An Experimental Assessment of Direct *vs.* Interlingual Translation for Cross-Language Information Retrieval

Michael Poprat, Udo Hahn

Text Knowledge Engineering Lab
Freiburg University
Werthmannplatz 1
D-79085 Freiburg, Germany
hahn@coling.uni-freiburg.de

Joachim Wermter, Stefan Schulz, Kornél Markó

Medical Informatics Department
Freiburg University Hospital
Stefan-Meier-Str. 26
D-79104 Freiburg, Germany
stschulz@uni-freiburg.de

Abstract

We introduce an interlingua-based approach to cross-language information retrieval, in which queries, as well as documents, are mapped onto a language-independent concept layer and retrieval operations are performed at the level of that interlingua. This approach is contrasted with one which operates without such an intermediary concept level. Non-English queries (German ones, in our experiments) are directly translated to English queries which, subsequently, are processed on English documents. We provide an empirical evaluation of both alternatives on a large medical document collection.

Introduction

Medical document retrieval (Hersh 2002) and text retrieval on the WWW (Kobayashi & Takeda 2000) share many challenges for the design and implementation of retrieval systems. First, clinical document collections are usually very *large* and *dynamic*, with estimates ranging, e.g., for a single clinical site, on the order of millions of documents in total, and hundreds to thousands new documents being added every day. Similar observations can be made for biomedical publications which are increasingly available on the Web. From the viewpoint of the design of information retrieval systems, this rules out, on a broader scale, the reuse of many of the sophisticated statistical approaches which perform so well under small-scale experimental conditions such as the vector space model, latent semantic indexing, or even more sophisticated probabilistic models (for a brief survey, cf. Manning & Schütze (1999, Ch. 15)). The reason for this is that currently no search engine is capable of routinely maintaining high-dimensional document-term vectors ($n \gg 100,000$) for such an enormous volume of documents and high rate of update frequencies.

Second, medical and Web document collections are truly *multi-lingual*. While clinical documents are typically written in the physicians' native language, searches in major bibliographic databases (such as MEDLINE) require substantial knowledge of (expert-level) English medical terminology. Hence, some sort of bridging between synonymous or 'related' terms from different languages has to be realized to make proper use of the information these databases hold.

Third, the user population of medical document retrieval systems and their search strategies are really *diverse*. Not only physicians, but also nurses, medical insurance companies and patients are increasingly getting access to these resources, with the Web adding an even more scattered crowd of searchers. Hence, mappings between different *jargons and sublanguages* are required to serve the needs of such a heterogeneous search community. The simplicity of the content representation of the documents, as well as automatically performed intra- and interlingual lexical alignments of semantically equivalent expressions then become crucial issues for an adequate methodology of information retrieval.

Our approach is intended to meet these particular challenges. At its core lies a new type of dictionary, in which the entries are *equivalence classes* of subwords, i.e., semantically minimal content descriptors. These equivalence classes capture intralingual as well as interlingual synonymy. As equivalence classes abstract away from subtle particularities within and between languages and reference to them is realized via a language-independent code system, they form an *interlingua*. Compared with relationally richer, e.g., WORDNET based, interlinguas used for cross-language retrieval (Gonzalo, Verdejo, & Chugur 1999; Ruiz, Diekema, & Sheridan 1999), we use a truly limited set of semantic relations and pursue a more restrictive approach to synonymy. In particular, we do not claim to cover general language but rather restrict ourselves to the sublanguage used in the context of the medical domain. We contrast this interlingua-based retrieval approach to one which relies on a direct, i.e., term-based translation of non-English (here, German) queries to English ones for subsequent processing on the English document collection and evaluate these two alternatives on a large medical document collection.

Morpho-Semantic Document Retrieval

Subwords as Morpho-Semantic Description Units

Our work is based on the assumption that neither fully inflected nor heuristically stemmed words – such as common in many text retrieval systems – constitute the appropriate granularity level for content description. Especially in scientific and technical sublanguages, we observe that basic semantic entities are chained in complex word forms such as in *'pseudo⊕hypo⊕para⊕thyroid⊕ism'*, *'gluco⊕corticoid⊕s'*,

or *'pancreat⊕itis'* ('⊕' denotes the concatenation operator). Domain-specific suffixes (e.g., *'-itis'*) and single-word compounds are even more accentuated in many languages other than English, e.g., German. In order to deal with these phenomena, we introduce *subwords* as self-contained, semantically minimal units and motivate their existence by their usefulness for document retrieval rather than by linguistic considerations.

