothesis that the term *pressure*, when used in any of these collocations, is used with the meaning corresponding to the same semantic type. This allows deductions of the following form:

Collocation	bar pressure, mean pressure
Semantic type	Quantitative Concept
Sense of *pressure*	C0033095, physical pressure

Since nearly all English and German multiword technical medical terms are head-final, it follows that the a multiword term is usually of the same semantic type as its head, the final word. (So for example, *lung cancer* is a kind of cancer, not a kind of lung.) For English, UMLS 2001 contains over 800,000 multiword expressions the last word in which is also a term in UMLS. Over 350,000 of these expressions have a last word which on its own, with no other context, would be regarded as ambiguous (has more that one CUI in UMLS). Over 50,000 of *these* multiword expressions are unambiguous, with a unique semantic type which is shared by only one of the meanings of the potentially ambiguous final word. The ambiguity of the final word in such multiword expressions is thus resolved, providing over 50,000 'seed collocations' for use in semantically annotating documents with disambiguated word senses.

4.1 Results for collocational disambiguation

Unfortunately, results for collocational disambiguation (Table 3) were disappointing compared with the promising number of seed collocations we expected to find. Precision was high, but comparatively few of the collocations suggested by UMLS were found in the Springer corpus.

	Precision	Recall	Coverage
English	79%	3%	4%
German	82%	1%	1.2%

Table 3: Results for collocational disambiguation

In retrospect, this may not be surprising given that many of the 'collocations' in UMLS are rather collections of words such as

C0374270 intracoronary percutaneous placement s single stent transcatheter vessel

which would almost never occur in natural text. Thus very few of the potential collocations we extracted from UMLS actually occurred in the Springer corpus. This scarcity was especially pronounced for German, because so many terms which are several words in English are compounded into a single word in German. For example, the term

C0035330 retinal vessel

does occur in the (English) Springer corpus and contains the ambiguous word *vessel*, whose ambiguity is successfully resolved using the collocational method. However, in German this concept is represented by the single word

C0035330 Retinagefaesse

and so this ambiguity never arises in the first place.

It should still be remarked that the few decisions that were made by the collocational method were very accurate, demonstrating that we can get some high precision results using this method. It is possible that recall could be improved by relaxing the conditions which a multiword expression in UMLS must satisfy to be used as a seed-collocation.

5 Disambiguation using related UMLS terms found in the same context

While the collocational method turned out to give disappointing recall, it showed that accurate information could be extracted directly from the existing UMLS and used for disambiguation, without extra human intervention or supervision. What we needed was advice on how to get more of this high-quality information out of UMLS, which we still believed to be a very rich source of information which we were not yet exploiting fully. Fortunately, no less than 3 additional sources of information for disambiguation using related terms from UMLS were suggested by a medical expert.[5] The suggestion was that we should consider terms that were linked by conceptual relations (as given by the MRREL and MRCXT files in the UMLS source) and which were noted as coindexing concepts in the same MEDLINE abstract (as given by the MRCOC file in the UMLS source). For each separate sense of an ambiguous word, this would give a set of related concepts, and if examples of any of these related concepts were found in the corpus near to one of the ambiguous words, it might indicate that the correct sense of the ambiguous word was the one related to this particular concept.

This method is effectively one of the many variants of Lesk's (1986) original dictionary-based method for disambiguation, where the words appearing in the definitions of different senses of ambiguous words are used to indicate that those senses are being used if they are observed near the ambiguous word. However, we gain over purely dictionary-based methods because the words that occur in dictionary definitions rarely correspond well with those that occur in text. The information we collected from UMLS did not suffer from this drawback: the pairs of coindexing concepts from MRCOC were derived *precisely* from human judgements that these two concepts both occured in the same text in MEDLINE.

The disambiguation method proceeds as follows. For each ambiguous word w, we find its possible senses $\{s_j(w)\}$. For each sense s_j, find all CUI's in MRREL, MRCXT or MRCOC files that are related to this sense, and call this set $\{c_{\mathrm{rel}}(s_j)\}$. Then for each occurrence of the ambiguous word w in the corpus we examine the local context to see if a term t occurs whose sense[6] (CUI) is one of the concepts in $\{c_{\mathrm{rel}}(s_j)\}$, and if so take this as positive evidence that the sense s_j is the appropriate one for this context, by increasing the score of s_j by 1. In this way, each sense s_j in context gets assigned a score which measures the number of terms in this context which are related to this sense. Finally, choose the sense

[5] Personal communication from Stuart Nelson (instrumental in the design of UMLS), at the MUCHMORE workshop in Croatia, September 2002.

[6] This fails to take into account that the term t might itself be ambiguous — it is possible that results could be improved still further by allowing for *mutual* disambiguation of more than one term at once.

with the highest score.

One open question for this algorithm is what region of text to use as a context-window. We experimented with using sentences, documents and whole subdomains, where a 'subdomain' was considered to be all of the abstracts appearing in one of the journals in the Springer corpus, such as *Arthroskopie* or *Der Chirurg*. Thus our results (for each language) vary according to which knowledge sources were used (Conceptually Related Terms from MRREL and MRCXT or coindexing terms from MRCOC, or a combination), and according to whether the context-window for recording cooccurence was a sentence, a document or a subdomain.

5.1 Results for disambiguation based on related UMLS concepts

The results obtained using this method (Tables 5.1 and 5.1) were excellent, preserving (and in some cases improving) the high precision of the bilingual and collocational methods while greatly extending coverage and recall. The results obtained by using the coindexing terms for disambiguation were particularly impressive, which coincides with a long-held view in the field that terms which are topically related to a target word can be much richer clues for disambiguation that terms which are (say) hierarchically related. We are very fortunate to have such a wealth of information about the cooccurence of pairs of concepts through UMLS, which appears to have provided the benefits of cooccurence data from a manually annotated training sample without having to perform the costly manual annotation.

