Ontologies in Cross-Language Information Retrieval

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

Špela Vintar, Paul Buitelaar
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
{vintar, paulb}@dfki.de

Abstract: We present an approach to using ontologies as interlingua in cross-language information retrieval in the medical domain. Our approach is based on using the Unified Medical Language System (UMLS) as the primary ontology. Documents and queries are annotated with multiple layers of linguistic information (part-of-speech tags, lemmas, phrase chunks). Based on this we identify medical terms and semantic relations between them and map them to their position in the ontology.

The paper describes experiments in monolingual and cross-language document retrieval, performed on a corpus of medical abstracts. Results show that semantic information, specifically the combined use of concepts and relations, increases the precision in monolingual retrieval. In cross-language retrieval the semantic annotation outperforms machine translation of the queries, but the best results are achieved by combining a similarity thesaurus with the semantic codes.

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. 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 multilingual ontologies or thesauri to bridge the gap between surface linguistic form and meaning. The latter seem particularly appropriate for domains (and languages) for which extensive multilingual ontologies are available, such as UMLS (Unified Medical Language System) in the medical domain.

The experiments reported in this paper were performed within the MUCHMORE project[1], which aims at systematically comparing concept-based and corpus-based methods in cross-language medical information retrieval.

[1]MUCHMORE is a European Union project under grant IST-1999-11438 which also cooperates with two US partners. For details see http://muchmore.dfki.de.

2 Related Work

Many authors have experimented with machine translation or dictionary look-up for CLIR (see [KH98]). In a comparison of such methods in both query and document translation, Oard [Oar98] 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. An important problem in the translation of short queries is the lack of context for the disambiguation of words that have more than one meaning and therefore may correspond to more than one translation.

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 by [ERS98] who describes the use of the UMLS MetaThesaurus for French and Spanish queries on the OHSUMED text collection, a subset of MEDLINE. He uses the thesaurus as a source for compiling a bilingual lexicon, which is then used for query translation. Next to domain-specific thesauri more general semantic resources such as EuroWordNet [Vos97] have been used in both mono- and cross-language information retrieval.

The work we describe here is 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 for an accurate semantic annotation of relevant terms and relations in both the queries and the documents.

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[2]. 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 special tags and symbols in order to produce a clean, plain text version of each abstract, consisting of a title, text and keywords. The corpus was then linguistically annotated using standard tools for shallow processing: a tokenizer, a statistical part-of-speech tagger, a morphological analyser and a chunker for phrase recognition.

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

3.2 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). A simplified entry for *HIV pneumonia* in the Metathesaurus main termbank (MRCON) looks like this:

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

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 identifier, and the string itself.

In addition to the mapping of terms to concepts, the Metathesaurus - a true ontology - organizes concepts into a hierarchy by specifying relations between concepts. These are thesaurus-type generic relations like *broader_than, narrower_than, parent, sibling* etc. 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 (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.

3.2.1 Terms and Concepts

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.

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.

