Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval

(Version of September 5, 2002,

submitted to the International Journal of Medical Informatics)

Martin Volk, Bärbel Ripplinger[1]
Eurospider Information Technology AG
Schaffhauserstrasse 18
CH-8006 Zürich, Switzerland
volk@eurospider.com

Špela Vintar, Paul Buitelaar, Diana Raileanu, Bogdan Sacaleanu
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
{vintar, paulb, raileanu, bogdan}@dfki.de

Keywords: Natural Language Processing, Linguistics, Semantics, Information Systems, Unified Medical Language System

Abstract

We present a framework for concept-based cross-language information retrieval in the medical domain, which is under development in the MUCHMORE project. Our approach is based on using the Unified Medical Language System (UMLS) as the primary source of semantic data. Documents and queries are annotated with multiple layers of linguistic information. Linguistic processing includes part-of-speech tagging, morphological analysis, phrase recognition and the identification of medical terms and semantic relations between them.

The paper describes experiments in monolingual and cross-language document retrieval, performed on a corpus of medical abstracts. Results show that linguistic processing, especially lemmatization and compound analysis for German, is a crucial step to achieving a good baseline performance. On the other hand they show that semantic information, specifically the combined use of concepts and relations, increases the performance in monolingual and cross-language retrieval.

[1] Bärbel Ripplinger is now at oiz Stadt Zürich and can be contacted via baerbel.ripplinger@oiz.stzh.ch.

1 Introduction

The task of finding relevant information from large, multilingual and domain-specific text collections is a field of active research within the information retrieval and natural language processing communities [19]. Methods of Cross-Language Information Retrieval (CLIR) are typically divided into: approaches based on bilingual dictionary look-up or Machine Translation (MT); corpus-based approaches utilizing a range of IR-specific statistical measures; and concept-driven approaches, which exploit semantic information (thesauri) to bridge the gap between surface linguistic form and meaning. The latter seem particularly appropriate for domains (and languages) for which extensive, multilingual, semantic resources are available, such as UMLS (Unified Medical Language System) in the medical domain.

The experiments reported in this paper were performed within the framework of the MUCHMORE project[2], which aims at systematically comparing concept-based and corpus-based methods in cross-language medical information retrieval. Primary goals of the project are:

1. Developing and evaluating methods for the effective use of multilingual thesauri in the semantic annotation of English and German medical texts;
2. Subsequently evaluating and comparing the impact of such semantic information for the purpose of Cross-Language Information Retrieval.

The paper is organized as follows. We first give an overview of related research in concept-based cross-language information retrieval. In section 3 we present our resources and approaches for linguistic and semantic annotation. In section 4 we describe the retrieval experiments with different indexing features.

2 Related Work

Many authors have experimented with machine translation or dictionary look-up for CLIR (see [11] or [12]). In a comparison of such methods in both query and document translation, Oard [18] found that dictionary-based query translation seems to work best for short queries while for long queries machine translation of the queries performs better than dictionary look-up. However, machine translation of the documents outperforms all other methods with long queries. An important problem in the translation of short queries is the lack of context for diambiguation of words that have more than one meaning and therefore may correspond to more than one translation [13] [24]. Therefore, in the case of short queries mostly all translations are considered instead of trying to disambiguate between them. Alternatively, the user is asked to select the appropriate translation in an interactive setting [7].

Ambiguity is also of importance to interlingua approaches to CLIR that use multilingual thesauri as resources for a language-independent (semantic) representation of both queries and documents. Domain-specific multilingual thesauri have been used for English-German CLIR within social science [9], while [6] describes the use of the UMLS MetaThesaurus for French and Spanish queries on the OHSUMED text collection, a subset of MEDLINE. Both of these approaches use the thesaurus as a source for compiling a bilingual lexicon, which is then used for query translation. A different use of multilingual thesauri is in combination with document classification techniques, such as Latent Semantic Indexing [14] and the Generalized Vector Space Model [5], both of which depend on available parallel

[2] MUCHMORE is sponsored by the European Union under grant IST-1999-11438. The European project partners are DFKI GmbH, LT Department, Saarbrücken; XEROX Research Centre Europe, Grenoble; ZInfo, Klinikum der J.W. Goethe-Universität Frankfurt; and Eurospider Information Technology AG, Zurich. The project also includes two partners in the US: Carnegie Mellon University, LT Institute, and Stanford University, CSLI. For details see http://muchmore.dfki.de.

corpora. Finally, next to domain-specific thesauri also more general semantic resources such as EuroWordNet [30] have been used in both mono- and cross-language information retrieval [10].

The work we describe here is also primarily an interlingua approach to CLIR in the medical domain, in which we use both domain-specific (UMLS) and general language semantic resources (EuroWordNet). Central to the approach is the use of linguistic processing that includes part-of-speech tagging, morphological analysis (including compound analysis for German), and phrase recognition for an accurate semantic annotation of relevant terms and relations in both the queries and the documents. Specifically for morphologically rich languages such as German it is important to extend linguistic processing beyond primitive stemming.

We strictly apply semantic annotation in a monolingual way, in which information available from parallel documents is not considered. This is to ensure that the approach will be applicable to any multilingual document collection (for our purposes here in English and German) and not only to parallel document collections.