In particular, for English (Table 5.1), results were actually better using only coindexing terms rather than combining this information with hierarchically related terms: both precision and recall are best when using only the MRCOC knowledge source. As we had expected, recall and coverage increased but precision decreased slightly when using larger contexts.

The German results (Table 5.1) were slightly different, and even more successful, with nearly 60% of the evaluation corpus being correctly disambiguated, nearly 80% of the decisions being correct. Here, there was some small gain when combining the knowledge sources, though the results using only coindexing terms were almost as good. For the German experiments, using larger contexts resulted in greater recall *and* greater precision. This was unexpected — one hypothesis is that the sparser coverage of the German UMLS contributed to less predictable results on the sentence level.

These results are comparable with some of the better SENSEVAL results (Kilgarriff and Rosenzweig, 2000) which used fully supervised methods, though the comparison may not be accurate because we are choosing between fewer senses than on avarage in SENSEVAL, and because of the doubts over our interannotator agreement.

Comparing these results with the number of words disambiguated in the whole corpus (Table 1), it is apparent that the average coverage of this method is actually *higher* for the whole corpus (over 80%) than for the words in the evaluation corpus. It is possible that this reflects the fact the the evaluation corpus was specifically chosen to include words with 'interesting' ambiguities, which might include words which are more difficult than average to disambiguate. It is possible that over the whole corpus, the method actually works *even better* than on just the evaluation corpus.

This technique is quite groundbreaking, because it shows that a lexical resource derived almost entirely from English data (MEDLINE indexing terms) could successfully be used for automatic disambiguation in a German corpus. (The alignment of documents and their translations was not even considered for these experiments so the results do not depend at all on our having access to a parallel corpus.) This is because the UMLS relations are defined between concepts rather than between words. Thus if we know that there is a relationship between two concepts, we can use that relationship for disambiguation, even if the original evidence for this relationship was derived from information in a different language from the language of the document we are seeking to disambiguate. We are assigning the correct senses based not upon how terms are related in language, but how *medical concepts* are related to one another.

It follows that this technique for disambiguation should be applicable to any language which UMLS covers, and applicable at very little cost. This proposal should stimulate further research, and not too far behind, successful practical implementation.

6 Summary and Conclusion

We have described three implementations of unsupervised word-sense disambiguation techniques for medical documents. The bilingual method relies on the availability of a translated parallel corpus: the collocational and relational methods rely solely on the structure of UMLS, and could therefore be applied to new collections of medical documents without requiring any new resources. The method of disambiguation using relations between terms given by UMLS was by far the most successful method, achieving 74% precision, 66% coverage for English

ENGLISH RESULTS	Related terms (MRREL)			Related terms (MRCXT)			Coindexing terms (MRCOC)			Combined (majority voting)		
	Prec.	Rec.	Cov.	Prec.	Rec.	Cov.	Prec.	Rec.	Cov.	Prec.	Rec.	Cov.
Sentence	50	14	28	60	9	15	78	32	41	74	32	43
Document	48	24	50	63	22	35	74	46	62	72	45	63
Subdomain	51	33	65	64	38	59	74	49	66	71	49	69

Table 4: Results for disambiguation based on UMLS relations (English)

GERMAN RESULTS	Related terms (MRREL)			Related terms (MRCXT)			Coindexing terms (MRCOC)			Combined (majority voting)		
	Prec.	Rec.	Cov.	Prec.	Rec.	Cov.	Prec.	Rec.	Cov.	Prec.	Rec.	Cov.
Sentence	64	24	38	75	11	15	76	29	38	77	31	40
Document	68	43	63	75	27	36	79	52	66	79	53	67
Subdomain	70	51	73	74	52	70	79	58	73	79	58	73

Table 5: Results for disambiguation based on UMLS relations (German)

and 79% precision, 73% coverage for German on the evaluation corpora, and achieving over 80% coverage overall. This result for German is particularly encouraging, because is shows that a lexical resource giving relations between concepts in one language can be used for high quality disambiguation in another language.

Acknowledgments

This research was supported in part by the Research Collaboration between the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University, and by EC/NSF grant IST-1999-11438 for the MUCHMORE project.

We would like to thank the National Library of Medicine for providing the UMLS, and in particular Stuart Nelson for his advice and guidance.

References

P. Brown, S. de la Pietra, V. de la Pietra, and R Mercer. 1991. Word sense disambiguation using statistical methods. In *ACL 29*, pages 264–270.

Paul Buitelaar and Bogdan Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proceedings of WordNet and Other Lexical Resources*, NAACL 2001 Workshop, Pittsburgh, PA, June.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge MA.

W. Gale, K. Church, and D. Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Nancy Ide and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, March.

Nancy Ide. 1999. Parallel translations and sense discriminators. In *Proceedings of the ACL SIGLEX workshop on Standardizing Lexical Resources*, pages 52–61.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34(1-2):15–48, April.

M. E. Lesk. 1986. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC conference*. ACM.

Špela Vintar, Paul Buitelaar, Bärbel Ripplinger, Bogdan Sacaleanu, Diana Raileanu, and Detlef Prescher. 2002. An efficient and flexible format for linguistic and semantic annotation. In *Third International Language Resources and Evaluation Conference*, Las Palmas, Spain.

Dominic Widdows, Diana Steffen, Scott Cederberg, Chiu-Ki Chan, Paul Buitelaar, and Bogdan Sacaleanu. 2003. Methods for word-sense disambiguation. Technical report, MUCHMORE project report.

David Yarowsky. 1993. One sense per collocation. In *ARPA Human Language Technology Workshop*, pages 266–271, Princeton, NJ.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.