Cross-Lingual Medical Information Retrieval through Semantic Annotation

Špela Vintar[§], Bärbel Ripplinger[*], Paul Buitelaar[§]

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

*Eurospider Information Technology AG
Schaffhauserstrasse 18
CH 8006 Zürich, Switzerland
ripplinger@eurospider.com

Abstract

We present a framework for concept-based, cross-lingual information retrieval (CLIR) 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, whereby documents and queries are annotated with multiple layers of linguistic information. Linguistic processing includes POS-tagging, morphological analysis, phrase recognition and the identification of medical concepts and semantic relations between them.

The paper describes experiments in mono- and bilingual document retrieval, performed on a parallel English-German corpus of medical abstracts. Results show on the one hand that linguistic processing, especially lemmatisation and compound analysis, is a crucial step to achieving good baseline performance. On the other hand we show that semantic information, specifically the combined use of concepts and relations, significantly increases performance in cross-lingual retrieval.

Cross-Lingual Medical Information Retrieval through Semantic Annotation

1 Introduction

The task of finding relevant information from large, multilingual and domain-specific text collections is a field of active research within the NLP community. 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 utilising 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) where extensive, multilingual, semantic resources are available, such as UMLS[1] (Unified Medical Language System) in the medical domain. In addition to simply applying such existing semantic or terminological resources, efforts should also be directed towards developing ways to identify novel semantic information - concepts, relations - with which to extend these resources.

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-lingual medical IR. Among the chief goals of the project are therefore to: 1. Develop and evaluate methods in using UMLS for semantic annotation of English and German medical texts, covering terms and concepts at the MetaThesaurus level and semantic types and relations at the Semantic Network level; 2. Subsequently evaluate and compare the significance of such semantic information for the purposes of cross-lingual medical IR.

The rest of the paper is organized as follows. We first give a brief overview of related research, followed by a short description of the corpus used. Section 4 presents the resources and approach used in semantic annotation. Section 5 describes the experiments with the indexing, retrieval and weighting methods used, as well as an outline of the testing and evaluation scenario. Finally, Section 6 gives a comparison of the results obtained with different combinations of parameters.

2 Related Work

Many authors have experimented with MT or dictionary-based approaches to CLIR [6] [3], whereby dictionary-based query translation seems to work best for short queries while for long queries machine translation of the documents significantly outperforms other methods. Our work is more closely related to projects using multilingual thesauri or other structured semantic databases. Gonzalo et al [5] report on the use of EuroWordnet as a general language semantic resource both for mono- and cross-lingual IR. Exploiting and evaluating EuroWordnet annotation for CLIR in the medical domain is among the objectives of the MUCHMORE project, however within the scope of this paper we focus on experiments based on UMLS.

Domain-specific semantic resources have been used in TREC-8 for English-German CLIR within social science [4], while Eichmann and Ruiz [2] 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, whereas our approach is based on semantic annotation of both queries and documents, thereby enabling cross-lingual matching on the concept level. In addition, semantic annotation in our approach is not based merely on primitive stemming that allows term recognition via the UMLS word index, but rather on the use of sophisticated linguistic pre-processing that includes part-of-speech tagging, morphological analysis (including compound analysis for German) and phrase recognition.

[1] http://umls.nlm.nih.gov

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

3 Corpus and Linguistic Pre-Processing

The main corpus used in the MUCHMORE project is a parallel corpus of English-German scientific medical abstracts obtained from the Springer Link web site[3]. The corpus consists approximately of 1 million tokens for each language. Abstracts are taken from 41 medical journals (e.g. Der Nervenarzt, Der Radiologe, etc), each of which constitutes a relatively homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.).
Corpus preparation included removing HTML-tags, removing English segments from German abstracts and vice versa, deleting names of authors, addresses, etc., removing or converting symbols and other non-ASCII elements and producing a clean, plain text version of each abstract, consisting of a title (if available), 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, TnT [1] for part-of-speech tagging, Mmorph [7] for morphological analysis and Chunkie [8] for phrase recognition.

4 Semantic Annotation

The basic resource for semantic annotation is UMLS, which is organised in three parts: the Specialist Lexicon (covering lexical information, such as part-of-speech and morphology), the MetaThesaurus (covering terms and corresponding concepts), and the Semantic Network (covering semantic types and relations between them). The UMLS2001 version includes 1,734,706 terms mapped to 797,359 concepts, of which 1,462,202 entries are English and 66,381 German. Because only the MeSH part of the MetaThesaurus covers both German and English, we only use MeSH terms (564,011 term entries for English and 49,256 for German) in annotation.

Semantic relations are annotated on the basis of the UMLS Semantic Network, which defines 52 domain-specific hierarchically organised relations between Semantic Types. Since the semantic types are rather general (e.g. Pharmacological Substance, Patient or Group), the relations - when mapped to text - are often found to be very vague or even incorrect. This partly explains the unsatisfactory performance of semantic relations in the results of our experiments reported here. Therefore, we are developing ways of filtering existing relations to those most relevant for CLIR, as well as discovering new relations based on co-occurrence analysis and clustering.