3 Corpus Processing and Annotation

3.1 Linguistic Processing

The main document collection used in the MUCHMORE project is a parallel corpus of English-German scientific medical abstracts obtained from the Springer web site[3]. The corpus consists of approximately 9000 documents with a total of one million tokens for each language. Abstracts are taken from 41 medical journals (e.g. *Der Nervenarzt, Der Radiologe,* etc.), each of which constitutes a homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.). Corpus preparation included removing HTML-tags, English segments from German abstracts and vice versa, and names of authors, addresses, etc. It also included removing or converting special symbols and other non-ASCII elements in order to produce a clean, plain text version of each abstract, consisting of a title, text and keywords (if available). The corpus was then linguistically annotated using ShProT, a shallow processing tool that consists of four integrated components: the SPPC tokenizer [21], TnT [1] for part-of-speech tagging, Mmorph [20] for morphological analysis and Chunkie [28] for phrase recognition.

3.1.1 Part-of-Speech Tagging

TnT is a statistical part-of-speech tagger based on a Hidden Markov Model and trained on general language corpora (the NEGRA[4] corpus for German, the SUSANNE[5] corpus for English). In order to perform in an optimal way, we adapted it to the medical domain. Two approaches were considered: First, retraining the tagger on an annotated domain-specific corpus, or second, an update of its underlying lexicon. As medical corpora with manually controlled part-of-speech information are difficult to obtain, we decided to adapt the lexicon with information from the UMLS English and German Specialist Lexicons (German lexicon constructed by and available through ZInfo). Because of a similar syntax for general language and the medical language used in our corpus of scientific abstracts, we obtained good results without retraining.

3.1.2 Morphological Analysis

Morphological analysis is based on a full-form lexicon generated by Mmorph. For each token we look for a matching entry that will provide its morphological information. If no valid word form is found, the token is analyzed as a potential compound and decomposition is attempted. For example, the German compound *Hausstaubmilbenallergiker* (engl. *house*

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

[4] http://www.coli.uni-sb.de/sfb378/negra-corpus/

[5] http://www.cogs.susx.ac.uk/users/geoffs/Rsue.html

dust mite allergy patient) is segmented into four components *Haus + Staub + Milbe + Allergiker*.

Initial lemmatization experiments produced poor results on our corpus, particularly in the analysis of medical compounds for German and English. We therefore decided to update the existing lexicon in two steps. First, we updated it with additional medical entries for both German and English. This allowed us to avoid the incorrect decomposition of the following words:

medical term	incorrect segmentation
zoonoses	zoo + nose
Epicillin	epic + ill + in
endoral	end + oral
trypan	try + pan

In a second step we removed general-language word forms that function as prefixes in the medical domain (e.g. *auto, post, radio*), in order to avoid the incorrect decomposition of words such as:

medical term	incorrect segmentation
autoimmune	auto + immune
postinflammatory	post + inflammatory
radiogram	radio + gram

Another bootstrapping approach, yet to be explored, uses the nouns in our corpus as segmentation elements for compounds. For example, we can segment *Pertussisantigene* into *Pertussis* and *Antigene* since these two words occur stand-alone in the corpus. And we can then lemmatize the plural form *Antigene* into *Antigen*.

3.1.3 Chunking

The UMLS Metathesaurus provides an extensive inventory of technical terms for the medical domain, but we also extract novel terms from our corpus. As technical terms are mainly noun phrases, the recognition of these syntactic structures, also called chunks, is a necessary task. Our tool in this process, Chunkie, is a Hidden Markov Model-based partial parser that goes beyond simple bracketing and is capable of recognizing not only the boundaries, but also the internal structure of simple as well as (most) complex noun phrases, prepositional phrases and adjective phrases. Similar to the TnT tagger, the performance of Chunkie will improve by adaptation to the medical domain. The only possible approach to this is retraining it on a domain-specific collection of syntactically annotated sentences (often called a treebank). However, as we are not aware of any existing treebank in the medical domain, we decided to use Chunkie as is, trained on general language data.

3.2 Semantic Annotation using EuroWordNet

In addition to annotation with UMLS, terms are annotated with EuroWordNet senses [30] to compare domain-specific and general language use. EuroWordNet is a multilingual database for several European languages and is structured similar to the Princeton WordNet [17]. Each language-specific (Euro)WordNet is linked to all others through the so-called Inter-Lingual-Index, which is based on WordNet1.5. The languages are interconnected via this index, 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 components 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 all the synsets that contain it (see example in section 3.4). Disambiguation will reduce these to a single sense if possible. A term can be simple (e.g. *man*) or complex (e.g. *rock and roll*). Because meanings between languages cannot be mapped one-on-one, more than one synset within a language may be mapped to the same concept in the Inter-Lingual-Index. In order to distinguish between these, every synset was given a unique identifier (ID), called offset[6], as shown in the following example:

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

If an English query contains *knuckle*, this will be mapped to offset *3824895*, allowing us to retrieve German documents containing terms mapped to the same offset, like *Knöchel*. The same applies for a German query to English documents.

3.3 Semantic Annotation using UMLS and MeSH

The essential part of any concept-based CLIR system is the identification of terms and their mapping to a language-independent conceptual level. Our basic resource for semantic annotation is UMLS, which is organized in three parts.

The **Specialist Lexicon** provides lexical information: a listing of word forms and their lemmas, part-of-speech and morphological information.

Second, the **Metathesaurus** is the core vocabulary component, which unites several medical thesauri and classifications into a complex database of concepts covering terms from 9 languages. Each term is assigned a unique string identifier, which is then mapped to a unique concept identifier (CUI). An entry for *HIV pneumonia* in the Metathesaurus main termbank (MRCON) looks like this:

```
C0744975 | ENG | P | L1392183 | PF | S1657928 | HIV pneumonia | 3|
```

The fields in this entry specify (from left to right), the concept identifier, the language of the term, the term status, the term identifier, the string type, the string identifier, the string itself, and a restriction level.