The minimality criterion for subwords is difficult to define in a general way. Given the token *'diaphysis'*, e.g., a linguistically plausible morpheme-style decomposition might lead to *'dia⊕phys⊕is'*. From a medical perspective, a coarser segmentation into *'diaphys⊕is'* seems much more reasonable, because the canonical linguistic decomposition is far too fine-grained and likely to create many subword ambiguities. Comparable 'low-level' decompositions of semantically unrelated tokens such as *'dia⊕lyt⊕ic'*, *'phys⊕iol⊕ogy'* postulate *'dia'* and *'phys'* as morpheme-style units, which unwarrantedly will then match *'dia⊕phys⊕is'*, too. The (semantic) self-containedness of a subword is, however, often supported by the existence of a synonym, e.g., *'shaft'* for *'diaphys'*.

Subwords are assembled in a multilingual lexicon and thesaurus. For each subword entry, special attributes of and semantic relations between subwords are specified. The lexicon and the thesaurus are both constructed manually, based on the following considerations:

- Subwords are registered, together with their attributes such as language (English or German), subword type (stem, prefix, suffix, invariant), comments and user metadata. Each lexicon entry is assigned a unique identifier and one synonymy class (identified by a so-called *equivalence class identifier, EC-ID*), which contains this entry as its only member.
- Synonymy classes which contain intralingual synonyms and interlingual translations of subwords are fused. Intra- and interlingual semantic equivalence are judged within the context of medicine only.
- Semantic links between synonymy classes are added. We subscribe to a shallow approach in which semantic relations are restricted to:
  (i) a paradigmatic relation *has-meaning*, which relates one ambiguous class to its specific readings, e.g.:
  $\{head\} \Rightarrow \{kopf,zephal,caput,cephal\}$ *OR* $\{leader, boss\}$.
  (ii) a syntagmatic relation *expands-to*, which consists of predefined segmentations for utterly short subwords, such as $\{myalg\} \Rightarrow \{muskel,muscle\} \oplus \{schmerz,pain\}$.

We refrain from introducing hierarchical relations between EC-IDs, because such links are acquired from domain-specific vocabularies (Markó *et al.* 2003), e.g., the Medical Subject Headings (MESH 2001).

Maintenance of the Subword Repositories

The manual construction of the lexicon and the thesaurus has consumed, up until now, about three person-years. The combined subword lexicon (release 08/2003) contains 43,165 entries, 21,237 for English[1] and 21,928 for German. Given our observations over time, we found a well-known logarithmic growth behavior as far as the number of subword entries are concerned (Schulz & Hahn 2000). All of these entries are related in the subword thesaurus by 17,805 equivalence classes. The average cardinality per equivalence class is 2,42. Furthermore, there are 760 ambiguity links between classes, together with 269 expansion links.

The process of lexicon construction is a challenging task which requires in-depth knowledge of biomedical terminology. In the development workflow, the effects of lexicon and thesaurus updates are immediately fed back to the developers using word lists to test both the segmentation and the assignment of EC-IDs. Furthermore, a collection of parallel texts (abstracts of medical publications in English and German as shown in Table 1) are used to detect errors in the assignment of EC-IDs. In order to impose common policies on the lexicon builders, we developed a maintenance manual which contains 31 rules. The most critical tasks they cover are listed below:

- The proper delimitation of subwords (e.g., *'compat⊕ibility'* vs. *'compatib⊕ility'*);
- The decision whether an affix introduces a new meaning which would justify a new entry (e.g., *'neur⊕osis'* vs. *'neuros⊕is'*);
- Data-driven decisions, e.g., adding *'-otomy'* as a synonym of *'-tomy'* in order to block erroneous segmentations such as *'nephrotomy'* into *'nephr⊕oto⊕my'* (assuming, of course, that *'-oto'* and *'-my'* are already in the lexicon);
- The decision to exclude short stems from segmentation (such as *'my-'*, *'ov-'*) in order to block false segmentations;
- The decision to locate the appropriate level of semantic abstraction when defining equivalence classes, e.g., by grouping {*hyper-, high, elevate*} into the same class;
- The decision which function words and affixes are excluded from indexing, such as *'and'*, *'-ation'*, *'-able'*, and those which are not, e.g., *'dys-'*, *'anti-'*, *'-itis'*.

