Subword Clusters as Light-Weight Interlingua for Multilingual Document Retrieval

Udo HAHN
Jena University
Language & Information Engineering Lab
Fürstengraben 30
D-07743 Jena, Germany
hahn@coling.uni-freiburg.de

Kornél MARKÓ **Stefan SCHULZ**
Freiburg University Hospital
Medical Informatics Department
Stefan-Meier-Str. 26
D-79104 Freiburg, Germany
marko@coling.uni.freiburg.de

Abstract

We introduce a light-weight interlingua for a cross-language document retrieval system in the medical domain. It is composed of equivalence classes of semantically primitive, language-specific subwords which are clustered by interlingual and intralingual synonymy. Each subword cluster represents a basic conceptual entity of the language-independent interlingua. Documents, as well as queries, are mapped to this interlingua level on which retrieval operations are performed. Evaluation experiments reveal that this interlingua-based retrieval model outperforms a direct translation approach.

1 Introduction

Medical document retrieval presents a unique combination of challenges for the design and implementation of retrieval engines. Clinical document collections have increasingly become available in electronic form (e.g., as Electronic Patient Records (EPRs)) and their size is rapidly growing, with estimates ranging, for a single clinical site, on the order of millions of documents in total, including hundreds to thousands new documents being added every day. Hence, there is a growing demand for automatic support for content-oriented access and browsing of electronic files and EPRs.

Furthermore, medical document collections are inherently *multi-lingual*. While clinical texts are usually written in the native language of the country, searches in major bibliographic databases (e.g., MEDLINE) require a substantial proficiency of (expert-level) English medical terminology. Non-native speakers of English, however, often lack this particular competence. Hence, some sort of bridging between synonymous or, at least, related terms from different languages has to be provided to make full use of the information these databases hold.

Finally, 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 diversified crowd of searchers. Hence, mappings between different linguistic *registers* are inevitable to serve the needs of such a heterogeneous search community. Therefore, automatically performed intra- and interlingual lexical mappings and transformations of equivalent expressions become an obvious necessity to support these different user groups in an adequate manner.

We here propose an approach which is intended to meet these particular challenges. At its core lies a new type of interlingua the basic entities of which are composed of semantically minimal *subwords*. From a linguistic perspective, subwords are often closer to formal Porter-style stems (Porter, 1980) rather than to lexicologically orthodox basic forms, e.g., of verbs or nouns or linguistically plausible stems. Hence, their merits have to be shown in (retrieval) experiments. These language-specific subwords form semantically defined equivalence classes which capture intralingual as well as interlingual (near) synonymy between all subwords in a single cluster. Thus, they abstract away from subtle particularities within and between languages. We do not claim to cover the lexicon of general language but rather restrict ourselves to the terminology used in the medical domain.

In Section 2, we elaborate on the lexicological foundation of this interlingua, i.e., the format of subwords and their synonymy relations, and their role in the process of morphosemantic normalization. The usefulness of subwords will be shown in retrieval experiments (Section 3), in which we contrast our interlingua-based retrieval approach to one which relies on direct translation only (Section 4).

2 Light-Weight Interlingua

We here introduce the notion of subwords (Section 2.1), their organization in terms of an interlingua (Section 2.2), some principles underlying the creation and maintenance of the lexicon as well as the interlingua resource (Section 2.3), and the basic procedure for morphosemantic analysis (Section 2.4).

2.1 Subwords

From a linguistic perspective, the proper choice of the granularity of the basic lexical units is usually guided by syntactic considerations, i.e., the syntax of words (e.g., inflection or derivation) or the syntax of sentences (e.g., in terms of subcategorization or valency frames). For the proper choice of subwords, however, semantic considerations are key. Especially in scientific and technical sublanguages, we observe that semantically non-decomposable entities and domain-specific suffixes (e.g., *'-itis'* (Pacak et al., 1980)) are chained in complex word forms such as in *'pseudo⊕hypo⊕para⊕thyroid⊕ism'*, *'pancreat⊕itis'* or *'gluco⊕corticoid⊕s'*.[1] We refer to these self-contained, semantically minimal units as *subwords* and motivate their status primarily by their usefulness for document retrieval rather than by linguistic arguments.