In addition to the mapping of terms to concepts, the Metathesaurus organizes concepts into a hierarchy by specifying relations between concepts. These are thesaurus-type generic relations like *broader_than, narrower_than, parent, sibling* etc. Another component of the Metathesaurus provides information about the sources and contexts of the concepts. The UMLS 2001 version includes 1.7 million terms mapped to 797,359 concepts, of which 1.4 million entries are English and only 66,381 German. Only the MeSH (Medical Subject Heading) part of the Metathesaurus covers both German and English, therefore we only use MeSH terms (564,011 term entries for English and 49,256 for German) for corpus annotation.

The third part is the **Semantic Network**, which provides a grouping of concepts according to their meaning into 134 semantic types (TUI). The concept above would be assigned to the class *T047, Disease or Syndrome*. The Semantic Network then specifies potential relations between those semantic types. There are 54 hierarchically organized domain-specific relations, such as *affects, causes, location_of* etc.

[6] In our case only for German, as the English synsets correspond to the Inter-Lingual-Index directly.

3.3.1 Terms and Concepts

At the level of terms, we used the following semantic information in the annotation of our documents and queries:

- Concept Unique Identifier (CUI)
- Type Unique Identifier (TUI)
- Medical Subject Headings ID - an alternative code to the CUIs
- Preferred Term - a term that is marked as the preferred name for a particular concept

An example of an annotated term in our XML annotation scheme is presented in section 3.4.

The identification of UMLS terms in the documents was based on morphological processing of both the term bank and the document, so that term lemmas were matched rather than word forms. The preparation of the term bank included filtering and normalization procedures, such as case folding, removal of long terms, inversion of term variants with commas (*Virus, Human Immunodeficiency* → *Human Immunodeficiency Virus*), conversion of special characters etc. The annotation tool matches terms of lengths 1 to 3 tokens, based on lemmas if available and word forms otherwise. Term matching on the sub-token level is also implemented to ensure the identification of terms that are a part of a more complex compound, which is crucial for German.

The decision to use MeSH codes in addition to concept identifiers (CUIs) was based on our observation that the UMLS Semantic Network, especially the semantic types and relations, does not always adequately represent the domain-specific relationships. MeSH codes on the other hand have a transparent structure, from which both the semantic class of a concept and its depth in the hierarchy can be inferred. For example, the terms *infarction* (C23.550.717.489) and *myocardial infarction* (C14.907.553.470.500) both belong to the group of diseases, but the node of the first term lies higher in the hierarchy as its code has fewer fields.

There are several possible levels of **ambiguity** for terms and concepts: a single term may be assigned several concept identifiers, and a single concept identifier may be mapped to several MeSH codes. Since UMLS is designed to become the unified ontological resource for the medical domain, the MeSH hierarchy is viewed as a subset of the UMLS Metathesaurus. This relationship is reflected in our annotation scheme in the following way: We treat terms with several concept identifiers as ambiguous readings, and therefore annotate each reading as a separate element with its corresponding information.

However, if a concept identifier can be mapped to several MeSH codes, we annotate those as possible mappings subordinate to the concept identifiers. Since retrieval experiments show better results for MeSH codes than for conceptual identifiers (CUIs) (cf. section 4.2), and furthermore the transparency of MeSH codes facilitates the discovery of new semantic relations, we are considering to reorganize this structure in the future and use MeSH codes as primary conceptual nodes in our scheme.

Another ambiguity occurs on the level of semantic types. Thus, for example, the term *Type I Collagen* with concept identifier C0041455 can have the semantic type T116 or T123, meaning *Amino Acid, Peptide or Protein* or *Biologically Active Substance* respectively. But since in this case we are dealing with a single concept, which can be viewed from different perspectives depending on the context, we do not consider multiple semantic types to represent real ambiguities and thus do not treat them as different readings of a term.

It should be noted that semantic annotation is purely monolingual. This means that the annotation of a document in one language is based only on this document. The parallel document in the other language which could be used to complement or to compare the semantic annotation is ignored. This was done to ensure that the evaluation results are applicable to any document collection and not only to parallel document collections.

3.3.2 Semantic relations

Semantic relations are annotated on the basis of the UMLS Semantic Network, which defines binary relations between semantic types (TUIs) in the form of triplets, for example *T195 - T151 - T042* meaning *Antibiotic - affects - Organ or Tissue Function.* We search for all pairs of semantic types that co-occur within a sentence, which means that we can only annotate relations between items that were previously identified as UMLS terms. According to the Semantic Network relations can be ambiguous, meaning that two concepts may be related in several ways. For example:

```
Diagnostic Procedure | analyzes            | Antibiotic
Diagnostic Procedure | assesses_effect_of  | Antibiotic
Diagnostic Procedure | measures            | Antibiotic
Diagnostic Procedure | uses                | Antibiotic
```

Since the semantic types are rather general (e.g. *Pharmacological Substance, Patient or Group*), the relations are often found to be vague or even incorrect when they are mapped to a document. Given the ambiguity of relations and their generic nature, the number of potential relations found in a sentence can be high, which makes their usefulness questionable. A manual evaluation of automatic relation tagging by medical experts showed that only about 17% of relations were correct, of which only 38% were perceived as significant in the context of information retrieval. On the other hand, low term coverage - particularly for German - severely limits the number of relations that we can identify in the described way. Retrieval experiments performed with German queries over English documents showed that an evaluation of semantic relations in this context is almost impossible due to the fact that few sample queries contain more than one concept identifier, and consequently very few semantic relations can be established.