Morpho-Semantic Normalization

Table 1 illustrates how source documents from different languages are converted into an interlingual representation by a three-step procedure. The first step deals with **orthographic normalization** (cf. Table 1, second column). A preprocessor reduces all capitalized characters from input documents to lower-case characters and, additionally, performs language-specific character substitutions (such as the normalization of German umlauts) in order to ease the matching of (parts of) text tokens and entries in the lexicon.

The next step in the pipeline is concerned with **morphological segmentation**. The system decomposes the ortho-

[1] Just for comparison, the size of WORDNET assembling the lexemes of general English in the 2.0 version is on the order of 152,000 entries (`http://www.cogsci.princeton.edu/~wn/doc.shtml`, last visited on January 3, 2004)

Original Document	Orthographic Normalization	Morphological Segmentation	Semantic Normalization
High TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism.	high tsh values suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.	high tsh value s suggest the diagnos is of primar y hypo thyroid ism while a suppress ed tsh level suggest s hyper thyroid ism.	**#up# tsh #value#** #suggest# **#diagnost# #primar# #small# #thyre# #suppress# tsh #nivell#** #suggest# **#up# #thyre#** .
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose, ein supprimierter TSH-Spiegel spricht dagegen für eine Schilddrüsenüberfunktion.	erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose, ein supprimierter tsh-spiegel spricht dagegen fuer eine schilddruesenueberfunktion.	er hoeh te tsh - wert e erlaub en die diagnos e einer primaer en hypo thyre ose, ein supprim iert er tsh - spiegel spricht dagegen fuer eine schilddrues en ueber funktion.	**#up# tsh - #value#** #permit# **#diagnost# #primar# #small# #thyre#** , **#suppress# tsh** - {#mirror# **#nivell#**} #speak# **#thyre# #up#** #function# .

Table 1: Morpho-Semantic Indexing for English (row 1) and German (row 2) (Bold EC-IDs co-occur in both documents.)

graphically normalized input stream into a sequence of subwords (corresponding to the entries in the subword lexicon) and lexical remainders (not in the lexicon) (cf. Table 1, third column). The segmentation results are checked for morphological plausibility using a finite-state automaton in order to reject invalid segmentations (e.g., segmentations without stems or beginning with a suffix). If there are ambiguous valid readings or incomplete segmentations (due to missing entries in the lexicon), a series of heuristic rules are applied, which prefer those segmentations with the longest match from the left, the lowest number of unspecified segments, etc. Whenever the segmentation algorithm fails to detect a valid reading, the original word is restituted.

In the final step, **semantic normalization**, each subword is substituted by its corresponding EC-ID (cf. Table 1, column 4). After that step, all synonyms within a language and all translations of semantically equivalent subwords from different languages are represented by the same code item in that target representation. Composed terms (such as *'myalg⊕y'*) which are linked to their components by the *expands-to* relation are substituted by the EC-IDs of their components, in the same way as if this were performed by the segmenter. Ambiguous classes, i.e., those related by a *has-meaning* link to two or more classes, produce a sequence of their related EC-IDs (cf. the set notation in Table 1, column 4, row 2).[2] The result generated by this pipeline is a morpho-semantically normalized document in a language-independent, interlingual representation.

Experimental Setting

Document Corpus

Our experiments were run on the OHSUMED corpus (Hersh *et al.* 1994), which constitutes one of the standard IR testbeds for the medical domain. It is a subset of the MEDLINE database which contains bibliographic information (author, title, abstract, index terms, etc.) of life science and biomedicine articles. We only considered that OHSUMED subset where each bibliographic unit contained a title and an abstract field, and came up with a document collection comprised of 233,445 texts. Our test collection is made of roughly 42 million tokens, and the average document length is 179.6 tokens (with a standard deviation of 76.4).

In the OHSUMED corpus, 106 queries are available (actually 105, because for one query no relevant documents could be found), including associated relevance judgments. The average number of query terms is 5.1 (with a standard deviation of 1.8). The following is a typical query: *"Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy?"*. In Figure 1, the result of processing this query and an extract of one retrieved document illustrate the two alternative approaches we discuss. (Bold terms co-occur in queries and the document fragment.)

