Project ref. no.	*IST-1999-11438*
Project acronym	**MUCHMORE**
Project full title	Multilingual Concept Hierarchies for Medical Information Organization and Retrieval

Security (distribution level)	*Public*
Contractual date of delivery	*Month 24 (June 2002)*
Actual date of delivery	*Month 35 (May 2003)*
Deliverable number	*D5.1*
Deliverable title	*WSD Methods*
Type	*PU*
Status & version	*Final*
Number of pages	*45*
WP contributing to the deliverable	*WP5.1*
WP / Task responsible	*CSLI*
Author(s)	*Dominic Widdows, Stanley Peters, Scott Cederberg, Chiu-Ki Chan (CSLI); Diana Steffen, Paul Buitelaar, Bogdan Sacaleanu (DFKI)*
EC Project Officer	*Yves Paternoster*
Keywords	*Unsupervised Word Sense Disambiguation, UMLS, EuroWordNet, GermaNet, Evaluation Corpora, Semantic and Linguistic Annotation, Parallel corpora, Collocations, Related Terms, Domain Specific Sense, Instance Based Learning*
Abstract (for dissemination)	*The report describes the work in WP5 on Cross-Lingual Word Sense Disambiguation. Four types of methods were developed, two methods for UMLS (bilingual; collocation) and two for EuroWordNet (domain specific sense; instance-based learning), both for English and German. Evaluation corpora were produced to test these methods, which are described in terms of the annotation tool and guidelines used and the level of agreement between different annotators. Precision, recall and coverage results are described for each method.*

Introduction

The wider context of the work described here is 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.

Many words have more than one meaning, or sense. The different meanings of a word can range from being very closely related to having no apparent connection. A classic example of the latter extreme comes from the English word “bank”, which can refer either to a financial institution or to the side of a river. Another English example of sense ambiguity is the word “free”, which can either mean “gratis”, or without charge (“free beer”, a “free lunch”, a “free gift”) or can refer to freedom or liberty (“politically free”, “intellectually free”).

The task of determining which of its meanings an ambiguous word has in a particular instance is known as word-sense disambiguation, or WSD. This is typically performed by looking up the senses of a word in question in a dictionary, and computing the most likely sense.

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 meanings of the English word “free” discussed above are represented by the two Spanish words “gratis” and “libre”. 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 distinctions as mentioned above. Even short of the task of full translation, WSD may also be 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.

Because of this potential importance to cross-lingual language and information applications, WSD has been one of the areas of focus of the MUCHMORE project.

1 Language Resources used for WSD in MUCHMORE

In this section we describe the lexical resources used to give a list of possible senses for each term, and the corpus, which was marked up with senses from these resources. The task of disambiguation is then to remove inappropriate sense-labels.

1.1 The Lexical Resources

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: UMLS and EuroWordNet

1.1.1 UMLS

The Unified Medical Language System (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 UMLS MetaThesaurus. We consider the possible senses of a term to be equal to the set of concepts that this term can be mapped onto. A term can consist of one or more strings. For example, UMLS contains the term *trauma* as a possible realisation of the following two 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. These include:

- "Broader term" which is similar to the hypernymy relation in WordNet (Miller, 1997). In general, *x* is a 'broader term' for *y* if every *y* is also an *x*.
- More generally, "related terms" are listed, where possible relationships include *'is_like'*, *'is_clinically_associated_with'*.
- Co-occurring concepts, which are pairs of concepts, which occur 'together' in some information source. In particular, two concepts are regarded as co-occurring 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 to enable disambiguation.

1.1.2 EuroWordNet

EuroWordNet is a multilingual database with WordNets for a large number of European languages (Vossen, 1997). In addition to annotation with UMLS, terms in the corpus are annotated also with EuroWordNet to compare domain-specific and general language use. EuroWordNet is a multilingual database for several European languages and is structured in similar ways to the Princeton WordNet (Fellbaum, 1997). Each language specific (Euro)WordNet is linked to all of the others through the so-called Inter-Lingual-Index

(ILI), which is based on WordNet1.5. Via this index the languages are interconnected, so that it is possible to move from a word in one language to similar words in any of the other languages in the EuroWordNet database. For our current purposes we use only the German and English parts of EuroWordNet.

All information in (Euro)WordNet is centered around so-called synsets, which are sets of (near-) synonyms. The different senses of a term are therefore simply all the synsets that contain it. The goal of disambiguation is to narrow down these possibilities, ideally to a single sense. A term can be simple (*man*) or complex (*rock_and_roll*). A synset is identified by a unique identifier, called offset. Because meanings between languages cannot be exactly mapped one-to-one, there may be more than one synset within a language that is mapped on the same concept in the ILI. In order to distinguish between these, every synset was given a unique identifier (ID)[1], as shown in Table 1-1:

	Offset - ID	**Synset**
German	3824895 - 1	Fingergelenk
	3824895 - 2	Fingerknochen
	3824895 - 3	Knöchel
English	3824895	knuckle, knuckle joint, metacarpophalangeal joint

Table 1: EWN Example

1.2 The Springer Corpus

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.[2] 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 in MUCHMORE deliverable D4.1, *MUCHMORE Annotation Format.*. 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.

[1] In our case only for German, as the English synsets correspond to the ILI directly.

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

The number of tokens of UMLS terms included by this annotation process is given in Table 2. 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 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*).

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

Table 2 also shows how many tokens of UMLS concepts were in the annotated corpus *after* we applied the disambiguation process described in Section 3.3.2, 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 Evaluation Corpora

An important aspect of word sense disambiguation is the evaluation of different methods and parameters. To begin with, we define the terms 'Precision', 'Recall' and 'Coverage' which are used to measure and compare the effectiveness of different techniques. 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

$$Recall = Precision \times Coverage.$$

As described at the end of the previous section, it also makes sense to talk about the Coverage of a method over the whole corpus, since the Coverage score does not depend on whether a decision made by an automatic method for disambiguation was the same as that made by a human judge. But to compute Recall and Precision, we need evaluation

test sets where human annotators have judged that an ambiguous term in a given context has a particular meaning.

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.

We decided to construct a set of lexical sample corpora[3] 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 the Springer 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 developed two evaluation corpora for UMLS[4] (English and German).

This section describes our work in constructing these evaluation corpora. First we describe the annotation tool KiC that we developed for support of the annotation task, followed by an overview of the medical corpus used, the selection of ambiguous terms, our annotation guidelines and the resulting inter-annotator agreement.

2.1 Manual Annotation Tool

To support manual annotation we developed an annotation tool for lexical semantic tagging (KiC) that allows for fast and consistent manual tagging – see Figures 2-1 and 2-2 show screenshots from KiC applied to GermaNet, respectively UMLS (English).

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)

[3] See (Kilgarriff, 1998) 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.

