A systematic evaluation of concept-based cross-language information retrieval in the medical domain

Volk, Martin ; Buitelaar, P

A Systematic Evaluation of Concept-based Cross-Lingual Information Retrieval in the Medical Domain

Martin Volk
Eurospider Information Technology AG
Schaffhauserstrasse 18
CH-8006 Zürich, Switzerland
volk@eurospider.com

Paul Buitelaar
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
paulb@dfki.de

1 Introduction

The paper describes experiments and results of the MuchMore project[1], which is concerned with a systematic comparison of concept-based and corpus-based methods in cross-language information retrieval (CLIR) in the medical domain. Primary goals of the project are to develop and evaluate methods for the effective use of multilingual thesauri in the semantic annotation of English and German medical texts and subsequently to evaluate and compare the impact of such semantic information for the purpose of CLIR. In particular we describe work on semantic annotation with both domain-specific (UMLS, the Unified Medical Language system[2]) and general language semantic resources (EuroWordNet (Vossen, 1997)). Central to the approach is the use of linguistic processing (part-of-speech tagging, morphological analysis, phrase recognition and grammatical function analysis) for an accurate semantic annotation of relevant terms and relations in both the queries and the documents (Vintar et al., 2002). Especially for morphologically rich languages such as German it is important to extend linguistic processing beyond primitive stemming. All linguistic information was added to the documents into an XML format. The document collection used in the project is a parallel English-German corpus of approximately 9000 scientific medical abstracts with a total of one million tokens for each language.

2 Related Work

Many authors have experimented with machine translation or dictionary look-up for CLIR (see e.g. (Kraaij and Hiemstra, 1998)). In a comparison of such methods in both query and document translation, (Oard, 1998) 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 disambiguation of words that have more than one meaning and therefore may correspond to more than one translation (Sanderson, 1994). Therefore, in the case of short queries all translations are considered instead of trying to disambiguate between them.

Ambiguity is also important for 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 (Gey and Jiang, 1999), while (Eichmann et al., 1998) describe 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 and the Generalized Vector Space Model (Carbonell et al., 1997), both of which depend on parallel corpora. Finally, next to domain-specific thesauri also more general semantic resources such as EuroWordNet have been used in both monolingual and cross-language information retrieval.

In the MuchMore project we assign seman-

[1] http://muchmore.dfki.de
[2] http://umls.nlm.nih.gov

tic codes (MeSH, UMLS and EuroWordNet) to terms on the basis of a linguistic analysis. MeSH codes are assigned to terms in documents as well as in the queries. Annotation with UMLS codes is used for recognition and annotation of semantic relations. Finally, EuroWordNet senses are assigned to all (simple or complex) terms that are represented in this resource.

3 Annotation

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 for medical terms: 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). For example, the entry for *HIV pneumonia* in the Metathesaurus main termbank (MRCON) contains (among others) the concept identifier, the language of the term and the string:

```
C0744975 | ENG | HIV pneumonia
```

In addition to the mapping of terms to concepts, the Metathesaurus organizes concepts into a hierarchy by specifying relations between concepts. These are 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 for corpus annotation.

The third part is the **Semantic Network**, which provides a grouping of concepts according to their meaning into 134 semantic types. 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.

In the MuchMore project we assigned semantic codes to each sentence based on the linguistic information. MeSH codes were assigned to documents and to queries. UMLS concept identifiers were used as the basis for finding semantic relations. Appropriate EuroWordNet synset codes were assigned if a word or an expression belonged to a EuroWordNet synset (Buitelaar and Sacaleanu, 2001).

4 Evaluation

In order to evaluate whether semantic annotation results in a performance gain in information retrieval, several experiments were carried out. We used our corpus as document collection (the set of medical abstracts described above) in combination with a set of 25 queries and relevance assessments defined by medical experts that are partners in the project. MuchMore aims primarily at cross-language retrieval, but in order to assess CLIR performance, monolingual experiments in German and English were conducted as baselines for the cross-language experiments.

Below we present retrieval results in four columns. The first column contains the overall performance, measured as mean average precision (mAvP) as has become customary in the Text Retrieval Conference (TREC) experiments[3]. 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 integrates both precision and recall 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., 1998), the effectiveness within the high precision area is measured assuming that users are most interested in getting relevant documents ranked topmost in the result list. Because this number can

[3] http://trec.nist.gov/

vary substantially for different queries, we consider also the precision figure for the topmost documents retrieved (in column four). There we focus on the precision after the top 10 documents (P10).

4.1 Monolingual Evaluation Runs

For the retrieval experiments we used the commercial *relevancy* information retrieval system from Eurospider Information Technology AG. It is a vector-based retrieval system that can handle large document collections. 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 (Schäuble, 1997)).

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.

For each language, we produced a baseline performance by indexing only the tokens in both the documents and the queries. We call the German baseline DE-token. In addition an evaluation run based on linguistic stemming was produced which we termed DE-lemma. In table 1 we present the results of the monolingual German retrieval experiments.

In the baseline experiment for German (DE-token) the system finds only 322 relevant documents (out of 956). 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-lemma), which achieves a recall gain of 70% compared to DE-token. In parallel, the precision figures have improved substantially. Lemmatization was done in two steps. First we used a general-purpose (i.e. general vocabulary) morphological analyzer. It turned out that many medical terms were not lemmatized since they were not in the analyzer's lexicon. Therefore we developed heuristics for treating words that were unknown to the analyzer. Through these heuristics unknown adjectives were lemmatized by suffix truncation (*arthroskopischen* → *arthroskopisch*), and unknown nouns were decompounded if both compound parts were found as separate words in the corpus (*Nociceptinspiegel* → *Nociceptin Spiegel*). In this way the corpus itself was used as domain specific lexicon for decompounding.