3.2.2 Semantic relations

Semantic relations are annotated on the basis of the UMLS Semantic Network, which defines binary relations between semantic types 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
```

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 (cf. the results in section 4.2).

3.3 Semantic Annotation using EuroWordNet

In addition to annotation with UMLS, terms are annotated with EuroWordNet senses [Vos97] to compare domain-specific and general language use. 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.

4 Evaluation in Information Retrieval

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 set of 25 queries with human relevance assessments provided by the medical expert in the MUCHMORE project. In these assessments the number of relevant documents per query varies between 7 and 104. They add up to a total of 959 relevant documents for the 25 queries.

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

- *Arthroskopische Behandlung bei Kreuzbandverletzungen.*
  Arthroscopic treatment of cruciate ligament injuries.

4.1 Monolingual Evaluation Runs

MUCHMORE aims first and foremost at cross-language retrieval (CLIR). In order to set the CLIR performance into perspective, monolingual experiments in German and English were conducted acting as baselines for the cross-language experiments.

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 [Sch97]).

For the MUCHMORE evaluation runs we adapted the *relevancy* system so that it indexes the information provided by the annotated documents and queries: word forms (tokens) and their base forms (lemmas) for all indexable parts-of-speech. 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.

In table 1 we present the results of the monolingual English retrieval experiments. 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. This figure is computed as the mean of the precision scores after each relevant document retrieved. 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). 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 English (EN-token) we find 617 relevant documents (out of 956; cf. table 1). The mean average precision (mAvP) is 0.35, and the average precision

in the top ranks is high (AvP = 0.80). So, the few documents that are found are often ranked at the top of the list. On average there are 6.16 relevant documents among the 10 top ranked documents (P10).

Linguistic lemmatization (stemming) worsens the precision for English monolingual retrieval. But it does increase the recall when used in combination with tokens (see line EN-token-lemma). This is very different from German monolingual retrieval which clearly improves with lemmatization both for recall and precision. The additional benefit was particularly due to segmentation of German compounds.

The impact of the different types of semantic information was determined one by one, but always in combination with tokens. We wanted to support the hypothesis that semantic information will improve the precision over pure token information. It turns out that MeSH codes are the most useful indexing features among the semantic codes. Using MeSH codes slightly increases recall (from 617 to 637) but most impressively improves average precision (from 0.3455 to 0.3637). The positive impact of the UMLS terms is less visible.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
EN-token	0.3455	617	0.8077	0.6160
EN-lemma	0.3097	600	0.6632	0.5360
EN-token-lemma	0.3320	635	0.7543	0.5760
EN-token-EWN	0.2155	604	0.5847	0.4000
EN-token-UMLS	0.3455	617	0.8077	0.6160
EN-token-MeSH	0.3637	637	0.8259	0.6040

Table 1: Results of the monolingual English runs

Using the EuroWordNet terms (EWN) in this combination with tokens degrades the overall performance. We investigated this phenomenon and found that EuroWordNet terms in our queries are mostly general language words like *injury, complication* or *treatment*. By using these words as additional indexing features we give them more weight than content-bearing specific terms. This leads to a bias towards the general language words and thus to a loss in retrieval precision.

4.2 Cross-Language Evaluation Runs

For the Cross-Language Information Retrieval we assume that we have a document collection (i.e. a corpus) in one language and a query in another language. We used German queries to retrieve English documents.

As a baseline we investigated the use of Machine Translation (MT) for translating the queries. We employed 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. Many translations are incomplete or incorrect but still the automatically

translated queries scored well with regard to recall. In table 2, line DE2EN-MT-PT2002, we see that these queries lead to 440 relevant documents at a (rather low) mean average precision of 0.1381.

Now let us compare these results with the results based on 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. We are regarding the semantic codes as an interlingua to bridge the gap between German and English.

The second block in table 2 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).

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE2EN-MT-PT2002	0.1381	440	0.3747	0.2920
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 2: 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. 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 an automatically constructed 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: *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 2). 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. [BS00]) 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 2). And the figures for the high precision area (AvP and P10) are also outstanding.

5 Conclusions

We have explored the use of different kinds of semantic annotation derived from the UMLS ontology 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.

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.

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 improve the performance by the integration of disambiguation for UMLS and EuroWordNet terms as well as including novel extracted terms and relations for UMLS.

References

[BS00] Martin Braschler and Peter Schäuble. Using Corpus-Based Approaches in a System for Multilingual Information Retrieval. *Information Retrieval*, (3):273–284, 2000.

[ERS98] D. Eichmann, M. Ruiz, and P. Srinivasan. 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, 1998.

[KH98] W. Kraaij and D. Hiemstra. TREC6 Working Notes: Baseline Tests for Cross Language Retrieval with the Twenty-One System. In *TREC6 Working Notes*, Gaithersburg, MD, 1998. National Institute of Standards and Technology (NIST).

[Oar98] D. Oard. A Comparative Study of Query and Document Translation for Cross-Lingual Information Retrieval. In *Proc. of AMTA*, Philadelphia, PA, 1998.

[Sch97] Peter Schäuble. *Multimedia Information Retrieval. Content-based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, Boston, 1997.

[Vos97] Piek Vossen. EuroWordNet: A Multilingual Database for Information Retrieval. In *Proc. Of the DELOS Workshop on Cross-Language Information Retrieval.* Zurich, Switzerland, March, 5-7 1997.