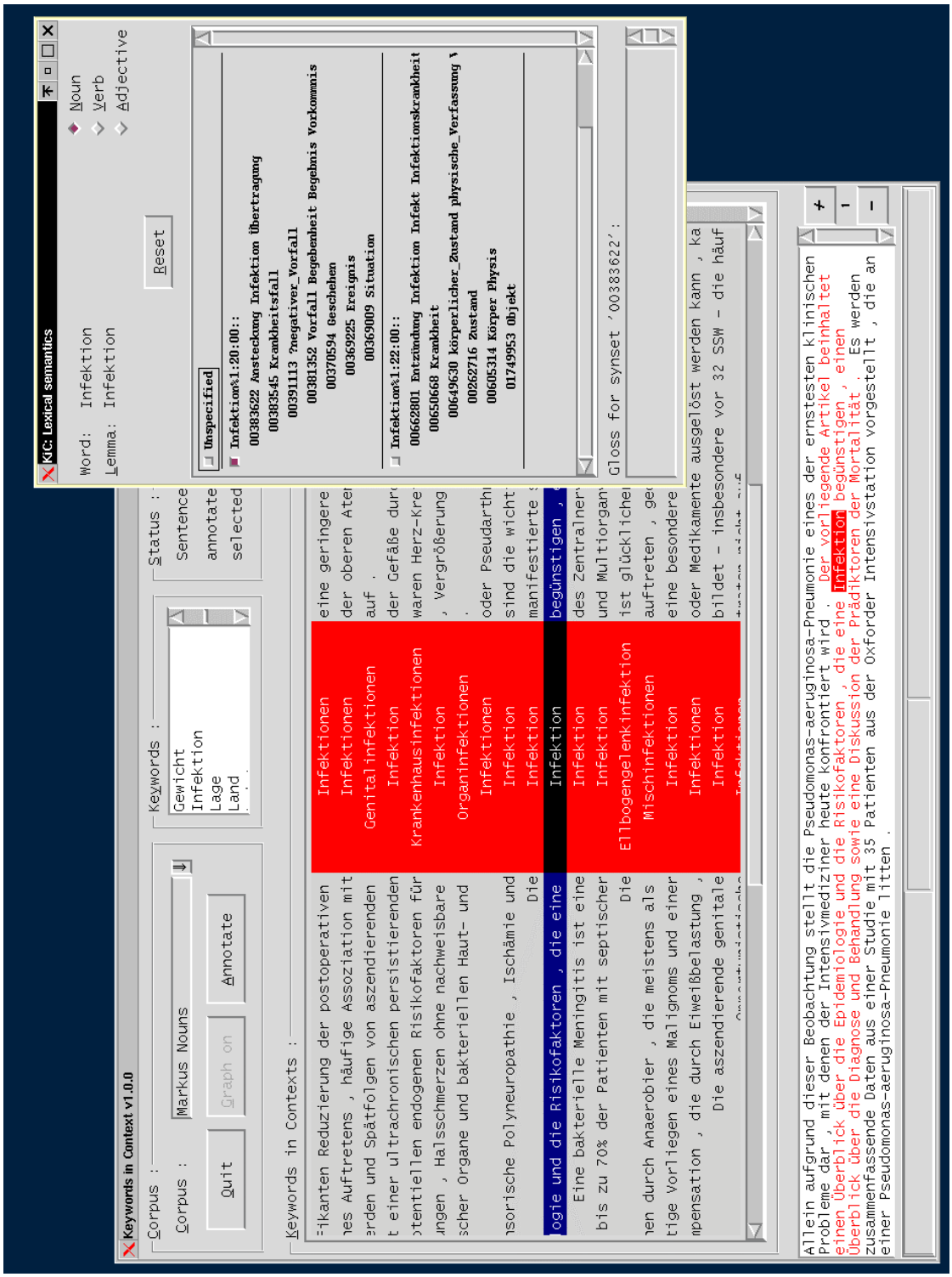

Figure 2-1: The annotation tool - GermaNet

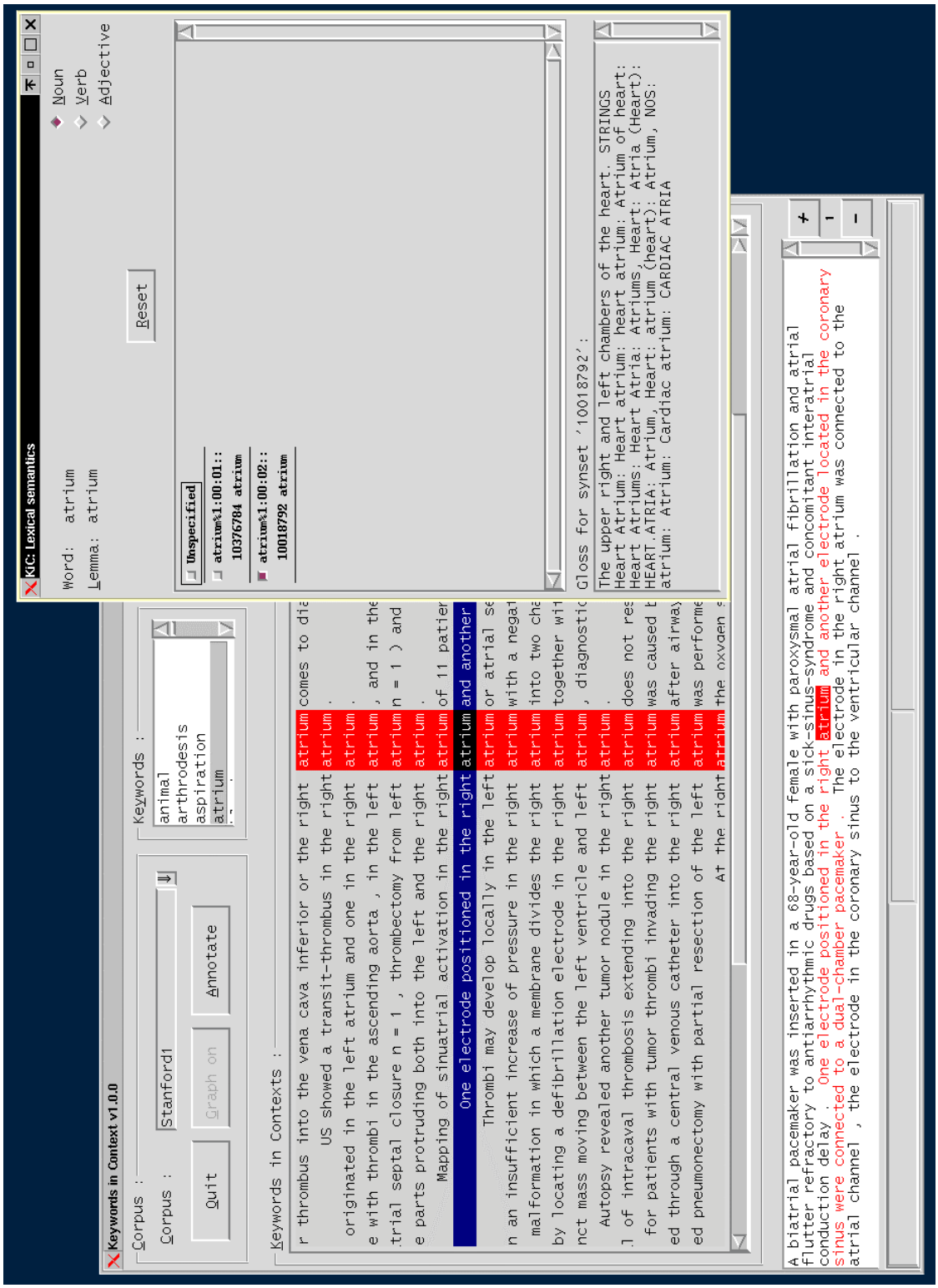

Figure 2-2: The annotation tool - UMLS (English)

Upon starting KiC, the annotator selects a particular corpus and receives a list of words/lemmas to be annotated[5]. After selecting a particular word, the annotator is displayed a list of sentences with this word in the sample of contexts in which it occurs.

Further, in selecting an occurrence, the annotator can see the extended context, that is, the left and right neighbor sentences in the medical corpus. The size of the extended context can be dynamically increased/decreased. 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[6].

2.2 Selection of Ambiguous Terms

GermaNet

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 automatic tf.idf-based procedure that compares relative word frequency between several domains (Buitelaar and Sacaleanu, 2001), which will be described in details in Section 3.3. 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. From this list we selected 40 terms, for each of which we then automatically extracted 100 occurrences at random. Table 2-1 gives an overview of the level of ambiguity (number of senses).

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

Table 3: Ambiguity Level in GermaNet Evaluation Corpus

[5] GermaNet is lemma-based whereas MeSH considers only full forms.
[6] The conversion was carried out manually and was only meant for the particular purpose described in this paper.

UMLS

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 annotation with UMLS concepts, we could automatically generate lists of all ambiguous UMLS terms (English and German) along with their token frequencies. Using these lists we selected a set of 70 frequent terms for English (token frequencies at least 28, 41 terms having token frequencies over 100). For German, we could only select 24 terms (token frequencies at least 11, 7 terms having token frequency over 100[7]), 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; only 7 English terms have 3 senses.

2.3 Annotation Guidelines

GermaNet

Three annotators, a medical expert and two linguistics students, were assigned the task of annotating the 40 ambiguous words. We also employed non-experts, as they would not have much difficulty in tagging occurrences in a medical corpus, because most of the terms express rather commonly known (medical or general) concepts. In order to tag an occurrence in the evaluation corpus they could use the information provided by KiC (see Picture 2-1):

- the small context: the sentence to which the occurrence belongs;
- the extended context: the neighbor sentences from the medical corpus;
- the GermaNet senses with their hierarchies and glosses;

In cases where annotators needed additional information to make a sense distinction, e.g. hyponyms, they could consult GermaNet directly through the standard GermaNet user interface. If none of the senses was appropriate in the particular context, they had to tag the occurrence with the label unspecified. The annotators were also allowed to annotate an occurrence with more than one sense[8] if several senses were appropriate for a particular context.