The OHSUMED corpus contains only English-language documents (and queries). This raises the question of how this collection (or MEDLINE, in general) can be accessed from other languages as well. It is a realistic scenario, because, unlike in sciences with English as a *lingua franca*, among medical doctors native languages are still dominant in their education and everyday practice. In order to solve this problem, medical practitioners might resort to translating their native-language search problem to English with the help of current Web technology, e.g., an automatic translation service available in a standard Web search engine. Its operation might be further enhanced by lexical ressources as available from the U.S. National Library of Medicine in support of various non-English languages, e.g. the UMLS Metathesaurus (UMLS 2003), which currently supports – with considerable differences in coverage – German, French, Spanish, Portuguese, Russian, and many others. Relying on the quality of the translation, this procedure then reduces the cross-language retrieval problem to a monolingual one. As an alternative, we introduce the Morpho-Semantic Indexing approach (henceforth, MSI). Both of them will then be evaluated on the same query and document set. As the baseline for our experiments, we provide a Boolean retrieval system operating with the Porter stemmer (Porter 1980) and language-specific stop word lists[3] so that the system runs on (original) English documents with (original) English queries.

[2] Work on a disambiguation metric is in progress.

[3] We used the stemmer available on `http://www.snowball.tartarus.org` (last visited on October, 16 2003). The stop word lists contained 172 English and 232 German entries.

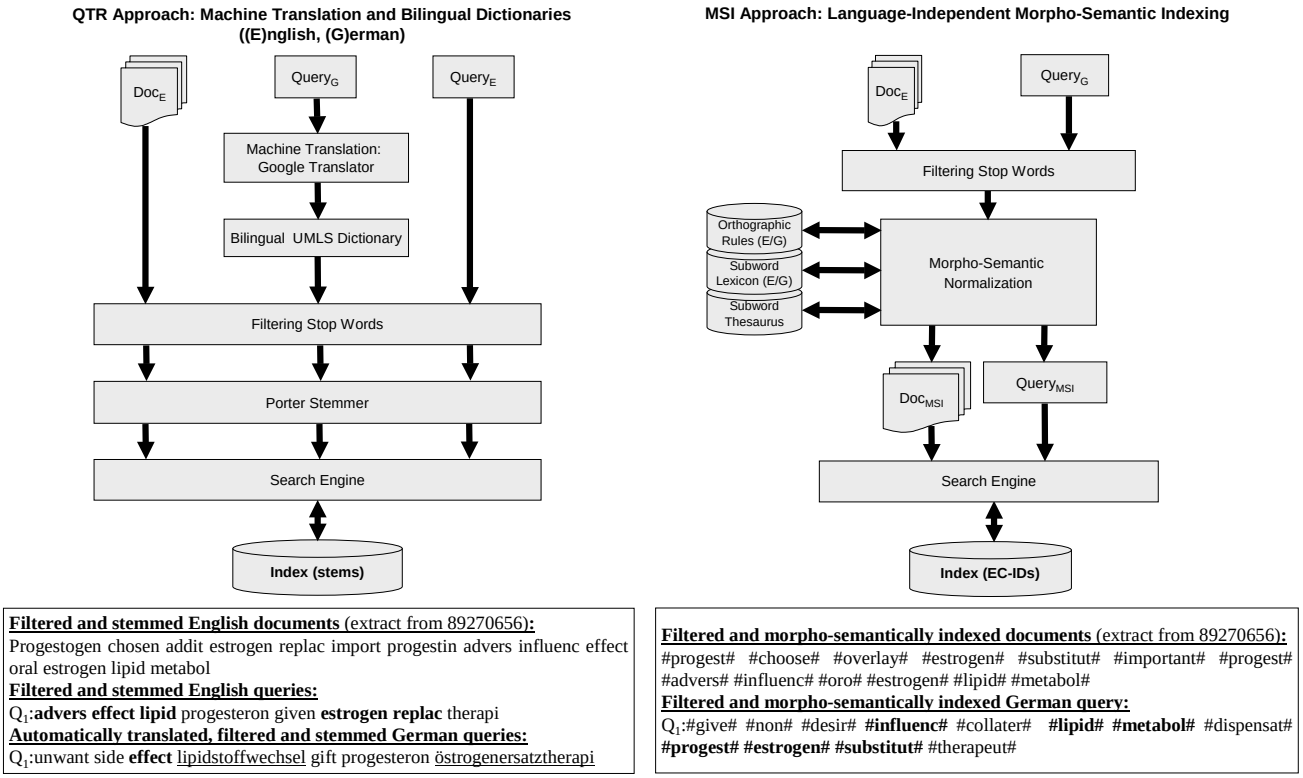

Figure 1: Steps for Automatic Direct Translation (left) and Morpho-Semantic Indexing (right)