The minimality criterion is often weaker than, e.g., for morphemes, though it is hard to define in a general way. For example, given the text token *'diaphysis'*, a linguistically plausible morpheme-style segmentation might lead to *'dia⊕phys⊕is'*. From a medical perspective, however, a segmentation into *'diaphys⊕is'* seems much more reasonable because the canonical linguistic decomposition is far too fine-grained and likely to create too many subword ambiguities (which would be harmful to precision). Comparable 'low-level' segmentations of semantically unrelated tokens such as *'dia⊕lyt⊕ic'*, *'phys⊕iol⊕ogy'* lead to morpheme-style subwords *'dia'* and *'phys'*, which unwarrantedly match *'dia⊕phys⊕is'*, too. The (semantic) self-containedness of the chosen subword is also often supported by the existence of a synonym, e.g., for *'diaphys'* we have *'shaft'*.

2.2 From Subwords to Interlingua

Subwords are assembled in a lexical repository, with the following considerations in mind:

- Subwords are listed, together with their attributes such as language (English, German, Portuguese, Spanish) or subword type (stem, prefix, suffix, invariant). Each subword is assigned one or more <u>m</u>orpho-semantic class <u>id</u>entifier(s), we call *MID*(s), representing the corresponding synonymy equivalence class.
- Intralingual synonyms and interlingual translation synonyms of subwords are assigned the same equivalence class (judged within the context of medicine only).
- Two types of meta relations can be asserted between synonymy classes:
  (i) a paradigmatic relation *has-meaning*, which relates one ambiguous class to its specific readings, as with:
  {*head*} ⇒ {*kopf,zephal,caput,cephal,cabec,cefal*} *OR* {*boss,leader,lider,chefe*}.
  (ii) a syntagmatic relation *expands-to*, which consists of predefined segmentations in case of utterly short subwords, such as:
  {*myalg*} ⇒ {*muscle,muskel,muscul*} ⊕ {*pain, schmerz,dor*}.

Compared with relationally richer, e.g., WORDNET based, interlinguas used for cross-language information retrieval (Gonzalo et al., 1999; Ruiz et al., 1999), we hence incorporate a much more limited set of semantic relations and pursue a more restrictive approach to synonymy. We also refrain from introducing additional hierarchical relations between MIDs because such links can be acquired from domain-specific vocabularies, e.g., the Medical Subject Headings (MeSH, 2004) (cf. experimental evidence from Markó et al. (2004)).

2.3 Engineering the Lexicon and Interlingua

In the development workflow, the effects of changes of subword size and granularity are immediately fed back to the developers using word lists to test and validate both the segmentation and the assignment of MIDs. A collection of parallel texts (abstracts of medical publications in English plus either German, Spanish or Portuguese) are used to detect errors in the assignment of MIDs. 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, such as to add *'-otomy'* as a synonym of *'-tomy'* in order to block erroneous segmentations such as *'nephrotomy'* into *'nephr⊕oto⊕my'*;
- 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 equivalence classes are formed, e.g., by grouping {*'hyper-'*, *'high'*, *'elevate'*} into the same class;

[1] '⊕' denotes the concatenation operator.

- The decision which function words and affixes are excluded from indexing, such as *‘and’*, *‘-ation’*, *‘-able’*, and those which are not *‘dys-’*, *‘anti-’*, *‘-itis’*.

In the meantime, the entire subword lexicon (as of July 2005) contains 68,615 entries, with 22,312 for English,[2] 23,600 for German, 14,892 for Portuguese, and 7,811 for Spanish. All of these entries are related in the thesaurus by 20,916 equivalence classes. We also found a well-known logarithmic growth behavior as far as the increase of the number of subwords are concerned (Schulz and Hahn, 2000). Under this observation, at least the English and German subword lexicons have already reached their saturation points.

Our project started from a bilingual German-English lexicon, while the Portuguese part was added in a later project phase (hence, its size still lags somewhat behind). All three lexicons and the common thesaurus structure were manually constructed, which took us about five person-years. While we simultaneously experimented with various subword granularities as well as weaker and stronger notions of synonymy, this manual approach was even heuristically justified. With a much more stable set of criteria for determining subwords emerging from these experiments, we recently switched from a manual to an automatic mode for lexicon acquisition. The Spanish sublexicon, unlike all other previously built sublexicons, was the first one generated solely by an automatic learning procedure. It makes initial use of cognate relations that can be observed for typologically related languages (Schulz et al., 2004) and has recently been embedded into a bootstrapping methodology which induces new subwords that cannot be found by considering merely cognate-style string similarities. This extended acquisition mode makes heavy use of contextual co-occurrence patterns in comparable corpora (Markó et al., 2005b).