UMLS

In the case of UMLS, medical experts are involved in the manual annotation, two for the German part and three[9] for the English part. The annotators have access to information on variants (including synonyms) of the ambiguous term as available in UMLS and on the

[7] 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.
[8] In fact, no term was tagged with more than two senses.
[9] We had two German annotators and an American annotator. The German ones annotated both the German and the English UMLS evaluation corpora, while the American annotator participated in only the English UMLS evaluation corpus.

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.

The annotation task consists of choosing one (or more - see below) appropriate UMLS concept(s) for each occurrence of every word in the evaluation corpus. In order to facilitate the annotation, the annotator has access to the following information.

- The context for each occurrence:
  - the sentence in which the word to be annotated occurs
  - the context of the sentence : one sentence before and one after

- The concepts to which this word in UMLS corresponds. A concept is defined by the set of its *variants*. (A set of variants is thus similar to a synset in WordNet parlance.) For example, the word `therapy` has two concepts (C0087111 and C0039798) in UMLS, with the following variants:

```
C0087111   Therapeutic procedure
C0087111   Therapies
C0087111   therapies
C0087111   Therapy
C0087111   therapy
C0087111   TREATMENT
C0087111   Treatment
C0087111   Treatments

C0039798   therapeutic aspects
C0039798   disease management
C0039798   therapy
C0039798   treatment
```

- Some concepts have also definitions. For instance for the concept C0039798, the definition is:

```
Used with diseases for therapeutic interventions except drug
therapy, diet therapy, radiotherapy, and surgery, for which
specific subheadings exist. The concept is also used for
articles and books dealing with multiple therapies.
```

In the annotation tool the annotation information is put together in a separate window called *KiC: Lexical semantics* (see Picture 2-2). In the top area of the window is the list of available concepts, and in the bottom area is the definition (if available), followed by the list of variants (referred to by the indicator STRINGS). Definition and variants will appear when the cursor moves over a particular concept. If the concept has no definition, then “NODEF" will appear. Different variants are separated by “:”. For the above mentioned concepts this looks like:

```
C0087111   NODEF
           STRINGS Therapeutic procedure: Therapeutic procedure,
           NOS: Therapy, NOS: therapy: Therapy: therapies:
           Therapies: Treatment: Treatments: TREATMENT:
           Therapeutic proced

C0039798   Used with diseases for therapeutic interventions
           except drug therapy, diet therapy, radiotherapy, and
           surgery, for which specific subheadings exist. The
           concept is also used for articles and books dealing
           with multiple therapies.
           STRINGS Therapeutic aspects: therapy: treatment:
           disease management
```

The context information mentioned above will appear in the main window, *Keywords in Context* (see Picture 2-2), when selecting an occurrence for the word to be annotated. In order to choose the most appropriate concept(s) for a certain occurrence, the annotator should proceed as follows:

- read the definition and then the variants
- determine the difference between concepts
- dependent on the context of the occurrence, decide which concept fits better

Further guidelines:

- If all available concepts are very similar:
  - ... select all of them, if they are suitable for the occurrence.
- If one or more concept definitions are missing:
  - ... select the appropriate concept(s) according to the variants information.
- If concept definitions are missing and variants are not helpful:
  - ... select `Unspecified`
- It is not allowed to select both `Unspecified` and one or more concepts for the same occurrence.

2.4 Inter-Annotator Agreement

The importance of inter-annotator agreement (IAA) has been discussed in detail in (Kilgarriff 98). For the first edition of SENSEVAL IAA was on average over 90% (Kilgarriff and Palmer, 2000). For the second edition it dropped to 80%[10]. The average IAA for the GermaNet evaluation corpus is 70%. The agreement numbers for every annotated word are shown in the third column of Table 2-2. They vary from very low to very high. There are several explanations for the very low agreement scores. In cases

[10] This happened because they used WordNets instead of full dictionaries as in the first edition.

where a word had two senses, and one of them was a hypernym of the other one, the annotators took always either the most specific one, or the most general one, or both of them. Some words were not a good choice for the medical domain, for the distinction between senses was not clear at all.

We intended to check the reliability of the judgments of the annotators, using the *kappa statistic*, as described in (Siegel and Castellan, 1988) and (Carletta 1996). Unfortunately 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. Therefore kappa scores cannot really say much about the reliability or the difficulty degree of the annotation. Another unfavorable aspect is that the algorithm assumes that an annotator can only choose one sense.

After the annotators finished the task, an arbitration step followed, where they settled the disagreement cases. Removing the occurrences annotated with *undef* (632 occurrences) from the resulted annotation gave us the gold standard for the GermaNet evaluation corpus (3343 occurrences), which we used to evaluate the disambiguation methods described in sections 3.3, 3.4 and 3.5.

The agreement scores for the UMLS evaluation corpora are shown in the third column of Table 2-3 (German), Table 2-4 and Table 2-5 (English). They vary also 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 Rosensweig, 2000, §7), and 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 on German UMLS concepts. The two annotators from Germany settled disagreements for 30 English terms, which are marked in the Tables 2-4 and 2-5 with “*”. This means, for this terms the agreement score is the agreement between the American annotator on one side and the collective annotation of the German annotators. The scores for the rest of 40 terms correspond to agreement between all annotators. To evaluate the disambiguation on English UMLS concepts, precision, recall and coverage scores were obtained for each of the annotators separately and average results reported.

Ambiguous term	**Occurrences**	**Agreement**
Abnahme (reduction)	100	94%
Abweichung (aberrance, anomaly)	100	0%
Anlage (predisposition, system)	100	88%
Anwendung (procedure, treatment)	99	97%
Art (species, way)	99	87%
Ausfall (outage, loss, failure)	100	92%
Band (tape, strap)	100	100%
Bereich (area, region, domain)	100	13%
Bewegung (motion, flow, stir)	100	35%
Differenz (difference)	100	0%
Eingriff (operation, procedure)	100	99%
Fall (drop, case, instance)	100	97%
Form (shape, mode, form)	100	95%
Gebiet (zone, region, field, area)	100	73%
Gefäß (jar, vessel)	100	100%
Gesellschaft (association, community, company)	99	100%
Gewicht (weight, importance)	83	94%
Infektion (infection)	100	43%
Lage (site, status, position, layer)	99	65%
Land (country, land)	100	96%
Leistung (service, power, activity)	100	38%
Menge (amount, mass)	100	95%
Modell (model)	100	31%
Operation (operation, surgery)	100	100%
Praxis (practice, experience)	100	70%
Programm (routine, manifesto)	100	85%
Prüfung (survey, tryout, checkup)	99	99%
Raum (space, room, range, cavity)	100	45%
Sicht (sight, prospect)	99	93%
Stand (status, profession, estate)	100	84%
System (system, scheme, regime)	100	39%
Untersuchung (probe, inquiry)	100	72%
Verbindung (contact, link, tie, bond)	100	70%
Verhältnis (rate, ratio, relation)	100	7%
Verlauf (process)	100	95%
Verletzung (injury, trauma)	100	100%
Versuch (trial, test, effort, experiment)	100	47%
Wahl (ballot, choice, option)	100	99%
Weg (way, method)	99	3%
Übertragung (transmission, transfer)	99	65%
All terms	*3975*	*70%*

Table 4: Ambiguous Terms in GermaNet Evaluation Corpus

Ambiguous term	**Occurrences**	**Agreement**
Antibiotikum_Therapie (antibiosis)	15	100%
Blut (blood)	100	61%
Chirurgie (surgery)	100	83%
Epidemiologie (epidemiology)	21	57%
Genetik (genetics)	15	87%
Geschichte (history)	29	86%
Heparin_Therapie (heparin therapy)	15	100%
Laser_Therapie (laser therapy	14	57%
Leber_Transplantation (liver transplantation)	37	100%
Marker (marker)	69	65%
Metastase (metastasis)	100	50%
Pathologie (pathology)	61	100%
Physiologie (physiology)	13	92%
Rehabilitation (rehabilitation)	100	100%
Schmerz_Therapie (analgesic therapy)	82	28%
Sepsis (sepsis)	100	100%
Standard_Therapie (standard therapy)	28	18%
Stoffwechsel (metabolism)	19	100%
Strahlentherapie (radiation therapy)	98	0%
Therapie (therapy)	100	7%
Transplantation (transplantation)	92	76%
Tumor_Chirurgie (cancer surgery)	16	0%
Vergiftung (toxication, poisoning)	11	91%
Verletzung (injury, trauma)	100	99%
All terms	*1335*	*65%*