In our experiments, the original English queries were first translated into German by medical experts (native speakers of German, with a very good mastery of both general and medical English).[4] In the second step, the manually translated queries were re-translated into English using the GOOGLE TRANSLATOR.[5] Admittedly, this tool may not be particularly suited to translate medical terminology (in fact, 17% of the German query terms were not translated). Hence, we additionally incorporated a bilingual lexeme dictionary derived from the UMLS Metathesaurus with about 26,000 German-English entries. If no English correspondence could be found, the terms were left untranslated (this, finally, happened to 7% of the German query terms). Just as in the baseline condition, the stop words were removed from both the documents and the automatically translated queries. The left side of Figure 1 visualizes this approach which we refer to as QTR.

Alternatively, we probed the MSI approach. Unlike QTR, the indexing of documents and queries using MSI (after stop word elimination) yields a language-independent, semantically normalized index format. The right side of Figure 1 visualizes the basic computation steps for MSI.

Search Engine

For an unbiased evaluation of both approaches, we basically used a simple Boolean search mechanism. The ranking algorithm keeps three different scores for each document. The first is the number of the different query terms, q_1, q_2, ..., q_n, that match the document. For each query containing n *OR*-ed terms, a document is assigned to one of $n + 1$ relevance levels (representing zero to n matches). The second score determines, within each relevance level, the absolute number of all occurrences of matching query terms within the document. Third, building on experience from previous experiments, we incorporated an adjacency constraint as an additional ranking criterion. The corresponding score is based on moving a window of size n (the total number of terms in the query) over the entire document. For each window starting at a document term which matches a query term, the extension of the window is searched for yet another matching query term. If there is a match, the 'window' score is incremented by '1', iterating over the entire window size. The window will then be moved to the next matching document term and the scoring procedure is started, again. The effect of the 'window' scoring method is to significantly increase the score of clustered matches. This allows further refinements in ranking, which become particularly important in the segmentation of complex word forms.[6]

[4] The (human or machine) translation of native-language queries into the target language of the document collection to be searched is a standard experimental procedure in the cross-language retrieval community (Eichmann, Ruiz, & Srinivasan 1998).

[5] `http://www.google.de/language_tools`, last visited on October 16, 2003.

[6] Otherwise, a document containing *'append⊕ectomy'* and *'thyroid⊕itis'*, and another one containing *'append⊕ic⊕itis'* and *'thyroid⊕ ectomy'* become indistinguishable after segmentation.

Retrieval Experiments

Three different test scenarios will be distinguished for our retrieval experiments:

- **BASE**line: Our baseline is given by both the Porter-stemmed English queries and the Porter-stemmed OHSUMED document collection.
- **QTR**: In this condition set, German queries are automatically translated into English ones (using the GOOGLE TRANSLATOR and the UMLS Metathesaurus), which are Porter-stemmed after the translation. These queries are evaluated on the Porter-stemmed document collection.
- **MSI**: This condition captures the automatic transformation of German queries into the language-independent MSI code schema (plus lexical remainders). The entire OHSUMED document collection is also submitted to the MSI code generation procedure. Finally, the MSI-coded queries are evaluated on the MSI-coded document collection, both at an interlingual representation level.

We use three types of measurement in comparing the performance of QTR and MSI. The first one (cf. Table 2 and its visualization in Figure 2) is the average of the precision values at all eleven standard recall points (0.0, 0.1, 0.2, ..., 1.0). We also calculate the average at the top two recall points (0.0 and 0.1). While this data was computed with consideration to the first 200 documents under each condition, we also calculated the exact precision scores for the top k ranked documents, for various $k \leq 100$ (cf. Table 3).

In Table 2, it is not surprising that the English-English baseline performs best with an 11pt average of 0.14. The German-English MSI result is almost on a par with the baseline (0.01 less (0.13)), whereas the German-English QTR result is more than 0.05 points worse (0.09). This means that the MSI approach achieved 93% of the baseline performance (quite a high score given CLIR standards), whereas the QTR approach scored far lower (62%), resulting in a 31 percentage points difference between the two approaches.

