Cross-Lingual Medical Information Retrieval through Semantic Annotation

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

Spela Vintar and **Paul Buitelaar**
DFKI GmbH
Stuhlsatzenhausweg 3,
D-66123 Saarbrcken, Germany
{*vintar*, *paulb*}@dfki.de

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, increases performance in cross-lingual retrieval at least in the high precision area.

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 (3)

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

(6), 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 et al. (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.

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 (9) for morphological analysis and Chunkie (13) 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

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

- 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.

5 Experiments

In order to evaluate whether the semantic annotations result in a performance gain, several experiments have been carried out. We used our own document collection (described above) as well as a query set selected by medical experts within the project. The OHSUMED text collection would not have been appropriate for the MUCHMORE project due to its monolingual nature (documents and queries are only available in English).
The MUCHMORE project focuses primarily on the retrieval of documents in a multilingual context. To evaluate the CLIR performance, monolingual experiments in German as well as English are conducted acting as baselines for the bilingual experiments. Most CLIR systems achieve only up to 75% of the monolingual IR (cf. (11), and one aim of MUCHMORE is to show whether semantic annotations could improve CLIR performance.

For the retrieval experiments we used the commercial RotondoSpider system from Eurospider. The system indexes tokens, extracted from documents and queries, using straight lnu.ltn weighting scheme (12). The tokens are normalised by applying a lexicon-based (Celex) stemmer for German, and a Porter-like stemmer for English (15); tokens and stems are equally indexed.
Due to the modular approach of RotondoSpider, this configuration could easily be adapted to index the information provided by the semantically annotated documents and queries: word forms (tokens), their stemmed forms (lemmas), terms, and semantic relations, both for German and English.

For each language, we produced a baseline run by indexing documents and queries without any normalisation steps, DEnostem and ENnostem. Some recent works (7), (8) have shown that at least for German a linguistic-based stemming and decompounding is beneficial, therefore another run based on linguistic stemming is performed, DELing. For similarity reasons, also the English documents were indexed based on linguistic stemming, ENLing. These runs could also be seen as a kind of baseline. The runs DESem and ENSem use the information extracted from the annotated documents (described above). In two further runs, DESemLing and ENSemLing, the tokens and lemmas are additonally linguistically analyzed with another morpho-syntactic analyzer as described in Section 3 in order to see how a good linguistic component influences the effectiveness of the retrieval.

For the bilingual runs we use German queries and retrieve English documents. These runs should not only be compared to the monolingual runs, our aim was rather to test and evaluate different approaches. For runs DEENWB and ENDEWEB the original RotondoSpider configuration is used and the stems are translated using a general lexicon (which contains also medical terms). For DEENMTLing and ENDEMTLing the query tokens are linguistically stemmed and translated with the MT system Systran. The next two runs use a conceptual transfer method based on terms and semantic relations: In the runs DEENSem and ENDESem the translation is only based on information about terms and semantic relations, lemmas and tokens are not translated. For DEENSemWB and ENDESemWB the lemmas and tokens are additionally translated using a dictionary built over the Springer abstracts based on a Similarity Thesaurus (10). Thus, query expansion – a well-known means in CLIR to improve effectiveness – based on various translations for a source language word is performed.

6 Results and Analysis

The overall performance is measured using the TREC-styled mean average precision (mAvP). According to Eichmann et al. (2), 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. Thus, we used the average precision at 0.1 recall (AvP01). Because this number could substantially vary for various queries we consider also the precision figures for the topmost documents retrieved. We also focus on the top 10 documents (P10) which in our case means that 50

For the experiments, we used the relevance assessments based on 25 queries provided by the medical experts in the MUCHMORE project. These are relevance assessments performed both for the German and English document collections. However, the number of documents classified as relevant for the two collections are quite different, 959 for German and 500 for English. The main reason for this discrepancy is the different profile of experts who performed 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 for the sake of data comparability.