Table 5: Ambiguous Terms in German UMLS Evaluation Corpus

Ambiguous term[11]	Occurrences	Agreement
Abnormality	100	43%
*Anatomy	100	100%
Animal	100	56%
Arthrodesis	47	100%
*Aspiration	67	100%
*Atrium	60	100%
*Blood	100	97%
*Callus	34	100%
Classification	100	99%
Compliance	69	91%
*Cost	100	98%
*Deafness	100	99%
*Development	100	65%
*Diagnosis	100	19%
*Dilatation	72	85%
Education	92	0%
*Enzyme	94	100%
Etiology	100	0%
*Female	100	21%
Geriatric	93	6%
Graft	100	58%
Guideline	100	0%
*Heat	61	95%
*Irradiation	96	78%
*Lupus	55	72%
*Male	100	3%
Metabolism	100	0%
*Neoplasm	100	97%
Nursing	77	4%
*Nutrition	66	92%
*Operation	100	2%
*Organization	43	7%
Oxygen	100	82%

Table 6: Ambiguous Terms in English UMLS Evaluation Corpus (continues on next page)

[11] For the words marked with '*' two annotators settled the disagreement cases (see Section 2.4). For the other words there was no arbitration.

Ambiguous term	Number of occurrences	Agreement
*Oxygenation	81	99%
*Pace	100	100%
*Para_thyroid	28	100%
Pathology	100	5%
Personnel	56	0%
*Pneumothorax	53	87%
*Plaque	87	99%
*Para_thyroid	28	100%
*Prostate	100	25%
Prosthesis	100	32%
*Radiography	65	0%
Radiology	33	27%
*Radiotherapy	100	100%
*Regulation	100	67%
*Rehabilitation	100	0%
Secondary	100	92%
Secretion	65	63%
Standard	100	0%
Supply	98	71%
Surgery	100	4%
Survival	100	0%
Tear	74	5%
Temperature	100	84%
Testis	32	100%
Therapy	100	38%
Thyroid	100	99%
Transplant	100	26%
Transplantation	100	1%
Trauma	100	0%
Treatment	100	22%
Ultrasound	100	0%
Urine	76	9%
Ventilation	100	11%
Vessel	100	99%
Water	92	94%
Weakness	48	92%
Weight	100	76%
x-ray	100	65%
All terms	*6014*	*51%*

Table 7: Ambiguous Terms in English UMLS Evaluation Corpus (continued from previous page)

3 Methods and results for disambiguation

3.1 Overview

Methods for disambiguation 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. (The exception to this is that it makes sense to use the output of unsupervised algorithms as training examples for algorithms, which benefit from having training examples available, as described in Section 3.6.)

The methods we have developed fall into the following categories. Bilingual methods (Section 3.2) take advantage of having a translated corpus, because knowing the translation of an ambiguous word can be enough to determine its sense. Dictionary based methods (Section 3.3) use relations between terms as deduced from a dictionary or some other semantic resource to determine which sense is being used in a particular instance. Domain-specific methods (Section 3.4) use the fact that certain meanings of general terms are far more important than others in specific domains (for example, in the medical domain, “operation” is far more likely to refer to a surgical operation than a military operation), a form of disambiguation that can also be regarded as lexical tuning. Instance-based learning (Section 3.5) is a machine-learning technique that we applied to unsupervised training in word-sense disambiguation.

3.2 Bilingual

The mapping between word-forms and senses differs across languages, and for this reason the importance of word-sense disambiguation has long been recognized 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 prescription drugs.

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 should “prendre” be translated as “take” or “make”), the number of acceptable translations produced by their system increased by 8%. Gale, Church and Yarowsky (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 bilingual disambiguation technique to entire dictionaries and corpora.

We used the bilingual Springer corpus in which both the English and German abstracts had been tagged with UMLS concept-unique-ID's (CUI's). We considered a term to be ambiguous if it had been assigned more than one CUI by this tagging. To disambiguate an instance of an ambiguous term, we consulted the translation of the abstract in which it appeared. We considered the translated abstract to disambiguate 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).

We consider these disambiguation criteria to be reasonable and relatively strict: that is, we would expect that when a term is judged to have been disambiguated according to the criteria we will have either a genuine, successful disambiguation or a store of assigned CUI's that is impoverished in one language with respect to the other. This assumption is discussed below in the context of the results obtained using this procedure.

Results for Bilingual Disambiguation

We applied this process to the 6374 German abstracts and their English translations in both directions. That is, 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 following results:

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

Table 8: Result s 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 as many correct decision as the English, but many more incorrect ones as well.

Our disambiguation results break down into three cases:

- 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.

- Terms ambiguous in one language that translate as multiple unambiguous terms in the other language; both of the terms are medical

- Terms that ambiguous between two meanings that are only very slightly different, or are difficult to distinguish without specialized medical knowledge.

One striking aspect of the results is 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 of ambiguity account for the vast majority of cases in which only one meaning is ever selected. It is easy to see why this would happen in the first category, and it is what we want to happen. For instance, the German term “Krebs” 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 finds the meanings of approximately one fifth of the ambiguous terms, and makes only a few mistakes for English but many more for German.

3.3 Dictionary (UMLS) based

3.3.1 Collocations

There is a strong and well-known tendency for words to express only one sense in a given collocation. For example, consider two definitions of the word “plant” (given by Merriam-Webster):

i. (a) a young tree, vine, shrub, or herb planted or suitable for planting (b) any of a kingdom (Plantae) of living things typically lacking locomotive movement or obvious nervous or sensory organs and possessing cellulose cell walls

ii. (a) the land, buildings, machinery, apparatus, and fixtures employed in carrying on a trade or an industrial business (b) a factory or workshop for the manufacture of a particular product (c) the total facilities available for production or service (d) the buildings and other physical equipment of an institution

In almost every instance, the phrase “plant life” will refer to a meaning of the word ‘plant’ from sense 1, and the phrase “manufacturing plant” will refer to a meaning of ‘plant’ from sense 2.

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) uses the one sense per collocation property as an essential ingredient for an unsupervised Word-Sense Disambiguation algorithm. To disambiguate between the above senses of “plant”, the collocations “plant life” and “manufacturing plant” are used as ‘seed-contexts’. The algorithm bootstraps from instances of the word “plant” in these collocations to obtain other classifiers, which indicate that one sense or the other is being used. For example, in Yarowsky's experiment the words “animal” and “species” often occur with the collocation “plant life” and the terms “equipment” and “employee” often occur with the collocation “manufacturing plant” (and rarely with the opposite collocations). These terms can then also be used to indicate which sense of “plant” is being used in a particular context. In effect, Yarowsky's algorithm uses instances of “plant” in the collocations “plant life” and “manufacturing plant” 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

i. to recognise which ambiguous terms are amenable to this technique,

and

ii. to choose appropriate “seed collocations” for each sense.

Thus the algorithm still requires expert human judgements, 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, Yarowsky 1992), using words appearing in the definitions as statistical classifiers.

The richly detailed structure of UMLS provides a special opportunity to combine both of these approaches. This is because many multiword expressions and collocations are included in UMLS as separate concepts.

Example

Consider the term ambiguous term “pressure”, which in UMLS can mean

```
i.   C0033095     Physical agent pressure, physical pressure
ii.  C0460139     Pressure - action
iii. C0234222     Baresthesia, pressure sense, sensation of
                  pressure
```

We can use the existing structure of UMLS to provide a method for disambiguating certain instances of the term “pressure”, using collocations which are themselves in UMLS.