We also experimented with a combination of token and lemmas. Both were combined as indexing terms of equal weights in the queries and the documents. This combination leads to a decrease in precision (see DE-token-lemma) and therefore the tokens were discarded in the subsequent runs.

The impact of the different types of semantic information was determined one by one, but always in combination with lemmas. We wanted to support the hypothesis that semantic information will improve the precision over pure lemma information. The results show 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 increases recall (from 591 to 601) and also average precision (from 0.2809 to 0.2873). As was to be expected the very specific semantic relations (Semrel) have hardly any impact. Using the EuroWordNet terms in combination with the lemmas degrades the overall performance.

4.2 Cross-language Evaluation Runs

The easiest approach to CLIR 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. Our approach however is different. 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.

A rough baseline for the cross-language task is

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE-token	0.1600	322	0.5622	0.4160
DE-lemma	0.2809	591	0.6759	0.5320
DE-token-lemma	0.2547	594	0.6744	0.5120
DE-lemma-EWN	0.2414	584	0.6140	0.4880
DE-lemma-MeSH	0.2873	601	0.6647	0.5280
DE-lemma-Semrel	0.2795	591	0.6474	0.5200

Table 1: Results of the monolingual German runs

to use 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 2).

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 latest version of the PC-based system PersonalTranslator (PT2002; 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. Still many translated queries are incomplete or incorrect but they scored surprisingly well with regard to recall. In table 2, line DE2EN-MT-PT2002, we see that the translated queries lead to 440 relevant documents at a rather low mean average precision of 0.1381.

Now we 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 2 has all the results. Again MeSH terms lead to the best results with respect to recall and 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 reach 404 relevant documents and a mean average precision of 0.1774. This precision clearly exceeds machine translation.

For the last two experiments we built a similarity thesaurus (SimThes) over the parallel corpus (Qui, 1995). Our similarity thesaurus contains German words (adjectives, nouns, verbs) from our corpus, each accompanied by a set of 10 English words that appear in similar contexts and are thus similar in meaning. The building of the similarity thesaurus can be understood as exchanging the roles of documents and terms in document retrieval. The documents now represent the indexing features and the terms are the retrievable items. (Schäuble, 1997) contains the technical details. In building a bilingual similarity thesaurus over a parallel corpus the term sets of two parallel documents are exchanged. Given a term from the source language we then compute similar terms from the target language.

If a similarity thesaurus is built over a monolingual corpus, it may 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. Each German word from the queries was then substituted by the English words of its similarity set. This resulted in the retrieval of 409 relevant documents and a relatively good mean av-

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE2EN-DE-token	0.0512	66	0.1530	0.1160
DE2EN-MT-PT2002	0.1381	440	0.3747	0.2920
DE2EN-EWN	0.0090	111	0.0311	0.0160
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 2: Results of the cross-language runs: German queries and English documents

erage precision of 0.2290 (see DE2EN-SimThes in table 2). Finally we checked the combination of all semantic annotations with the similarity thesaurus. Each query is now represented by its EuroWordNet, 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 1). And the figures for the high precision area (AvP and P10) are also outstanding. This result is approximating the results of monolingual retrieval with tokens, lemmas and semantic annotation.

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.

In cross-language retrieval the combination of all semantic information outperformed machine translation with respect to precision. It was only superseded by the use of a similarity thesaurus built over the parallel corpus where we used 10 similar words of the target language for each source language word. This means we included query expansion in combination with translation.

The highest overall performance resulted from a combination of the similarity thesaurus with the semantic information. This result is comparable to the German monolingual retrieval results in terms of precision but still 14% lower in the number of relevant retrieved items.

References

P. Buitelaar and B. Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proc. Of NAACL WordNet Workshop*, Pittsburgh.

J. Carbonell, 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*.

D. Eichmann, 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.

F.C. Gey 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)*, Gaithersburg, MD. National Institute of Standards Technology (NIST).

W. Kraaij and D. Hiemstra. 1998. TREC6 working notes: Baseline tests for cross language retrieval with the twenty-one system. In *TREC6 Working Notes*, Gaithersburg, MD. National Institute of Standards and Technology (NIST).

D. Oard. 1998. A comparative study of query and document translation for cross-lingual

information retrieval. In *Proc. of AMTA*, Philadelphia, PA.

Y. Qui. 1995. *Automatic Query Expansion Based on a Similarity Thesaurus*. Phd thesis, ETH Zurich.

M. Sanderson. 1994. Word sense disambiguation and information retrieval. In B. Croft and K. Van Rijsbergen, editors, *Proc. of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151. Springer-Verlag.

Peter Schäuble. 1997. *Multimedia Information Retrieval. Content-based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, Boston.

S. Vintar, P. Buitelaar, B. Ripplinger, B. Sacaleanu, D. Raileanu, and D. Prescher. 2002. An efficient and flexible format for linguistic and semantic annotation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, May 29-31.

Piek Vossen. 1997. EuroWordNet: A multilingual database for information retrieval. In *Proc. Of the DELOS Workshop on Cross-Language Information Retrieval.* Zurich, Switzerland, March, 5-7.