For the above reasons, the UMLS-based annotation of semantic relations needed to be improved and extended in several ways. The first task was to tackle relation ambiguity, i.e. to select correct and significant relations from the ones proposed by automatic UMLS lookup; a procedure we refer to as *relation filtering.* We select relations on the basis of two hypotheses:

1. Semantic relations are expressed via lexical markers, such as verbs.
2. Significant relations occur between significant concepts, i.e. concepts with a high inverse document frequency (IDF).

The implementation of the first hypothesis is a filtering method, based on a co-occurrence matrix of verbs and UMLS semantic relations. For each verb we compute a list of the top five most probable relations, whereupon relations not accompanied by any of the significant verbs are filtered out. This method reduces the number of ambiguous relations by 46%. The second filtering step selects relations on the basis of the significance of the related concepts, where significance is measured by IDF (Inverse Document Frequency). We decided to use IDF instead of the generally used TF-IDF, because term frequency (TF), if multiplied with IDF, will assign a higher score to frequent terms like *patient, therapy, disease.* However in our context these are precisely the terms we wish to rank lowest. Using an experimentally derived threshold, the IDF filtering method brings a further reduction of relations to 31 percent of the originally proposed ones. A comparison of these results with the set of manually filtered documents shows a high level of correspondence, which confirms the validity of both hypotheses.

Work is underway to tackle the second problem, namely low term coverage as well as inadequacy of the UMLS Semantic Network for retrieval purposes. We are exploring methods of discovering new relations between concepts, whereby the task is both finding new instances of known relations and identifying candidates of novel relations. As explained above, MeSH codes seem a more accurate representation of concepts than concept identifiers. Furthermore MeSH codes can be abstracted to a desired level of specificity. By computing co-occurrences of MeSH codes and filtering these according to a log-likelihood ratio, we assume that each frequent pair of MeSH codes indicates one or more relations. Contextual information can help us extract them, whereby we currently use verb and preposition frequencies as main attributes of each MeSH pair. Clustering those instances will assign MeSH pairs with similar context features to the same class, whereupon we need to find out which classes correspond to known UMLS relations and which indicate novel relations.

3.4 The XML Annotation Format

Both morpho-syntactic (part-of-speech, morphology, phrases) and semantic (terms, semantic relations) annotation are integrated in a multi-layered XML annotation format, which organizes various levels as separate tracks with options of reference between them via indices. The aim was to design an annotation format that would include all layers and adequately represent relationships between them, while at the same time remaining logical and readable, efficient for parsing and indexing as well as flexible for future additions and adjustments [29].

We will explain the annotation format with the following example sentence from an abstract in the field of psychiatry.

> *Balint syndrom is a combination of symptoms including simultanagnosia, a disorder of spatial and object-based attention, disturbed spatial perception and representation, and optic ataxia resulting from bilateral parieto-occipital lesions.*

Each document is split into sentences and the XML annotation is based on them. Each `<sentence>` contains a `<text>` block that holds the tokens as XML content, and both lemma and part-of-speech information as XML attributes.

```
<text>
  <token id="w1"  pos="NN">                         Balint         </token>
  <token id="w2"  pos="NN">                         syndrom        </token>
  <token id="w3"  pos="VBZ" lemma="be">             is             </token>
  <token id="w4"  pos="DT"  lemma="a">              a              </token>
  <token id="w5"  pos="NN"  lemma="combination">    combination    </token>
  <token id="w6"  pos="IN"  lemma="of">             of             </token>
  <token id="w7"  pos="NNS" lemma="symptom">        symptoms       </token>
  ...
  <token id="w20" pos="JJ"  lemma="spatial">        spatial        </token>
  <token id="w21" pos="NN"  lemma="perception">     perception     </token>
  <token id="w22" pos="CC"  lemma="and">            and            </token>
  <token id="w23" pos="NN"  lemma="representation"> representation </token>
  ...
</text>
```

The linguistic analyzer determines noun phrases, adjective phrases and prepositional phrases. In this example it determines - among others - a noun phrase (NP) for words `w1` and `w2` *Balint syndrom* and a more complex noun phrase from `w20` to `w23` *spatial perception and representation.*

```
 <chunk id="c1" from="w1"  to="w2" type="NP"/>
 <chunk id="c7" from="w20" to="w23" type="NP"/>
```

In addition each `<sentence>` contains semantic annotations. In a first block we store pointers to EuroWordNet (EWN) synsets. For the example sentence we determined that word `w21`, *perception*, has four EWN senses, related to *perceiving - sensing, perception*, and *perceptual experience*. We are currently working on a word sense disambiguation module to cut down on ambiguities concerning EWN senses, based on methods described in [3] and [4]. Evaluation of the disambiguation module is undertaken as part of the CLIR evaluation task (comparing disambiguated and non-disambiguated versions of the annotated document collection), as well as separately by using a manually tagged lexical sample corpus [23].

```
<ewnterm id="e5" from="w21" to="w21">
  <sense offset="487490"/>
  <sense offset="3890199"/>
  <sense offset="3955418"/>
  <sense offset="4002483"/>
</ewnterm>
```