In any case, it seems worth noting that at *no* recall point QTR values were higher than MSI values (cf. also Figure 2). The recall points at which the MSI approach performed significantly better than the QTR approach (at a significance level of 0.05 according to the Wilcoxon two-sided paired signed rank test) are marked with an asterisk in Figure 2. For German-English 7 out of 11 measurement points (0.1 through 0.7) are significant.

	English	German	
	BASE	QTR	MSI
0.0	.4000	.2667	.3429
0.1	.3228	.1922	.3046
0.2	.2282	.1448	.2237
0.3	.1779	.1176	.1618
0.4	.1425	.0926	.1306
0.5	.1174	.0668	.1039
0.6	.0829	.0380	.0747
0.7	.0394	.0184	.0437
0.8	.0155	.0098	.0181
0.9	.0041	.0013	.0128
1.0	.0000	.0000	.0071
11pt avr	.1392	.0862	.1294
top 2 avr	.3614	.2294	.3237
% BASE	100.0	61.9	**93.0**

Table 2: Standard 11pt Precision/Recall Table

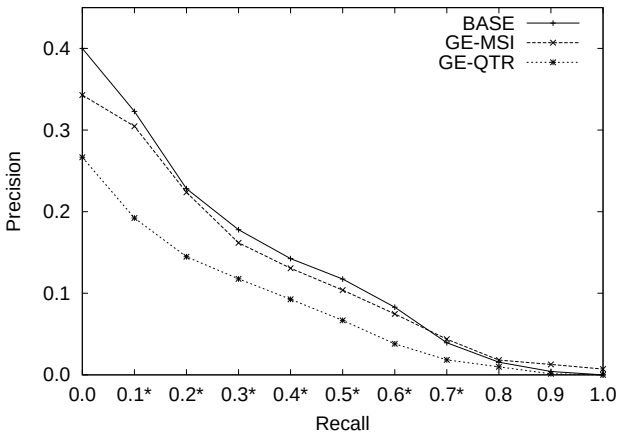

Figure 2: Precision/Recall Graph for German-English

Interesting from a realistic retrieval perspective is the average gain at the top two recall points. In Table 2, the German-English condition yields a precision value of 0.32 (90% of the baseline) for MSI. However, there may be considerable variation regarding the actual numbers behind these levels of recall.

Standard Web users or medical decision-makers under time pressure are more often interested in a few top-ranked documents. Thus, the exact precision scores for these documents are more indicative of the performance of the two approaches in a standard Web or medical retrieval context. Table 3 reveals that at almost all document cut-off points (we examined k = 5, 10, 15, 20, 30, 50, 100), the German-English MSI approach exceeds the others by more than 80% of the English baseline and thus improves, on the average, 20 percentage points over the QTR approach.

top	English	German	
	BASE	QTR	MSI
5	.3429	.2133 (62.2%)	.2952 (86.1%)
10	.2867	.1886 (65.8%)	.2495 (87.0%)
15	.2622	.1638 (62.5%)	.2165 (82.6%)
20	.2405	.1490 (62.0%)	.1919 (79.8%)
30	.2057	.1302 (63.3%)	.1717 (83.5%)
50	.1623	.1072 (66.1%)	.1411 (86.9%)
100	.1158	.0726 (62.7%)	.1038 (89.6%)

Table 3: Exact Precision Scores for Top k Documents

Related Work

The field of cross-language information retrieval (CLIR) is divided into dictionary-based *vs.* corpus-based approaches (Oard & Diekema 1998). Commercial or public-domain translation programs are usually marked by poor support of

technical sublanguages. McCarley (1999) reports on a translation model, which incorporates both query and document translation and outperforms either translation direction. His approach, however, depends on the availability of large parallel corpora, which is mostly not the case for technical sublanguages. Therefore, dictionary-based approaches are favored for domain-specific CLIR. For medical terminology, as well as for other sublanguages, non-specialized multilingual lexicons (based on WORDNET) also offer limited support only (Gonzalo, Verdejo, & Chugur 1999). Hence, we were faced with the need to construct a multilingual medical lexicon from scratch.