2.4 Morphosemantic Processing

Figure 1 depicts how source documents are converted into an interlingual representation by a three-step procedure. We start with **orthographic normalization**. A preprocessor reduces all capitalized characters from input documents to lower-case characters and, additionally, performs language-specific character substitutions[3] to ease the matching of (parts of) text tokens and entries in the dictionary.

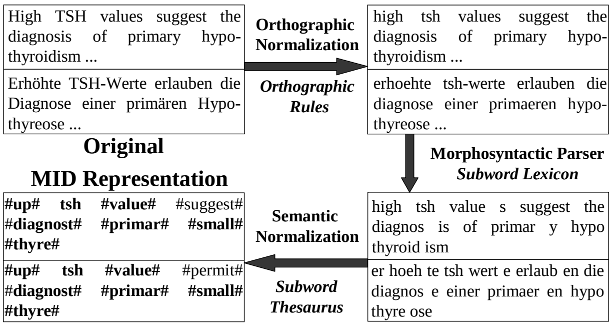

High TSH values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypo-thyreose ...
Orthographic Normalization
Orthographic Rules
high tsh values suggest the diagnosis of primary hypo-thyroidism ...
erhoehte tsh-werte erlauben die diagnose einer primaeren hypo-thyreose ...
Morphosyntactic Parser
Subword Lexicon
Original
MID Representation
#up# tsh #value# #suggest# #diagnost# #primar# #small# #thyre#
#up# tsh #value# #permit# #diagnost# #primar# #small# #thyre#
Semantic Normalization
Subword Thesaurus
high tsh value s suggest the diagnos is of primar y hypo thyroid ism
er hoeh te tsh wert e erlaub en die diagnos e einer primaer en hypo thyre ose

Figure 1: Morphosemantic Processing Pipeline

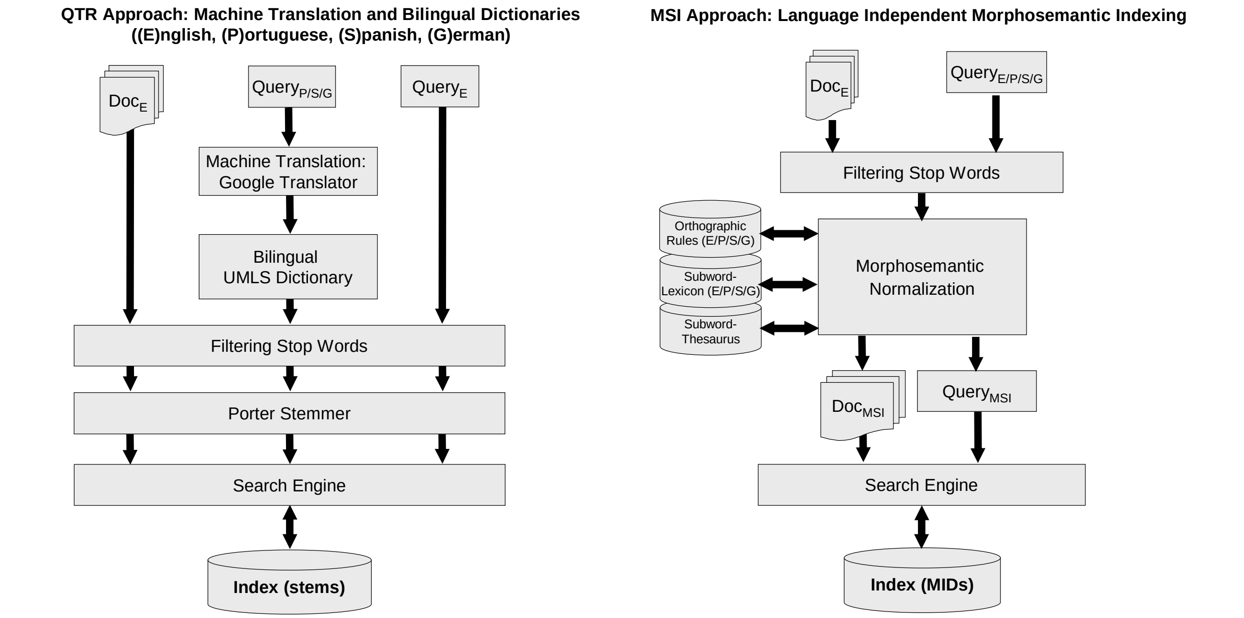
The next step in the pipeline is concerned with **morphosyntactic parsing**. The parser segments the orthographically normalized input stream into a sequence of subwords as found in the lexicon. The segmentation results stored in a chart are checked for morphological plausibility using a finite-state 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, all extracted stems of four characters or longer – if available – are preserved and the remaining fragments are discarded. Otherwise, if no stem longer than four characters can be determined during the segmentation, we recover the original word. This method was useful for the preservation of proper names, although a dedicated name recognizer is still a desideratum for our system.