UMLS classifies each of these definitions to a particular semantic type, as follows:

i.	Physical agent pressure, physical pressure	Quantitative Concept
ii.	Pressure - action	Therapeutic or Preventive Procedure
iii.	Baresthesia, pressure sense, sensation of pressure	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	bar pressure, mean pressure, peak pressure, population pressure
Therapeutic or Preventive Procedure	acupressure, orthostatic pressure, apply end expiratory negative pressure
Organ or Tissue Function	arterial pressure, lung pressure, intraocular pressure

This leads to the hypothesis that the term “pressure”, when used in any of the above 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 agent pressure, physical pressure

UMLS provides thousands of such examples. To obtain a reliable subset, we have proceeded as follows. Nearly all English and German multiword technical medical terms are head-final which the previous terms are modifying or making more specific. (So for example, “lung cancer” is a kind of cancer, not a kind of lung.) It follows that the a multiword term is usually of the same semantic type as its head, the final word.

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). Of these 350,000, over 50,000 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.

Results for collocational disambiguation

Unfortunately, results for collocational disambiguation 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 9: 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 even more 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 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.

3.3.2 Disambiguation using related UMLS terms found in the same context

While the method above 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, a new approach to extracting information for disambiguation from UMLS was suggested to us by an invited expert at the MuchMore workshop in Hvar, Croatia -- see semestrial report #5.

- What we were effectively doing with the collocational method was using UMLS to give information about ambiguous words and other words which, when they occurred with the ambiguous word, would help to predict the correct sense.
- There were many other sources of information in UMLS, which would give other words, which might indicate that an ambiguous term was being used with one a particular sense.
- In particular, we should consider terms that were linked by conceptual relations (as given by the MRREL and MRCON files) and which were noted as co-indexing concepts in the same MEDLINE abstract (as given by the MRCOC file).
- For each separate sense of an ambiguous word, this would give a set of related concepts
- If any of these related concepts could be 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. Effectively, a predesigned lexical resource is being used to give words that might be indicative of one sense or another.

This technique turned out to be particularly effective for the MUCHMORE project, once it was determined how to get such information from UMLS, which contains a great deal of information besides standard definitions. In particular, we gain over purely dictionary-based methods because the words that occur in dictionary definitions rarely correspond well with those that occur in text. On the other hand, the information we collected from UMLS, in particular the cooccuring concepts information, was derived precisely from knowing which concepts occurred together in similar contexts.

The disambiguation algorithm was thus as follows:

```
For each ambiguous word

     Find its possible senses (CUI's)

     For each sense

          find all CUI's in MRREL, MRCON or MRCOC files that
          are related to this sense.

For each occurrence of the word in the corpus

     Examine local context to see if any of the related CUI's
     appear
```

```
If so, assign this instance of the ambiguous word to the
sense related to this nearby concept.

If concepts related to more than one of the possible
senses occur, resolve the issue by majority voting
```

This algorithm fails to take into account the fact that the 'related concepts' might themselves be ambiguous, and so performance may be improved still further by allowing for the *mutual* disambiguation of more than one term at once, as implemented by Stevenson and Wilks (2001).

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 *DerChirurg*.

Thus our results (for each language) vary according to which knowledge sources were used (Conceptually Related Terms from MRREL and MRCXT or cooccuring indexing terms from MRCOC, or a combination), and according to whether the context-window for recording cooccurence was a sentence, a document or a subdomain.

Results for disambiguation based on UMLS related terms

The results obtained using this method have been 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 – this 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, 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 this knowledge source. As we had expected, recall and coverage increased but precision decreased slightly when using larger contexts.

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 10: Results for disambiguation based on UMLS relations (English)

The German results were slightly different, and even more successful, with nearly 60% of the evaluation corpus being successfully disambiguated, and 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 are 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.

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 11 Results for disambiguation based on UMLS relations (German)

Comparing these results with the number of words disambiguated in the whole corpus (Table 2) 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 that 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 is a very exciting proposal which should stimulate further research, and not too far behind, successful practical application.

3.4 Domain-Specific Sense

An ambiguous word can have general and domain specific senses, e.g. `Gewebe` (*tissue*) in Table 3-5.

Offset	Synset
607925	`Gewebe, Koerpergewebe` (*tissue, body tissue*)
1578773	`Kleiderstoff, Textilstoff, Gewebe, Webware, Stoff` (*tissue, cloth, textile*)

Table 12: Senses from "Gewebe"

When the word occurs in a domain specific corpus, it may have a strong preference for one of its domain specific senses, see (Cucchiarelli and Velardi, 1998), (Magnini et al., 2001). Starting from this idea (Buitelaar and Sacaleanu, 2001) developed a method that determines the domain specific relevance of a GermaNet synset on the basis of its statistical relevance across several domain specific corpora. This method is part of a larger effort to develop semi-automatic methods for domain specific lexicon construction that builds on the reuse of existing resources.

In a first step domain specific corpora are annotated with a shallow processing tool and frequency values are computed for all noun lemmas (terms). The relevance of a term for each domain is then computed using a slightly adapted version of standard tf.idf, as used in the vector-space models for information retrieval (Salton and Buckley, 1988). The relevance formula is shown in (3.1), where *t* is the term, *d* the domain and *N* the number of domain corpora. This measure gives full weight to terms that occur in just one domain and 0 weight to those occurring in all domains.

$$(3.1) \quad rlv(t \mid d) = \log(tf_{t,d}) \log(\frac{N}{df_t})$$

The relevance for a concept can be computed using the relevance values of the terms occurring in the corresponding synset. The intuitive way is to sum up the relevance values:

$$(3.2) \quad rlv(c \mid d) = \sum_{t \in c} rlv(t \mid d)$$

For some concepts, this does not seem to function properly. For example the concepts for the term Zelle look like this:

`[Zelle, Gefaengniszelle]`	*prison cell*
`[Zelle]`	*living cell*

Zelle has a high relevance in the medical domain while Gefaengniszelle is very unlikely to occur, so its relevance will be close to 0. Using the formula in (3.2), both concepts will

have the same relevance, which is wrong, because concept (2) is much more relevant to the medical domain than concept (1).

The formula in (3.2) is therefore reconsidered to take into account the number of concept terms that actually occur in the domain corpus (lexical coverage). The new formula is shown in (3.3), where *T* represents the lexical coverage and |*c*| the concept length:

$$(3.3) \quad rlv(c \mid d) = \sum_{t \in c} \frac{T}{|c|} rlv(t \mid d)$$

The intuition behind this formula is that the more concept terms occur in the domain corpus, the more relevant the concept is for that domain. However, the measure in (3.3) has also two handicaps. First, it has a preference for concepts of length 1, because the lexical coverage related to concept length is maximal. This means for example that the medical sense of `Zelle` will be always preferred in every domain, unless `Gefaengniszelle` actually occurs in the domain corpus. Secondly, if the concepts corresponding to a term with domain relevance have the same length, and the other terms (synonyms) do not occur in the domain corpus, the concepts are asigned the same relevance by the measure in (3.3). For example, the two senses of `Geschlecht` will get the same relevance if neither Haus nor Sexus occur in the domain corpus.

`[Geschlecht, Haus]` *family line*

`[Geschlecht, Sexus]` *gender*

In order to avoid these problems and to increase the number of terms to be found within a domain corpus, more lexical information is added to the relevance measure. The new formula (3.4) considers also the relevance values for all hyponyms for every term in the concept (*c*+ is the concept extended with hyponyms for every term):

$$(3.4) \quad rlv(c+ \mid d) = \sum_{t \in c+} \frac{T}{|c|} rlv(t \mid d)$$

Adding hyponyms does not change the lexical coverage, but increases the summed concept weight. The extended concepts for `Zelle` look like this:

`[Zelle, Gefaengniszelle, Todeszelle]`

`[Zelle, Koerperzelle, Pflanzenzelle]`

3.4.1 Application

The disambiguation algorithm using the domain specific sense method is as follows:

- For every GermaNet term in the medical corpus compute the relevance for the medical domain, using several domain specific corpora, as described in the previous section.
- For every occurrence in the evaluation corpus assign the sense with the highest relevance, if available.