The success of dictionary-based CLIR largely depends on the coverage of the lexicon, tools for conflating morphological variants, phrase and proper name recognition as well as word sense disambiguation (Pirkola *et al.* 2001). We optimize for lexical coverage by designing the lexicon around semantically relevant subwords of the medical domain. This also helps us in dealing with morphological variation, including single-word compounds. The latter is a very common phenomenon, especially in German medical terminology (Schulz & Hahn 2000) and cannot be sufficiently treated by dictionary-free techniques (Savoy 2002). This partially explains the poor results for German in the SAPHIRE medical text retrieval system which uses the UMLS Metathesaurus for semantic indexing (Hersh & Donohoe 1998).

The UMLS, together with WORDNET, is also the lexical basis of the approach pursued by the MUCHMORE project (Volk *et al*. 2002). Here, concept mapping occurs after various steps of linguistic pre-processing, including lemmatization. Although very good results are communicated, these are not comparable to ours because the authors use a homegrown document and query collection, as well as (not so reasonable) baselines diverging from ours.

Eichmann, Ruiz, & Srinivasan (1998) report on CLIR experiments for French and Spanish using the same test collection as we do (OHSUMED), and the UMLS Metathesaurus for query translation, achieving 71% of the baseline for Spanish and 61% for French. With the vector space engine they employ, their overall 11pt performance (0.24) is far above the one for the Boolean search engine (0.14) we use. This, however, does not compromise our results, because our experiments are aimed at comparing the performance of two different CLIR methods and not at comparing different search engine architectures. Moreover, our Boolean search engine is more in line with current clinical and Web retrieval engines and the requirements they have to fulfil.

Conclusions

We presented an interlingua approach to cross-language retrieval on a medical document collection. It uses a lexicon composed of equivalence classes of subwords which capture intra- and interlingual synonymy. Compared with direct translation in which queries are automatically translated by online translators, the interlingua approach outperformed the direct approach reaching 93% of the English baseline.

Acknowledgments. This research is supported by the grant *KL 640/5-1* from *Deutsche Forschungsgemeinschaft (DFG)*.

References

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*, 72–80.

Gonzalo, J.; Verdejo, F.; and Chugur, I. 1999. Using EUROWORDNET in a concept-based approach to cross-language text retrieval. *Applied Artificial Intelligence* 13(7):647–678.

Hersh, W., and Donohoe, L. 1998. SAPHIRE International: A tool for cross-language information retrieval. In *Proceedings 1998 AMIA Annual Fall Symposium*, 673–677.

Hersh, W.; Buckley, C.; Leone, T.; and Hickam, D. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM SIGIR Conference*, 192–201.

Hersh, W. R. 2002. *Information Retrieval. A Health and Biomedical Perspective*. New York: Springer, 2nd edition.

Kobayashi, M., and Takeda, K. 2000. Information retrieval on the Web. *ACM Computing Surveys* 32(2):144–173.

Manning, C., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Markó, K.; Daumke, P.; Schulz, S.; and Hahn, U. 2003. Cross-language MESH indexing using morpho-semantic normalization. In *Proceedings of the 2003 Symposium of the AMIA*, 425–429.

McCarley, J. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings 37th Annual Meeting of the ACL*, 208–214.

Oard, D., and Diekema, A. 1998. Cross-language information retrieval. In *Annual Review of Information Science and Technology, Vol. 33*. Information Today. 223–256.

Pirkola, A.; Hedlund, T.; Keskustalo, H.; and Järvelin, K. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval* 4(3/4):209–230.

Porter, M. 1980. An algorithm for suffix stripping. *Program* 14(3):130–137.

Ruiz, M.; Diekema, A.; and Sheridan, P. 1999. CINDOR conceptual interlingua document retrieval: TREC-8 evaluation. In *Proceedings of the 8th Text REtrieval Conference*.

Savoy, J. 2002. Recherche d'information dans des corpus plurilingues. *Ingénierie des Systèmes d'Information* 7(1/2):63–92.

MESH. 2001. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

UMLS. 2003. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

Schulz, S., and Hahn, U. 2000. Morpheme-based, cross-lingual indexing for medical document retrieval. *International Journal of Medical Informatics* 59(3):87–99.

Volk, M.; Ripplinger, B.; Vintar, S.; Buitelaar, P.; Raileanu, D.; and Sacaleanu, B. 2002. Semantic annotation for concept-based cross-language medical information retrieval. *Intl. J. of Medical Informatics* 67(1/3):79–112.