In the final step, **semantic normalization**, each subword recognized is substituted by its corresponding MID. After that step, all synonyms within a language and all translations of semantically equivalent subwords from different languages are represented by the same MID.

Composed terms (such as *‘myalg⊕y’*) which are linked to their components by the *expands-to* relation are substituted by the MIDs of their components, in the same way as if this were performed by the parser. Ambiguous classes, i.e., those related by a *has-meaning* link to two or more classes, produce a sequence of their related MIDs (for interlingua-based disambiguation, cf. Markó et al. (2005a)).

[2] 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://wordnet.princeton.edu/man/wnstats.7WN`, last visited on May 13, 2005). Linguistically speaking, the entries are basic forms of verbs, nouns, adjectives and adverbs.

[3] For German, e.g., *‘ß’* → *‘ss’*, *‘ä’* → *‘ae’*, *‘ö’* → *‘oe’*, *‘ü’* → *‘ue’* and for Portuguese *‘ç’* → *‘c’*, *‘ú’* → *‘u’*, *‘õ’* → *‘o’*.


QTR Approach: Machine Translation and Bilingual Dictionaries ((E)nglish, (P)ortuguese, (S)panish, (G)erman)
Doc_E
Query_P/S/G
Query_E
Machine Translation: Google Translator
Bilingual UMLS Dictionary
Filtering Stop Words
Porter Stemmer
Search Engine
Index (stems)
MSI Approach: Language Independent Morphosemantic Indexing
Doc_E
Query_E/P/S/G
Filtering Stop Words
Orthographic Rules (E/P/S/G)
Subword-Lexicon (E/P/S/G)
Subword-Thesaurus
Morphosemantic Normalization
Doc_MSI
Query_MSI
Search Engine
Index (MIDs)

Filtered and stemmed English documents (extract from 89270656)**:**
Progestogen chosen addit estrogen replac import progestin advers influenc effect oral estrogen lipid metabol
Filtered and stemmed English queries:
Q_1:**advers effect lipid** progesteron given **estrogen replac** therapi
Automatically translated, filtered and stemmed Portuguese queries:
Q_1:**advers effect lipid** exist progesteron given togeth spare therapi **estrogen**
Automatically translated, filtered and stemmed Spanish queries:
Q_1:**advers effect** exist **lipid** progesteron given **estrogen** therapi availabl
Automatically translated, filtered and stemmed German queries:
Q_1:unwant side **effect** lipidstoffwechsel gift progesteron östrogenersatztherapi

Filtered and morphosemantically indexed documents (extract from 89270656)**:**
#progest# #choose# #overlay# #estrogen# #substitut# #important# #progest# #advers# #influenc# #oro# #estrogen# #lipid# #metabol#
Filtered and morphosemantically indexed English query:
Q_1:**#advers# #influenc# #lipid# #progest#** #give# **#estrogen# #substitut#** #therapeut#
Filtered and morphosemantically indexed Portuguese query:
Q_1:#exist# **#influenc# #advers# #lipid# #progest#** #give# #linkag# #therapeut# **#substitut#** #rek# #posit# **#estrogen#**
Filtered and morphosemantically indexed Spanish query:
Q_1: **#influenc# #advers# #lipid# #progest#** #give# #therapeut# **#substitut# #estrogen#**
Filtered and morpho-semantically indexed German query:
Q_1: #give# #non# #desir# **#influenc#** #collater# **#lipid# #metabol#** #dispensat# **#progest# #estrogen# #substitut#** #therapeut#

Figure 2: Direct Query Translation (left: QTR) *vs.* Interlingual Morphosemantic Indexing (right: MSI)

3 Experimental Setting

3.1 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. OHSUMED is a subset of the MEDLINE database which contains bibliographic information (author, title, abstract, index terms, etc.) of life science and biomedicine articles. Because we only considered the title and abstract field for each bibliographic unit, we obtained a document collection comprised of 233,445 texts. (115,121 out of all 348,566 documents contain no abstract and were therefore ignored.) Our test collection is made of 41,924,840 tokens, and the average document length is 179.6 tokens (with a standard deviation of 76.4).