Evaluation metric

The measure used for the evaluation of the disambiguation results is the exact match criterion. The sense *s* assigned by the WSD system to an occurrence in the evaluation corpus is considered correct if:

- *s* is the same as the sense in the gold standard OR
- *s* belongs to the set of senses in the gold standard

For every experiment coverage, precision and an F-measure were computed:

$$coverage = \frac{|disambiguated\_occurrences|}{|all\_occurrences|}$$

$$precision = \frac{|correct\_disambiguated\_occurrences|}{|disambiguated\_occurrences|}$$

$$F = \frac{2 \times coverage \times precision}{coverage + precision}$$

Baseline

We decided to compare our results with a theoretical baseline. The precision of random sense assignment may be computed using the formula in (3.5), where *GS* means the gold standard.

$$(3.5) \quad prec_{random} = \frac{1}{|all\_occurrences|} \sum_{o \in GS} \frac{|gold\_st\_sense(s)|}{|GermaNet\_senses|}$$

That is, for every occurrence in the gold standard, the probability of assigning the correct sense(s) is computed by dividing the number of senses in the gold standard by the number of corresponding GermaNet senses. The average precision is the sum of all

probabilities divided by the number of all occurrences. For our evaluation corpus this gives a baseline precision of 36%, at coverage of 100%. The corresponding F-measure is 0.53.

Results

For all GermaNet senses in the training corpus a domain relevance score was computed. The experiments were conducted with different sets of domain specific corpora and with different corpora sizes. The corpora used are:

- medical corpora:
  - Springer (sp): medical abstracts
  - Radiology (rad): examination reports
- other domain corpora:
  - Deutsche Presse Agentur (dpa): news
  - Fussball (fb): soccer game reports
  - Wirtschaftswoche (wrt): economic news

In disambiguation, the sense with the highest domain relevance was selected. Because this sense depends on the domain and not on a particular context, all occurrences of an ambiguous word will be assigned the same sense. No decision was made in cases where (a) no sense had a relevance value or (b) two or more senses had the highest relevance value. Table 3-6 shows the evaluation results for different corpora sets and sizes.

Corpora	Size	Coverage	Precision	F
sp-dpa-fb-wrt	2 MB	12%	77%	0.20
sp-dpa	2 MB	6%	99%	0.11
sp-dpa	10 MB	17%	26%	0.20
rad-dpa-fb-wrt	2 MB	38%	44%	0.40
rad-dpa	2 MB	18%	50%	0.26
rad-dpa	10 MB	9%	34%	0.14
rad-dpa	20 MB	4%	31%	0.07

Table 13: Disambiguation Performance with Domain Specific Sense

The F-measure column indicates that no experiment improved on the baseline mentioned before. However, this method does not play the main role in our WSD system. It is meant to assist and improve the instance-based learning method. Nevertheless the results are interesting. We conducted these experiments not only to measure the concrete

performance but also to find out how performance changes for different values for parameters:

Springer vs. Radiology

The best F-measure is achieved with Radiology, but we are more interested in precision and its highest values are reached with Springer, on the cost of very low coverage. For Radiology coverage and precision are much closer to each other. An interesting question is why the coverage is much better for Radiology than for Springer, which is at the same time the test corpus. All the words from the evaluation corpus are contained in Springer. Unfortunately, these words have beside medical senses also general senses, so many of them occur also in the other domain specific corpora. According to the domain specific sense method, terms occurring in all corpora are assigned the weight 0, which means, no relevance is computed for them, and no disambiguation is possible. On the other side, Radiology has a much more restricted vocabulary and does not contain many of the evaluation words. So even if they appear in other domain specific corpora, they will still get relevance values, which leads to better coverage.

Number of different used corpora

Coverage grows with the number of domain specific corpora but unfortunately the precision gets lower. The hypothesis was that the more corpora the higher the precision, but even if the evaluation terms are specific for medicine, they also have other, more general interpretations, so the medical sense could get a lower relevance than the general one(s).

Corpora size

The performance is much lower for large corpora, which can be explained by the fact that they have a corresponding large set of common terms, which may influence coverage as well as precision.

3.5 Instance-Based Learning

The growing availability of large machine-readable corpora and the software and hardware performance improvements in the last decade initiated the use of statistical learning methods in natural language processing. The success of these statistical methods in speech recognition (Stolcke 1997, Jelinek 1998) motivated their application in other tasks like morphological and syntactic analysis (Charniak 1997), semantic disambiguation and interpretation, discourse processing and information extraction or machine translation (Knight 1997). The statistical methods use particular statistical techniques such as hidden Markov models, naive Bayes, maximum entropy, expectation maximization, probabilistic context-free grammars, etc. Another category of machine learning approaches employ typical learning paradigms like decision tree and rule-induction, neural networks, instance-based, Bayesian network learning, inductive logic programming, explanation-based learning, and genetic algorithms.

A machine-learning algorithm uses an internal representation. We can classify them according to the abstraction level of this representation. Some of the well-known

representations are: decision tables, decision trees, classification rules, association rules, rules with exceptions, rules involving relations, trees for numeric prediction, instance-based representation, clusters. A detailed description of representations and general methodologies can be found in (Witten and Eibe, 2000).

Instance-Based Learning

The input for a machine learning algorithm can be represented as a set of features or attributes, one of which identifies the class attribute. A particular input consists of a set of values for these attributes. We will call this set an instance. In a classifying task the system stores a set of training instances. Using this knowledge base, the system should then be able to assign a new instance with a missing value for the class attribute the corresponding attribute value. This algorithm is called *instance-based learning*, because it uses the instances themselves to represent what is learned, rather then inferring a more abstract internal representation.

In the nearest neighbour classification method, the instance which must be classified is compared with all training instances, using a distance metric, and the closest training instance is then used to assign the class to. The generalization of this method is the *k-nearest neighbour* method, where the class of the new instance is computed using the closest *k* training instances.

3.5.1 Training

The main method developed within our disambiguation system uses a k-nearest neighbour instance-based learning algorithm. To develop this method we used the WEKA[12] (**W**aikato **E**nvironment for **K**nowledge **A**nalysis) package, which implements several machine learning algorithms and methodologies in the Java programming language. This section describes the typical data structure for instance-based learning, as well as the training and disambiguation algorithms with the corresponding parameters.

Data structure

In the previous section we mentioned that the input for instance-based learning is represented by instances that are sets of attribute-value pairs, one of which identifies the class attribute. WEKA can only process instances in a particular format, called the ARFF format. To illustrate this format, consider the problem of deciding if an outside game should take place, given the weather conditions. The attributes describing the weather are: *outlook*, *temperature*, *wind intensity* and the class attribute is the *game status*. A possible training set in the ARFF format looks like in Table 3-7.

The ARFF format contains three main blocks:

- a generic task name (*weather*) introduced by **@relation**;
- an attribute block which defines name and type for each attribute (including the class attribute); every line starts with **@attribute**; the type can be numeric or nominal and in the second case all possible values must be listed;

[12] http://www.cs.waikato.ac.nz/~ml/weka/

- a data block introduced by **@data**, which lists attribute values for all training instances; missing values are represented by "?"; there is no distinction for the class attribute, because different tasks require different class attributes.

@relation weather
@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute windy { true, false}
@attribute play { yes, no }
@data
sunny, 85, false, no
sunny, 80, true, no
overcast, 83, false, yes
rainy, 70, false, yes
rainy, 65, true, no
overcast, 64, true, yes
sunny, 69, false, yes

Table 14: ARFF Format

Building instances

Before describing the training algorithm, we need to explain how our system constructs instances, given a particular input. Our input usually consists of sentence fragments, whose length depends on a particular parameter. Let *w* be the central noun in such an input. We can build several instances for *w* where the attributes are the lemmas of its left and right neighbour words in a context of size n, and the class attribute varies over its GermaNet synsets ids. If no lemma is available for a word, the value of the corresponding attribute is the word form itself. To illustrate this, let us consider the following sentence:

(3.6) *In dem Fall sind korrigierende* Eingriffe *nur eingeschränkt möglich.*

(*In this case, the possibility of corrective surgery is limited.*)

The word Eingriffe is ambiguous and has the following senses:

460326 [Operation, Eingriff]

(surgery, operation)

388935 [Eingriff, Intervention, Eingreifen]
(*intervention, invasion)*

Given the sentence (3.6), we can build the following instances for Eingriffe with context size 5 (two words left and 2 words right):

(3.7) *sein, korrigieren, nur, einschränken, 460326*
sein, korrigieren, nur, einschränken, 388935

Every context corresponds to a part-of-speech pattern, in our case the pattern is [- ADJ NN – VERB] with Eingriffe taking the position of NN ("-" stands for other parts-of-speech).