At the core of our semantic annotation are UMLS terms and MeSH codes. For the example sentence the words `w20` and `w21` point to the concept with a preferred name “Space Perception”, which corresponds to the CUI code C0037744 and TUI code T041 (i.e. Mental Process). In addition this concept is linked to two MeSH codes which stand for two positions of the term “Space Perception” in the MeSH tree of concepts, the first under the node “Perception” and the second under “Visual Perception”. And word `w26` *optic* triggered the concept “Optics” (with one corresponding MeSH code).

```
<umlsterm id="t7" from="w20" to="w21">
  <concept id="t7.1" cui="C0037744" preferred="Space Perception" tui="T041">
    <msh code="F2.463.593.778"/>
    <msh code="F2.463.593.932.869"/>
  </concept>
</umlsterm>

<umlsterm id="t8" from="w26" to="w26">
  <concept id="t8.1" cui="C0029144" preferred="Optics" tui="T090">
    <msh code="H1.671.606"/>
  </concept>
</umlsterm>
```

The most specific of our semantic information are the semantic relations that we derive from the UMLS Semantic Network. This network indicates that“Space Perception” is an issue in “Optics” which is coded in the following manner. Note that the XML attributes `term1` and `term2` point to the UMLS concepts introduced in the example above.

```
<semrel id="r7" term1="t7.1" term2="t8.1" reltype="issue_in"/>
```

4 Evaluation in Information Retrieval

4.1 The Set of Queries and Human Relevance Assessments

In order to evaluate whether the semantic annotations result in a performance gain in information retrieval, several experiments have been carried out. We used our own document collection (the set of medical abstracts described above) as well as a query set defined by medical experts. The OSHUMED collection would not have been appropriate for the MUCHMORE project due to its monolingual nature (documents and queries are only available in English).

For the experiments, we used the relevance assessments based on 25 queries provided by the medical expert in the MUCHMORE project. We obtained relevance assessments based on the German documents as well as based on the English documents from two teams of experts. However, the number of documents classified as relevant for the two collections were quite different, 959 for German and 500 for English. The main reason for this discrepancy is the different types of experts doing the assessments (medical experts for German, and students for English). Because the MUCHMORE corpus is parallel, we decided to use the German relevance assessments for all our experiments in order to get comparable data. In these assessments the number of relevant documents per query varies between 7 and 104.

The queries are short and usually consist of a complex noun phrase extended by attributes (including prepositional phrases) and coordination. Here are two examples.

- *Arthroskopische Behandlung bei Kreuzbandverletzungen.*
  Arthroscopic treatment of cruciate ligament injuries.
- *Indikation für einen implantierbaren Kardioverter-Defibrillator (ICD).*
  Indication for implantable cardioverter defibrillator (ICD).

4.2 Monolingual Evaluation Runs

MUCHMORE aims first and foremost at cross-language retrieval (CLIR). In order to assess the CLIR performance, monolingual experiments in German and English were conducted acting as baselines for the cross-language experiments. In the monolingual experiments the queries and the documents are of the same language. Most CLIR systems achieve only up to 75% precision compared to monolingual IR (cf. [25]), and one goal of MUCHMORE is to check whether semantic annotation improves CLIR performance.

For the retrieval experiments we used the commercial *relevancy* information retrieval system from Eurospider Information Technology AG. In regular deployment this system extracts word tokens from documents and queries and indexes them using a straight *lnu.ltn* weighting scheme (for the theoretical background of this scheme see [27] or [26]). In addition the system can index word stems derived from a lexicon-based (Celex) stemmer for German, and a Porter-like stemmer for English [31].

For the MUCHMORE evaluation runs we adapted the *relevancy* system so that it indexes the information provided by the XML annotated documents and queries: word forms (tokens) and their base forms (lemmas) for all indexable parts-of-speech both for German and English. The indexable parts-of-speech encompass all content words, i.e. nouns (including proper names and foreign expressions), adjectives, and verbs (excluding auxiliary verbs). All semantic information was indexed in separate categories each: EuroWordNet terms, UMLS terms, semantic relations, and MeSH terms.

For each language, we produced a baseline performance by indexing only the tokens in both the documents and the queries. We call these baselines DE-token and EN-token. Some recent works [15], [16] have shown, that at least for German a linguistic-based stemming and decompounding is beneficial for retrieval, and therefore two evaluation runs based on linguistic stemming were produced, which we termed DE-token-lemma and EN-token-lemma. In table 1 we present the results of the monolingual German retrieval experiments.

In this table and all subsequent tables we present the retrieval results in four columns. The first column contains the overall performance, measured as mean average precision (mAvP) as has become customary in the TREC experiments (cf. [8]). This figure is computed as the mean of the precision scores after each relevant document retrieved. The value for the complete evaluation run (i.e. the set of all queries) is the mean over all the individual mean average precision scores. This value contains both precision and recall oriented aspects and is the most commonly used summary measure. In the second column we present the absolute number of relevant documents retrieved, a pure recall measure. Third, we present the average precision at 0.1 recall (AvP01). According to Eichmann et al. [6], the effectiveness within the high precision area is measured assuming that users are most interested to get relevant documents ranked topmost in the result list. Because this number can vary substantially for different queries, we consider also the precision figures for the topmost documents retrieved (in column four). There we focus on the top 10 documents (P10).

In the baseline experiment for German (DE-token) we find only 322 relevant documents (out of 956; cf. table 1). The mean average precision is thus low (mAvP = 0.16), but the average precision in the top ranks is acceptable (AvP = 0.56). So, the few documents that are found are often ranked at the top of the list. On average there are 4.16 relevant documents among the 10 top ranked documents (P10).