Since the OHSUMED corpus was created specifically for IR studies, 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). This is a typical query: *"Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy?"*. In Figure 2, the results 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 the *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 further be enhanced by lexical resources as available from the U.S. National Library of Medicine in support of various non-English languages, e.g. the UMLS Metathesaurus (UMLS, 2004) (which currently supports – with considerable differences in coverage – German, French, Spanish, Portuguese, Russian, and many others). As a matter of fact, this procedure, direct Query Translation (QTR), reduces the cross-language retrieval problem to a monolingual one. As an alternative, we consider the interlingua-based cross-language approach in terms of Morphosemantic Indexing (MSI) as introduced in Section 2. Both approaches will then be evaluated on the same query and document set. As the baseline for our experiments, we provide a retrieval system operating with the Porter stemmer (Porter, 1980) and language-specific stop word lists[4] so that the system runs on (original) English documents with (original) English queries.

The (human or machine) translation of native-language queries into the target language of the document collection to be searched (QTR) is a standard experimental procedure in the cross-language retrieval community (Eichmann et al., 1998). In our experiments, the original English queries were first translated into Portuguese, Spanish and German by medical experts (all native speakers of those languages, with a very good mastery of both general and medical English). 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, 16% of the Portuguese, and 14% of the Spanish query terms were not translated). Hence, we additionally used bilingual lexeme dictionaries derived from the UMLS Metathesaurus with about 26,000 German-English entries, 14,200 entries for Portuguese-English, and 22,900 for Spanish-English. If no English correspondence could be found, the terms were left untranslated (this, finally, happened to 7% of the German, as well as 5% of the Portuguese and Spanish 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 2 visualizes this approach which we refer to as QTR.

As an alternative to QTR, we tested MSI, the approach as described in Section 2. 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 2 visualizes the basic computation steps for MSI.

3.2 Search Engine

For an unbiased evaluation, we ran several experiments with LUCENE,[6] a freely available open-source search engine which combines Boolean searching with a sophisticated ranking model based on TF-IDF. Beside its ranking facility, which achieves results that even can outperform advanced vector retrieval systems (Tellex et al., 2003), LUCENE has another advantage: it supports a rich query language, like multi-field search, and more than ten different query operators. In our experiments we made use of proximity search, which allows to find words within a specified window size. For example, given the query talar fracture$\sim$3, LUCENE finds documents containing the words *'talar'* and *'fracture'* within three words distance of each other and allows word swaps (e.g., *'fracture of the talar bone'*, *'talar bone fracture'*). In previous experiments, we discovered that this feature increases the retrieval performance in any scenario, including the baseline condition. Especially, the effect of considering a window of three items significantly increases the score of clustered matches. This becomes particularly important in the segmentation of complex word forms.[7]

4 Experimental Results

Three different test scenarios can now be distinguished for our retrieval experiments:

- **BASELINE**: The OHSUMED corpus serves as the baseline of our experiments both in terms of its Porter-stemmed English queries and its Porter-stemmed English document collection.
- **QTR**: German, Portuguese and Spanish 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 directly evaluated on the Porter-stemmed OHSUMED document collection.
- **MSI**: German, Portuguese and Spanish queries are automatically transformed into the language-independent MSI interlingua (plus lexical remainders). The entire OHSUMED document collection is also submitted to the MSI procedure. Finally, the MSI-coded

[4] We used the stemmer available on `http://www.snowball.tartarus.org`, last visited on January 2005. The incorporated stop word lists contained 172 English, 232 German, 220 Portuguese, and 329 Spanish entries.

[5] `http://www.google.de/language_tools`, last visited on January 2005.

[6] `http://jakarta.apache.org/lucene/docs/index.html`, last visited on January 2005.

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

queries are evaluated on the MSI-coded OHSUMED document collection, both at an interlingual representation level.

We take several measurements in comparing the performance of QTR and MSI. The first one is the average precision at all eleven standard recall points (0.0, 0.1, 0.2, ..., 1.0). These values are depicted in Figure 3 for all scenarios we considered. We also calculate the average at the top two recall points (0.0 and 0.1). While this data was computed with consideration of the first 200 documents under each condition, we also calculated the exact precision scores for the top five and top 20 ranked documents.

As shown in Table 1 (first row), the English-English baseline reaches 0.2 precision on the 11pt average. We also ran an experiment where we MSI-indexed the original OHSUMED corpus and english queries. This boosted the 11pt average to 0.22. Clearly, this approach cannot be taken as the baseline condition, since it confounds the notion of baseline with that of experimental conditions. It is interesting though, because it reveals some of the potential of MSI for (medical) indexing.