In Tables 1 and 2 the results of the monolingual German and English experiments are presented: The run (DESem) using semantic annotations only achieves 97% of the performance (measured by mean average precision) of the no-stemming run (DEnostem), and only 65% of the run based on linguistically produced stems. For German, using information about term and semantic relations, in addition to lemmas and tokens, degrades the overall performance. The importance of good linguistic stemming and decompounding is shown by the DESemLing run which achieves a performance gain of 30% compared to no-stemming, but looses 10% compared to the run which uses no conceptual information (DELing). One reason for this result is the different types of features to be indexed. While DEnostem and DELing used the RotondoSpider segmentation which removes sentence delimiters and other non-ASCII characters, the annotated documents still contain these and hence they are indexed as tokens. This negatively influences the weighting, and possible performance gain by using semantic information is diminished.

The figures for the high precision area show that semantic information brings a significant performance gain, 38% compared to nostemming, and 17% compared to the run using linguistically stemmed data. Also, more relevant documents are retrieved.

Table 1: Results of the monolingual German runs

	mAvP	Rel. Docs Retr.	AvP0.1	P10
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 runs in English using stemming (ENLing) versus nostemming (ENnostem) are much smaller (due to fewer inflected forms). In contrast to German, the run which uses semantic annotations achieves only 52% (mean average precision) of ENnostem, and only 47% of the performance of ENLing. Also the linguistic stemming of tokens and lemmas does not result in a substantial performance gain.

Table 2: Results of the monolingual English runs

	mAvP	Rel. Docs Retr.	AvP0.1	P10
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

In contrast to German, for English even the figures within the high precision area deteriorate showing a loss of 68% and 61% compared to the runs without semantic information (ENSem vs. ENnostem, vs. ENLing). The same observation can be made by considering the number of relevant documents retrieved.

For the bilingual runs, the results are given in Table 3. The performance figures show that the run doing a concept-based transfer (DEENSem) retrieves not only the most relevant documents but also shows a performance gain in the high precision area. The overall performance, however, achieve only 80% of the run where stems are translated by an MT system. Also the run with query expansion (DEENSemWB) does not achieve the performance of the MT run.

Table 3: Results of the bilingual run: German to English

	mAvP	Rel. Docs Retr.	AvP0.1	P10
DEENWB	0.0985	193	0.2483	0.1360
DEENMTLing	0.1863	305	0.3490	0.2080
DEENSem	0.1492	349	0.3547	0.2520
DEENSemWB	0.1796	306	0.3735	0.2320

One reason for this results is certainly that not all of the 25 queries used in the experiment are tagged with semantic annotations (actually only 20), thus for some of them no relevant document is found. This however influences the calculation of the average precision figure. Compared to the monolingual run ENSem, the bilingual run achieves 64

Another aspect of interest has been the contribution of the different types of semantic information to the performance. The figures presented in Table 4 shows a clear advantage of using term information.

Table 4: Results of the runs using different types of semantic information only

	mAvP	Rel. Docs Retr.	AvP0.1	P10
DEENT	0.1082	175	0.2420	0.1480
DEENSR	0.0515	26	0.1780	0.0840
DEENTSR	0.1227	175	0.2722	0.1640

7 Conclusions

The monolingual runs using conceptual information in the form of terms and semantic relations does not show a performance gain with respect to average precision figures. For German and English, using information about term and semantic relations, in addition to lemmas and tokens, degrades the overall performance. Moerover, 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 at least in the high precision area, and in the number of relevant documents retrieved. In terms of overall precision, the runs achieve the usual performance level of CLIR systems compared to monolingual IR systems. 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, and hope to further improve performance.

References

Brants T. 2000. TnT - A Statistical Part-of-Speech Tagger. In: *Proceedings of 6th ANLP*

Conference, Seattle, WA.

Eichmann D., M. Ruiz, and P. Srinivasan. 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.

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

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

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.

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

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

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.

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.

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

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

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

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

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

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