The importance of good linguistic stemming and decompounding is shown by the second experiment (DE-token-lemma), which achieves a recall gain of 60% compared to DE-token. In parallel, the precision figures have improved substantially. Lemmatization was done with a general-purpose morphological analyzer (as described in section 3.1.2). As a side issue we have also explored another linguistic lemmatizer that was especially tuned to medical terms. Using that lemmatizer resulted in an even bigger difference to the DE-token baseline than reported here. The additional benefit was particularly due to better compounding of word forms that contain specific medical morphemes.

The impact of the different types of semantic information was determined one by one, but always in combination with tokens and lemmas. We wanted to support the hypothesis that semantic information will improve the precision over pure token and lemma information. It turns out that the MeSH codes are the most useful indexing features whereas the EuroWordNet terms (EWN), without disambiguation in our current experiments (!), are the worst. Using MeSH codes slightly increases recall (from 516 to 526) but most impressively improves average precision (from 0.2180 to 0.2452). The positive impact of the UMLS terms is less visible and - as was to be expected - the very specific semantic relations (Semrel) have hardly any impact. Using the EuroWordNet terms in this combination with lemmas and tokens degrades the overall performance.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE-token	0.1600	322	0.5622	0.4160
DE-token-lemma	0.2180	516	0.5967	0.4720
DE-token-lemma-EWN	0.1980	500	0.5571	0.4520
DE-token-lemma-UMLS	0.2236	509	0.5895	0.4640
DE-token-lemma-MeSH	0.2452	526	0.6356	0.5120
DE-token-lemma-Semrel	0.2224	516	0.5841	0.4640

Table 1: Results of the monolingual German runs

Table 2 shows the results of the monolingual English evaluation runs. The difference between EN-token and EN-token-lemma is surprisingly small. This is due to the fact that English has fewer inflected forms and hardly any noun compounding. Interestingly, using English lemmas decreases precision significantly.

The performance level for English monolingual retrieval is significantly higher than for German. But when we add semantic indexing features in English, the general tendency is similar to German. MeSH leads to the best results both in recall and precision, UMLS is second best, and the semantic relations have almost no impact. The use of EuroWordNet terms as it stands without word sense disambiguation has a strong negative influence on the retrieval results.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
EN-token	0.3455	617	0.8077	0.6160
EN-token-lemma	0.3320	635	0.7543	0.5760
EN-token-lemma-EWN	0.2565	616	0.6025	0.4640
EN-token-lemma-UMLS	0.3415	641	0.7516	0.5840
EN-token-lemma-MeSH	0.3543	648	0.7748	0.6000
EN-token-lemma-Semrel	0.3272	637	0.7279	0.5520

Table 2: Results of the monolingual English runs

4.3 Cross-Language Evaluation Runs

The cheapest way of Cross-Language Information Retrieval is monolingual retrieval over a parallel corpus. This means that we would search German documents with a German query and simply display those English documents that are known to be correspondences of the found German documents. This is not what we do here. Instead, we assume that we have a document collection (i.e. a corpus) in one language and a query in another language.

For the cross-language evaluation runs we used German queries to retrieve English documents. These results should not only be compared to the monolingual runs but also different approaches should be evaluated.

A rough baseline for the cross-language task is using the tokens of the German queries directly for retrieval of the English documents. The idea is that the overlap in technical vocabulary between these languages will lead to relevant documents. And indeed, this approach finds 66 relevant documents (cf. DE2EN-DE-token in table 3). The best queries were those with the acronym *HIV* (which is the same in German and English) and with the Latin expression *diabetes mellitus*. For both these queries more than half the relevant documents were retrieved.

It might be surprising that the overlap in technical vocabulary does not carry further than merely 66 documents. But one must consider that often the roots of the words are identical but the forms do not match because of differences in spelling and inflection (e.g. *arthroskopische* vs. *arthroscopic*). Stemming combined with some letter normalization (e.g. $k = c = z$) would lead to an increased recall, but has not been explored here.

As a second baseline we investigated the use of Machine Translation (MT) for translating the queries. We employed the PC-based system PersonalTranslator 2001 (PT2001; linguatec, Munich) to automatically translate all queries from German to English. PersonalTranslator allows to restrict the subject domain of the translation and we selected the domains medicine and chemistry. This restriction helps the system to choose the subject-specific interpretation if multiple interpretations for a given lexical entry are available.

Although the system contains medical vocabulary, many words from our queries are not in its lexicon and remain untranslated (see the first example query below). Unfortunately the system does not segment compounds if it lacks knowledge of some of their parts. Therefore the word *Myokardinfarkts* is not segmented although *Infarkt* is in the system's lexicon and could have been translated. Other queries are fully translated and almost perfect (see the second example query).

- *DE: Behandlung des akuten Myokardinfarkts.*
  PT2001: Treatment of the acute Myokardinfarkts.
  EN: Treatment of acute myocardial infarction.

- *DE: Möglichkeiten der Korrektur von Deformitäten in der Orthopädie.*
  PT2001: Possibilities of the correction of deformities in orthopedics.
  EN: Approach of the correction of deformities in orthopedics.

Many translations are incomplete or incorrect but still the automatically translated queries scored well with regard to recall. In table 3, line DE2EN-MT-PT2001, we see that these queries lead to 376 relevant documents at a (rather low) mean average precision of 0.1184.