The training algorithm

We can now present the training algorithm. Given a training corpus annotated with part-of-speech and morphology, for any ambiguous word *w* from the evaluation corpus and its set of synset ids *S* do the following:

- determine all part-of-speech patterns of size *n* in which *w* occurs in the evaluation corpus;
- for every part-of-speech pattern :
    - extract all contexts in the training corpus;
    - for every context build the corresponding instances, under the constraint that the value of the class attribute belongs to *S*;
    - collect all instances from all contexts in a training set *I*(*w*, *p*);
    - eliminate duplicates;

When the training process is done, we will have for every ambiguous word in the evaluation corpus several training sets in ARFF format, one for every part-of-speech pattern, in which the word occurs.

A training set for Eingriff in the pattern [- ADJ NN - VERB] is shown in Table 3-8.

Parameters
The training process implements several parameters as follows:

- *Pattern frequency*: the minimal frequency a part-of-speech pattern must have to be considered in the training corpus; if a pattern has to few realizations in the training corpus, they can not generate a reliable training set;
- *Training corpus*: the medical corpus used for training: Springer or Radiology;
- *Relevant attributes*: relevant attributes when building instances:

- all attributes (i.e. all parts-of-speech);
- only attributes which correspond to NN/ADJ/VERB parts-of-speech; in this case the value for non-relevant attributes is null;

- *Context size*: this parameter says how many left and right neighbours of the ambiguous word are considered when building training instances; **3** means one neighbour left and one neighbour right; **5** means two neighbours left and two neighbours right;

@relation dataset
@attribute att1 {ein,der,oder,und,bei,nach,sein} @attribute att2 {offen, therapeutisch, planen, diagnostisch, modern, begrenzen, nochmalig, endonasaler, zweit, verbal, unterschiedlich, orrigieren, chirurigischen} @attribute att3 {zu, nicht, und, wieder, werden, aufeinander, nur} @attribute att4 {erfassen, ermöglichen, bewerten, schaffen, sichern, bevorzugen, vorstellen, ersparen, beziehen, quantifizieren, bedachen, einschränken, profitieren} @attribute att5 {388935,460326}
@data und,diagnostisch,und,ermöglichen,388935 und,diagnostisch,und,ermöglichen,460326 der,planen,zu,bewerten,388935 der,planen,zu,bewerten,460326 bei,modern,zu,schaffen,388935 ein,begrenzen,und,sichern,388935 ein,nochmalig,wieder,bevorzugen,460326 nach,endonasaler,werden,vorstellen,460326 ein,zweit,zu,ersparen,388935 ein,zweit,zu,ersparen,460326 und,verbal,aufeinander,beziehen,388935 nach,unterschiedlich,zu,quantifizieren,460326 ein,therapeutisch,nicht,bedachen,388935 sein,korrigieren,nur,einschränken,388935 sein,korrigieren,nur,einschränken,460326 und,therapeutisch,werden,vorstellen,460326 ein,chirurigischen,nicht,profitieren,388935 oder,offen,zu,erfassen,460326

Table 15: Training Set for "Eingriff" in the Pattern [- ADJ NN:*Eingriff* – VERB]

Application

After collecting training sets for all part-of-speech patterns for all words we want to disambiguate, we can start the disambiguation and disambiguation process. For every occurrence of an ambiguous word from the evaluation corpus do the following:

- determine the part-of-speech pattern *p* of length *n*;
- extract the corresponding training set *I*(*w*, *p*);
- delete all instances corresponding to the occurrence itself, generating *I'*(*w*, *p*);
- create a new instance *i* for the occurrence, with a missing value for the class attribute;
- ask the WEKA system to classify *i* by searching for the most similar instance in *I'*(*w*, *p*);
- analyze the probability distribution provided by WEKA for all senses for *w*, and if there is a sense with a highest probability, assign it to the occurrence.

Some of these steps can be illustrated by continuing the example from the previous section. We can disambiguate the occurrence of `Eingriff` in (3.6) using the training set in Table 3-8 - *I*(*w*, *p*). To obtain *I'*(*w*, *p*) we have to delete from *I*(*w*, *p*) the instances shown in (3.7). This step is important because the instance, which must be classified, should not be in the training set already. Because the system is unsupervised, the instances corresponding to an occurrence of an ambiguous word are identical except for the value of the class attribute (sense). If we try to classify this occurrence, all senses will get the same probability, so there is no real disambiguation. In the next step a new instance is created, with the missing value for the class attribute:

```
sein, korrigieren, nur, einschränken, ?
```

The new instance can then be classified using *I'(w,p)*. This algorithm guarantees that the training set used for classifying a new instance contains no identical instances. If the attribute values from the new instance do not occur at all in the training set, the instance is hard to classify.

The disambiguation process uses the same parameters as the training process. We have to use the same pattern frequency, relevant attributes, and context size, otherwise the disambiguation can not take place. Even when the same parameters are used in both training and disambiguation, there are three reasons why an occurrence of an ambiguous word from the evaluation corpus cannot be disambiguated:

- the part-of-speech pattern of the occurrence has a very small frequency, therefore no training set can be built for it;
- the left or right context of the occurrence is to small; if the word is the first or last in the sentence, it has no left or no right neighbours, therefore no instance can be build for it; if the word is the second or the last but one, it has a context of size 3 but not of size 5;

- the occurrence has a normal context and a training set was built for it, but in the classification process all senses get the same probability.

Results

This disambiguation method was evaluated for different values of training and disambiguation parameters and the results are shown in Table 3-9. These experiments were made for k = 1 (k nearest neighbours).

Training corpus	Context size	Pos	Coverage	Precision	F
Springer corpus	3	All	62%	49%	0.55
	3	N/V/A	39%	43%	0.41
	5	All	33%	54%	0.41
	5	N/V/A	44%	47%	0.45
Radiology corpus	3	All	49%	43%	0.46
	3	N/V/A	30%	44%	0.36
	5	All	31%	42%	0.36
	5	N/V/A	33%	48%	0.39

Table 16: Disambiguation Performance with IB1

Training Corpus

We were interested to see how well the system performs when training and application use the same corpus (Springer) compared to when the training corpus is different from the test corpus, but still belonging to the same domain (Radiology). As expected, precision and coverage are better in the first case.

Context Size

We experimented with contexts of size 3 and 5. For smaller contexts the coverage is much better, but precision reaches its highest values for contexts of size 5. Larger contexts contain more relevant information, which can contribute to the selection of a particular sense. For contexts larger than 5 the training instances become too sparse and the coverage gets very low.

Part-of-Speech Selection

Here two cases were considered: (a) *all* - all attributes are relevant; (b) *N/V/A* - only attributes corresponding to nouns, verbs and adjectives are relevant. With contexts of size 3 precision values are better when all attributes are relevant. This makes sense because in many small contexts no nouns, verbs, or adjectives occur, so no useful training instances can be built. With context size 5, the results are different for different training corpora.

For Springer, precision is better when using all parts-of-speech are used (54% vs. 47%) while for Radiology filtering out attributes corresponding to other parts-of-speech than N/V/A leads to a better precision (48% vs. 42%).

The best performance was reached with Springer as training corpus, context of size 3 and using all attributes (coverage: 62%, precision: 49%). The F-measure (0.55) is better than for the baseline (0.53). The results in the other experiments (except for one) are below the baseline.

The next set of experiments was made using the Springer corpus, considering all parts-of-speech relevant and varying *k*. The results are shown in Table 3-10.

Springer corpus	Context: 3			Context: 5		
K	Coverage	Precision	F	Coverage	Precision	F
1	62%	49%	0.55	33%	54%	0.41
3	62%	49%	0.55	35%	55%	0.43
6	65%	48%	0.55	41%	53%	0.46
9	66%	48%	0.55	45%	51%	0.48
12	67%	48%	0.56	46%	51%	0.48
15	68%	47%	0.55	48%	51%	0.49
18	69%	47%	0.56	49%	51%	0.50

Table 17: Disambiguation Performance with IBk

Increasing *k* leads in general to better values for coverage, precision and F-measure. For contexts of size 3 the coverage gets better (+ 7%), while the precision gets very little worse and the F-measure remains constant. For contexts of size 5 the precision gets a little bit worse, but the coverage and the F-measure get much better (+16 % respectively +9%).