Semantic annotation involves the following steps:

- each term is labelled with its Concept Unique Identifier (CUI), taken from the MRCON database,
- each CUI is mapped to its Semantic Type (TUI) on the basis of the MRSTY database,
- each pair of TUI's occurring within the same sentence is looked up in the SRSTRE1 database of the Semantic Network and in case of an existing semantic relation we annotate the related term pair and the type of relation found.

The identification of UMLS terms in the text is based on morphological pre-processing of both the term bank and the text, so that term lemmas are matched rather than word forms. The preparation of the term bank included other filtering and normalisation procedures, such as case folding, removal of very 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 length from 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 especially crucial for German.

Both morpho-syntactic (part-of-speech, morphology, phrases) and semantic (terms, semantic relations) annotation are integrated in a multi-layered XML annotation format, which organises various levels as separate tracks with options of reference between them via indices.

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

5 Experiments

In the retrieval experiments we used the commercial RotondoSpider system from Eurospider[4], which indexes tokens, extracted from documents and queries, using straight lnu.ltn weighting scheme [8]. For our purposes, morphological stemming has not been performed with the Spider facilities, but stems were instead extracted from the XML-annotated documents and queries. Similarly, the information about terms and semantic relations are extracted and indexed. Four different types of information are indexed (using the lnu.ltn weighting): word forms (tokens), their stemmed forms (lemmas), terms (CUI), and semantic relations, both for German and English.

To find out which information is most useful, for monolingual as well as for bilingual searches, a set of experiments has been carried out. For each language, two baseline runs have been conducted: The first run (DEnostem) indexes only the tokens of documents and queries. Because stemming and decomposition are known means to improve performance, in the second run (DEling) a linguistic stemming of the tokens using a morphosyntactic analysis (different from the one described above) has been applied.

To evaluate the use of the semantic annotations, the following runs have been conducted: The run (DESem) uses information extracted from the annotations as described above. In the second run (DESemLing) the token and lemma information are indexed and additionally linguistically stemmed. The same set of runs was performed for English, i.e. ENnostem, ENling, ENSem, ENSemLing.

In accordance with the main aims of the MUCHMORE project we also focus on semantic annotations for concept-based CLIR. Bilingual runs are performed using German queries to retrieve English documents, for which another two baseline runs were conducted: First, the lemmas were translated using a large translation dictionary DEENWB, and in a second run (DEENMTLing) the machine translation system SYSTRAN based on linguistically stemmed data has been used.

To evaluate the benefit of semantic information we carried out five runs: Three of them use semantic information only, i.e. no translation of lemmas has been done. To get a more detailed insight in the usefulness of information provided by terms and/or semantic relations, each type is evaluated separately: DEENT uses only term information, DEENSR only semantic relations, and DEENTSR both terms and semantic relations.

Because the queries used in the experiments do not necessarily contain identifiable medical concepts, in two further runs we additionally implement lemma translation using a large transfer lexicon. In DEENSemWB tokens and lemmas as provided by the annotated documents are translated, and in DEENSemWBLing these are further stemmed using a morphosyntactic analyzer (see above) before translation.

6 Results and analysis

For our experiments, we had access to a list of 126 English and German short but realistic queries provided by medical experts in the project. Of these, we used only 25 because at this time only for these queries relevance assessments existed for both German and English. As performance measurement we applied the average precision known from the TREC evaluations. Additionally we looked at the total recall, the interpolated precision at 0.1 (according to Eichmann and Ruiz [2]) and at the precision after 10 documents, where 10 documents means 50% of relevant documents per query.

Monolingual Experiments

Using information about term and semantic relations, in addition to lemma and token, degrades the overall performance to 65% of that of the best baseline run, and 97% of no stemming. Using linguistically stemmed data, DESemLing achieves 92% of the best baseline run. One reason for this result is the different types of features to be indexed. While the baseline runs used the RotondoSpider segmenter, which removes sentence delimiters and other non-ASCII characters, the annotated documents still contain these and hence they are indexed as tokens and lemmas. This negatively

[4] http://www.eurospider.com

influences the weighting, and thus possible performance gain by using semantic information is diminished.

	Average Precision	Recall	Interpolated Recall Precision Average at 0.1	Precision at 10 Docs
DEnostem	0.2557	470	0.6341	0.5080
DEling	0.3624	674	0.7468	0.5920
DESem	0.2494	629	0.8741	0.6960
DESemLing	0.3319	678	0.7898	0.6160

Due to the terminological density of medical texts the differences between the baseline runs in English using stemming or no stemming are much smaller (less inflected forms). In contrast to German, using additional semantic annotations achieves only 50% of the baseline runs whether linguistically stemmed or not. Working with a parallel corpus this result may seem surprising because the annotations of documents and queries should be equal for both languages. There is also a difference in the relevance assessments, for the 25 queries 959 German documents are relevant but only 500 in English. This does not influence the performance figures (except recall) but a direct comparison of the runs for each language is not possible.