Now let us compare these results with the semantic codes annotated in our corpus and queries. This means we are using the semantic annotation of the German queries to match the semantic annotation of the English documents. One could say that we are now using the semantic annotation as an interlingua or intermediate representation to bridge the gap between German and English.

The third block in table 3 has all the results. This time the UMLS terms lead to the best results with respect to recall, but MeSH is (slightly) superior regarding precision. EuroWordNet leads to the worst precision and the semantic relations have only a minor impact due to their specificity. If we combine all semantic information, we achieve the best recall (404) and mean average precision (0.1774). This clearly outperforms machine translation.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE2EN-DE-token	0.0512	66	0.1530	0.1160
DE2EN-MT-PT2001	0.1184	376	0.3382	0.2520
DE2EN-EWN	0.0090	111	0.0311	0.0160
DE2EN-UMLS	0.1620	366	0.3724	0.2800
DE2EN-MeSH	0.1699	304	0.3888	0.2600
DE2EN-Semrel	0.0229	23	0.0657	0.0480
DE2EN-all-combined	0.1774	404	0.3872	0.2720
DE2EN-SimThes	0.2290	409	0.4492	0.3640
DE2EN-SimThes+all-comb.	0.2955	518	0.5761	0.4600

Table 3: Results of the cross-language runs: German queries and English documents

For the last two experiments we have built a similarity thesaurus (SimThes) over the parallel corpus. The similarity thesaurus contains words (adjectives, nouns, verbs) from our corpus, each accompanied by a set of words that appear in similar contexts and are thus similar in meaning. A similarity thesaurus can be built over a monolingual corpus. It may then serve for query expansion in monolingual retrieval. In our case we built the similarity thesaurus over the parallel corpus. We were interested in German words and their similar counterparts in English. The similarity thesaurus is thus a bilingual lexicon with a broad translation set (in our case 10 similar English words per German word). For example, for the German word *Myokardinfarkt* the similarity thesaurus contains the following 10 words in decreasing degrees of similarity:

Similarity Thesaurus: *infarction, acute myocardial infarction, myocardial, thrombolytic, acute, thrombolysis, crs, synchronisation, cardiogenic shock, ptca*

We used these words for cross-language retrieval. Each German word from the queries was substituted by the words of its similarity set. This resulted in a recall of 409 relevant documents found and a relatively good mean average precision of 0.2290 (see DE2EN-SimThes in table 3). Note that unlike in our previous experiments, we have now exploited the parallelism of the documents in our corpus for the construction of the similarity thesaurus. The bilingual similarity thesaurus is only available if we have a parallel or comparable corpus (cf. [2]) whereas the semantic annotations will also be applicable for a monolingual document collection.

Finally we checked the combination of all semantic annotations with the similarity thesaurus. Each query is now represented by its EuroWordNet, UMLS, MeSH and semantic relations codes as well as by the words from the similarity thesaurus. This combination leads to the best results for CLIR. We retrieved 518 relevant documents with a mean average precision of 0.2955 (cf. the last line DE2EN-SimThes+all-combined in table 3). And the figures for the high precision area (AvP and P10) are also outstanding. Note that this result is comparable to German monolingual retrieval with tokens, lemmas and semantic annotation (cf. table 1).

5 Conclusions

We have explored the use of different kinds of semantic annotation for both monolingual and cross-language retrieval.

In monolingual retrieval (for both English and German) semantic information from the MeSH codes (Medical Subject Headings) were most reliable and resulted in an increase in recall and precision over token and lemma indexing. Moreover, the monolingual experiments show that high-quality linguistic analysis is crucial for a good retrieval performance, which indicates that further work is needed to improve the compatibility and quality of morphological analysis both on the side of document and query processing and indexing. This is a prerequisite for a good baseline.

In cross-language retrieval the combination of all semantic information outperformed machine translation. It was only superseded by the use of a similarity thesaurus built over the parallel corpus. The highest overall performance resulted from a combination of the similarity thesaurus with the semantic information. This result was comparable to the German monolingual retrieval results.

If we compare the monolingual and cross-language retrieval results, it is striking that the best semantic sources in the monolingual experiments were also the best in the cross-language task. This indicates that monolingual results for semantic annotations can be extrapolated to cross-language retrieval if no cross-language test set is available.

So far, semantic annotation in our approach was based on the use of existing resources (UMLS and EuroWordNet) without applying disambiguation. In future work we hope to further improve performance by the integration of disambiguation for UMLS and EuroWordNet terms as well as including novel extracted terms for EuroWordNet and extracting more relevant novel relations for UMLS.

References

[1] Brants T. 2000. TnT - A Statistical Part-of-Speech Tagger. In: *Proc. of the 6th ANLP Conference*, Seattle, WA.

[2] Braschler, M. and P. Schäuble. 2000. Using Corpus-Based Approaches in a System for Multilingual Information Retrieval. *Information Retrieval*, 3, 273-284.

[3] Buitelaar, P. and B. Sacaleanu. 2001. Ranking and Selecting Synsets by Domain Relevance. In: *Proc. of NAACL WordNet Workshop.*

[4] Buitelaar P., J. Alexandersson, T. Jaeger, S. Lesch, N. Pfleger, D. Raileanu, T. von den Berg, K. Klöckner, H. Neis, and H. Schlarb. 2001. An Unsupervised Semantic Tagger Applied to German. In: *Proc. of Recent Advances in NLP (RANLP)* , Tzigov Chark, Bulgaria, 5-7 September.