3.6 Combined methods

The next step was to combine the methods described in Sections 3.4 and 3.5, trying to improve the performance obtained in previous experiments. The disambiguation algorithm is as follows:

- decide in which order the methods should be applied;
- for every occurrence from the evaluation corpus do:
  - apply the first method;
  - if a decision is made, assign the resulting sense to this occurrence;
  - otherwise, apply the second method;

When combining the two methods no additional training step is necessary because the methods are independent of each other, therefore we can use the training results from the previous experiments. The additional disambiguation parameter is the order in which the two methods are applied.

Results

For these experiments we used the domain relevance values which led to the best results in the experiments with the domain-specific method (first row in Table 3-6) and the sets of training instances generated with the instance-based learning method (k = 1). For every occurrence of an ambiguous word we applied the two methods disjunctively, that is, if the first method could not make any decision, the second one was applied. We varied the order in which the methods were applied. Table 3-11 shows the results for this set of experiments.

Training corpus	**Context size**	**Pos**	IB1→ Domain Specific Sense			Domain Specific Sense → IB1		
			Cov.	**Prec.**	**F**	**Cov.**	**Prec.**	**F**
Springer corpus	3	All	67%	52%	0.58	67%	53%	0.59
	3	N/V/A	47%	50%	0.48	47%	51%	0.49
	5	All	41%	60%	0.49	41%	60%	0.49
	5	N/V/A	52%	53%	0.52	52%	53%	0.52
Radiology corpus	3	All	64%	45%	0.53	64%	44%	0.52
	3	N/V/A	55%	44%	0.49	55%	43%	0.48
	5	All	53%	46%	0.49	53%	44%	0.48
	5	N/V/A	59%	49%	0.53	56%	46%	0.50

Table 18: Disambiguation Performance with Combined Methods

This set of experiments produced the best performance for *k*=1. In particular the experiments with Springer as training corpus, using contexts of size 3 and all attributes are significantly above the baseline (F-measure 0.58 and 0.59 compared to 0.53). The F-measure got also much better for all others experiments, still remaining below the baseline. This proves that even if domain specific sense is not very performing when applied alone, it brings much improvement when assisting another WSD method. The two methods were only used disjunctively but we would expect a better precision when using them conjunctively, which is assigning an instance a sense when both methods agree, in cases where both methods can make a decision. For cases where just one method can say something, the system would just accept the respective decision.

Applying the instance-based method first produces slightly better results than applying the domain specific method first. This may result from the fact that the latter always

selects the same sense for every occurrence of an ambiguous word, whereas the first selects a sense depending on a particular context.

Comparing the results from the combination of methods with the results produced by the two methods separately shows that both precision and coverage are better than for instance-based learning, but the best precision (60%) does not reach the highest values (77%, 99%) with the domain specific sense method.

4 Conclusions

The main conclusion from this research is that high precision, broad coverage disambiguation of medical documents can be achieved without the costly annotation of many training examples. The best results for precision ranged from 74% (English) to 79% (German), achieved by the UMLS related terms method (Section 3.3.2) on the UMLS evaluation corpus, and from 77%-99% achieved by the Domain Specific Sense method (Section 3.4) on the GermaNet evaluation corpus. The best results for Coverage range from 67% achieved by Instance Based Learning (Section 3.5) on the GermaNet evaluation corpus, to 83% (English) and 87% (German) achieved by the UMLS related terms method (Section 3.3.2) on the whole Springer corpus.

While none of these methods required manually annotated training data, the more precise methods relied on other sources of knowledge for their success. In particular, the UMLS related terms method (Section 3.3.2) made use of the detailed structure of UMLS and the way UMLS terms have been used by human experts to index MedLine articles. The collocational method (Section 3.3.1) also relied on UMLS as a knowledge source, and the bilingual method (Section 3.2) relied on the availability of a parallel corpus – it would be impractical to construct these resources purely for the sake of disambiguation. The Domain Specific Sense method (Section 3.4) and the Instance-Based Learning method (Section 3.5) were less resource intensive, using only the structure of GermaNet and domain specific corpora for training.

The best single method was the UMLS related terms method (Section 3.3.2), which achieved excellent results for precision and coverage *in both languages*, even though the knowledge source used to give related terms was based entirely on relations in *English* documents. It follows that this method could be applied in exactly the same way to *all* the languages covered by UMLS, without the need for any extra resources for training. Document retrieval experiments are planned to test whether the disambiguation provided by this method is beneficial for information retrieval – if so, this powerful technique could be implemented relatively easily to improve access to medical information in several European languages.

References

Brown P., de la Pietra S., de la Pietra V., Mercer R. *Word Sense Disambiguation Using Statistical Methods*, ACL 29, pages 264-270, 1991.

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

Carletta J.C. *Assessing Agreement on Classification Tasks: the Kappa Statistic*. In: Computational Linguistics 22(2), pages 249-254, 1996.

Charniak E. *Statistical Techniques for Natural Language Parsing*. In: AI Magazine 18(4), pages 33-43, 1997.

Cucchiarelli A., Velardi P. *Finding a Domain-Appropriate Sense Inventory for Semantically Tagging a Corpus*. In: Journal of Natural Language Engineering, 1998.

Fellbaum C. *WordNet: An Electronic Lexical Database.* MIT press, 1997

Gale W., Church K., Yarowsky D. *A Method for Disambiguating Word Senses in a Large Corpus*. Computers and the Humanities 26, pages 415-439, 1992.

Hamp B., Feldweg H. *GermaNet: A Lexical-Semantic Net for German*. In: P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.): Proceedings of the ACL/EACL97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, Spain, 1997.

Ide N. *Parallel Translations as Sense Discriminators*. SIGLEX99: Standardizing Lexical Resources, ACL99 Workshop, College Park, Maryland, pp52-61, 1999.

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

Jelinek F. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1998.

Kilgarriff A. *Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs*. In: R. Gaizauskas (ed): Computer Speech and Language 12(4), Special Issue on Evaluation of Speech and Language Technology, pages 453-472, 1998.

Kilgariff A., Palmer M. *Introduction to the Special Issue on SENSEVAL*. In: Computers and the Humanities (34), pages 1-13, 2000.

Kilgarriff A., Rosenzweig J. *Framework and results for English SENSEVAL*. In:

Computers and the Humanities (34), pages 15-48, 2000.

Knight K. *Automatic Knowledge Acquisition for Machine Translation*. In: AI Magazine 18(4), pages 81-96, 1997.

Lesk M.E. *Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cone.* In: Proceedings of the SIG-DOC Conference, 1986.

Magnini B., Strapparava C., Pezzulo G., Gliozzo A.. *Using Domain Information for Word Sense Disambiguation.* In: Proceedings of SENSEVAL-2, ACL, Toulouse, France, 2001.

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

Miller G.A. *Nouns in WordNet.* Chapter 1 in Fellbaum 1997, op. cit.

Plaehn P., Brants Th. *Annotate – An Efficient Interactive Annotation Tool.* In: Proceedings of the 6th Conference on Applied Natural Language Processing ANLP, Seattle, WA, 2000.

Salton G., Buckley C. *Term-Weighting Approaches in Automatic Text Retrieval*. In: Information Processing & Management 24(5), pages 515-523, 1988.

Siegel S., Castellan N.J. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Second Edition, 1988.

Stevenson M. and Wilks Y. *The Interaction of Knowledge Sources in Word Sense Disambiguation*. In: Computational Linguistics 27(3), 2001.

Stolcke A. *Linguistic Knowledge and Empirical Methods in Speech Recognition*. In: AI Magazine 18(4), pages 25-31, 1997.

Vintar S., Buitelaar P., Ripplinger B., Sacaleanu B., Raileanu D., Prescher D. *An Efficient and Flexible Format for Linguistic and Semantic Annotation*. In: Proceedings of LREC2002, Las Palmas, Canary Islands – Spain, May 29-31, 2002.

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

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

Witten I.H., Eibe F. *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA, 2000.

Yarowsky D. *Word-Sense Disambiguation Using Statistical Models of Roget's Categories*. In: Proceedings of COLING '92, Nantes, France, pages 454-460, 1992.

Yarowsky D. *One Sense Per Collocation*. In: Proceedings of the ARPA Human Language Technology Workshop, Morgan Kaufman, San Francisco, CA, 1993.

Yarowsky D. *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods*. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, MA, pp. 189-196, 1995.