	Average Precision	Recall	Interpolated Recall Precision Average at 0.1	Precision at 10 Docs
ENnostem	0.4473	433	0.7481	0.5240
ENling	0.4883	824	0.8219	0.6520
ENSem	0.2318	336	0.5030	0.3560
ENSemLing	0.2447	339	0.5102	0.3760

For the monolingual runs, the results given in the tables above show that a good linguistic stemming (including decomposition) is necessary to attain a certain performance level. However, in the high precision area, the use of semantic information at least in German shows a slightly better performance than the baseline runs.

Bilingual Experiments

In contrast to the monolingual experiments, the bilingual runs achieve a certain performance gain using semantic information. As shown in the table below, concept-based CLIR (run DEENTSR) using term as well as semantic relations outperforms the baseline run using a transfer dictionary for general language (gain of 25%). Furthermore, as a detailed analysis of run DEENSR and DEENT shows, term information seems to contribute mostly to the achieved performance gain.

As mentioned before not all queries contain identifiable medical concepts, and thus a lemma translation is highly recommended and results in a much better performance compared to the pure concept-based runs.

Similar to the monolingual experiments, the contribution of a precise lemmatisation and compound analysis has been evaluated. The run where the lemma and tokens are additionally linguistically analysed shows a performance gain of 30% compared to the respective baseline run (DEENMTLing), and of 90% compared to the run without translation (DEENTSR). Considering recall and high precision area this run is superior to all other runs.

The outcome of these experiments gives clear hints in which direction future work will be trended. First, an improvement of the morphological analysis used, and a more precise term tagging. Possible enhancements can also come from a better, i.e. domain specific transfer lexicon.

	Average Precision	Recall	Interpolated Recall Precision Average at 0.1	Precision at 10 Docs
DEENWB	0.0985	193	0.2483	0.1360
DEENMTLing	0.1863	305	0.3490	0.2080
DEENT	0.1082	175	0.2420	0.1480
DEENSR	0.0515	26	0.1780	0.0840
DEENTSR	0.1227	175	0.2722	0.1640
DEENSemWB	0.1796	306	0.3735	0.2320
DEENSemWBLing	0.2327	349	0.5456	0.3120

7 Conclusions

The monolingual experiments show that high-quality linguistic analysis is crucial for 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 performance, which would enable us to evaluate the gain of using semantic information in more detail. However, the concept-based method used for cross-lingual retrieval in MUCHMORE already shows a performance gain, for the monolingual retrieval the effectiveness increases at least in the high precision area. So far semantic annotations were based only on existing resources (UMLS), but in future we envisage the integration of novel extracted terms and relations as well as relevance filtering of existing relations.

References

1. Brants T. 2000. *TnT - A Statistical Part-of-Speech Tagger*. In: Proceedings of 6th ANLP Conference, Seattle, WA.
2. Eichmann D., Ruiz M. and Srinivasan P. 1998. Cross-Language Information Retrieval with the UMLS Metathesaurus. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia.
3. Gaussier E., Grefenstette G., Hull D.A. and Schulze B.M. 1998. Xerox TREC-6 site report: Cross language text retrieval. In: Proceedings of The Sixth Text Retrieval Conference (TREC-6). Gaitersburg, MD: National Institute of Standards Technology (NIST).
4. Gey F.C. and Jiang H. 1999. English-German Cross-Language Retrieval for the GIRT Collection - Exploiting a Multilingual Thesaurus. In: The Eighth Text REtrieval Conference (TREC-8), draft notebook proceedings.
5. Gonzalo J., Verdejo F. and Chugur I. 1999. *Using EuroWordNet in a Concept-based Approach to Cross-Language Text Retrieval* Applied Artificial Intelligence:13, 1999.
6. Oard D. 1998. *A comparative study of query and document translation for cross-lingual information retrieval* In: Proceedings of AMTA, Philadelphia, PA.
7. Petitpierre D. and Russell G. 1995. *MMORPH - The Multext Morphology Program.* Multext deliverable report for the task 2.3.1, ISSCO, University of Geneva.
8. Singhal A., C. Buckley, and M. Mitra. *Pivoted Document Length Nomalization*. In: Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21-29, 1996, Zurich.
9. Skut W. and Brants T. 1998. *A Maximum Entropy partial parser for unrestricted text.* In Proceedings of the 6th ACL Workshop on Very Large Corpora (WVLC), Montreal.
10. Vossen P. 1997. *EuroWordNet: a multilingual database for information retrieval.* In: Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.