[5] Carbonell J., Y. Yang, R. Frederking, R. D. Brown, Y. Geng, and D. Lee. 1997. Translingual Information Retrieval: A Comparative Evaluation. In: *Proc. of the Fifteenth International Joint Conference on Artificial Intelligence.*

[6] Eichmann D., M. Ruiz, and P. Srinivasan. 1998. Cross-Language Information Retrieval with the UMLS Metathesaurus. In: *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia.

[7] Erbach G., G. Neumann, and H. Uszkoreit. 1997. MULINEX Multilingual Indexing Navigation and Editing Extensions for the World-Wide Web, AAAI Symposium on Cross-Language Text and Speech: American Association for Artificial Intelligence. 1997. pp. 22-28. ISBN: 1-57735-040-5; Technical Report: SS-97-05. http://www.clis.umd.edu/dlrg/filter/sss/papers/

[8] Gaussier E., G. Grefenstette , D. A. Hull, and B. M. Schulze. 1998. Xerox TREC-6 site report: Cross language text retrieval. In: *Proc. of the Sixth TExt Retrieval Conference (TREC-6).* National Institute of Standards Technology (NIST), Gaithersburg, MD.

[9] Gey F. C., and H. Jiang. 1999. English-German Cross-Language Retrieval for the GIRT Collection - Exploiting a Multilingual Thesaurus. In: *Proc. of the Eighth Text REtrieval Conference (TREC-8)*, National Institute of Standards Technology (NIST), Gaithersburg, MD.

[10] Gonzalo J., F. Verdejo, and I. Chugur. 1999. Using EuroWordNet in a Concept-based Approach to Cross-Language Text Retrieval, *Applied Artificial Intelligence:13*, 1999.

[11] Hull D. A., and G. Grefenstette. 1996. Querying Across Languages: A Dictionary based Approach to Multilingual Information Retrieval. In: *Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: ACM SIGIR.* 1996. http://www.xerox.fr/people/grenoble/hull/papers/sigir96.ps

[12] Kraaij, W. and D. Hiemstra. 1998. TREC6 Working Notes: Baseline Tests for Cross Language Retrieval with the Twenty-One System. In: *TREC6 working notes.* National Institute of Standards and Technology (NIST), Gaithersburg, MD.

[13] Krovetz, R. and B. Croft. 1992. Lexical Ambiguity and Information Retrieval. In: *ACM Transactions on Information Systems*, Vol. 10, No. 2, pp. 115-141.

[14] Landauer T. K., and M. L. Littman. 1990. Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing. In: *Proc. of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research.* Waterloo, Ontario: UW Centre for the New OED and Text Research.1990. pp. 31–38. http://www.cs.duke.edu/ mlittman/docs/x-lang.ps

[15] Monz, C., and M. de Rijke. 2001. The University of Amsterdam at CLEF2001. *CLEF 2001 Working Notes.*

[16] Moulinier, J. A. McCulloh, and E. Lund. 2000. West Group at CLEF 2000: Non-English Monolingual Retrieval. In *Cross-Language Information Retrieval and Evaluation*, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Springer Verlag.

[17] Miller G.A., R. Beckwith, C. Fellbaum, D. Gross, K. Miller. 1993. *Five papers on Wordnet.* Technical report, Cognitive Science Laboratory, Princeton University, August. Revised version.

[18] Oard D. 1998. A comparative study of query and document translation for cross-lingual information retrieval In: *Proc. of AMTA*, Philadelphia, PA.

[19] Oard, D. and A. Diekema. 1998. Cross-Language Information Retrieval. Annual Review of Information Science (ARIST), Vol. 33, Martha Williams (Ed.), Information Today Inc., Medford, NJ.

[20] Petitpierre D., and G. Russell. 1995. MMORPH - The Multext Morphology Program. *Multext deliverable report for the task 2.3.1*, ISSCO, University of Geneva, Switzerland.

[21] Piskorski, J. and G. Neumann. 2000. An intelligent text extraction and navigation system. In: *Proc. of the 6th RIAO*. Paris.

[22] Qui, Y. 1995. Automatic Query Expansion Based on a Similarity Thesaurus. *PhD thesis*, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.

[23] Raileanu D., P. Buitelaar, J. Bay, and S. Vintar. 2002. Evaluation Corpora for Sense Disambiguation in the Medical Domain. In: *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

[24] Sanderson, M. 1994. Word Sense Disambiguation and Information Retrieval. In: Croft, B. and K. van Rijsbergen. (Eds.), In: *Proc. of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.* pp. 142-151. Springer-Verlag.

[25] Schäuble, P., and P. Sheridan. 1998. Cross-language information retrieval (CLIR) track overview. In: *Proc. of The Sixth Text Retrieval Conference (TREC-6).* National Institute of Standards Technology (NIST), Gaithersburg, MD.

[26] Schäuble, P. 1997. Multimedia Information Retrieval. Content-Based Information Retrieval from Large Text and Audio Databases. Kluwer Academic Publishers.

[27] Singhal A., C. Buckley, and M. Mitra. 1996. Pivoted Document Length Normalization. In: *Proc. of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21-29, Zurich, Switzerland.

[28] Skut W. and T. Brants. 1998. A Maximum Entropy partial parser for unrestricted text. In: *Proc. of the 6th ACL Workshop on Very Large Corpora (WVLC)*, Montreal, Canada.

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

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

[31] Wechsler, M., P. Sheridan, and P. Schäuble, 1997. Multi-Language Text Indexing for Internet Retrieval. In: *Proc. of the RIAO'97 Computer-Assisted Information Searching on the Internet*, Montreal, Canada.