The German-English MSI result is almost on a par with the baseline (0.01 less (0.19)), whereas the German-English QTR result is more than 0.08 points worse (0.12). Hence, the MSI approach achieved 95% of the baseline performance (quite a high score given CLIR standards), whereas QTR scored far lower (60%), resulting in a 35 percentage points difference between the two approaches.

The difference turns out to be less dramatic, but still noticeable, in comparing the Portuguese-English MSI and QTR results with the baseline (78% for MSI and 52% for QTR, hence, 26 percentage points difference). We may speculate that this result is not due to any particularities of the Portuguese language, but rather to the uneven investment of effort in building various lexicons (the size of the Portuguese subword lexicon is only two thirds of the corresponding German and English ones; cf. Section 2). For Spanish, QTR precision averages 40% of the monolingual baseline (0.08), whilst 69% of the baseline is reached for MSI (0.14). This is a significant win given that the Spanish dictionary was built in a fully automatic way.

Interesting from a realistic retrieval perspective is the average gain on the top two recall points. In Table 1 (second row), the German-English MSI condition achieves a precision of 0.41 (92% of the baseline), the Portuguese-English condition yields a precision value of 0.36 (80% of the baseline). For Spanish, still 78% of the monolingual baseline precision is reached.

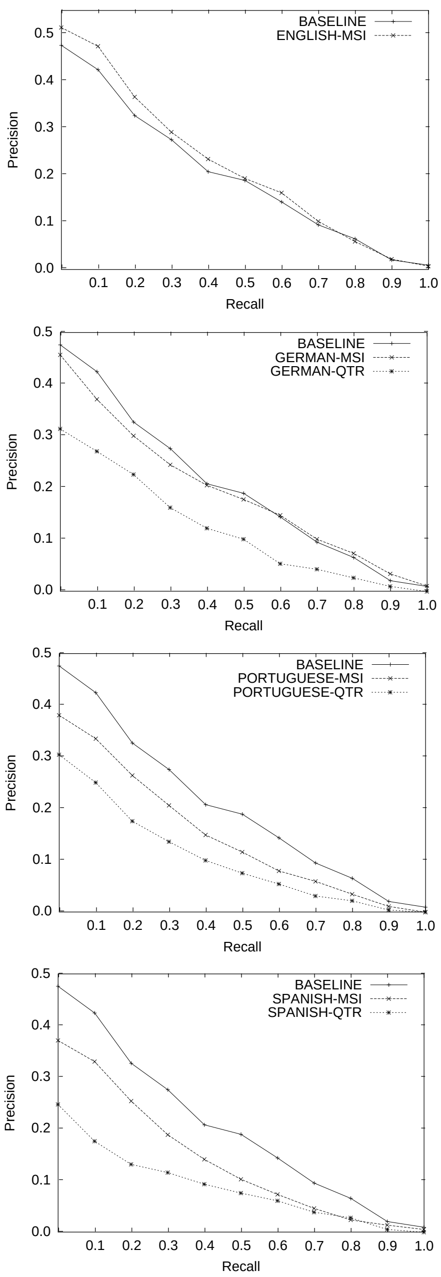


Figure 3: Precision/Recall Graphs for, from top to bottom, the English-English baseline, German-English, Portuguese-English, Spanish-English

Medical decision-makers 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 such a standard medical retrieval context (see Table 1, third and fourth row). MSI exceeds QTR by 12-17 percentage points for German, 9-10 percentage points for Portuguese, and even 21-25 percentage points for Spanish, considering the top 5, respectively top 20, ranked documents.

	English		German		Portuguese		Spanish	
	BASE	MSI	QTR	MSI	QTR	MSI	QTR	MSI
11pt	.2031	.2211 (**108.9**)	.1209 (59.5)	.1930 (**95.0**)	.1051 (51.7)	.1489 (**77.6**)	.0811 (39.9)	.1409 (**69.0**)
top 2pt	.4503	.4944 (**109.8**)	.2926 (65.0)	.4142 (**92.0**)	.2779 (61.7)	.3583 (**79.6**)	.1900 (42.2)	.3509 (**77.9**)
top 5	.7566	.7301 (**96.5**)	.4603 (60.8)	.5528 (**73.0**)	.4396 (58.1)	.5094 (**67.3**)	.3566 (47.1)	.5170 (**68.3**)
top 20	.6033	.6066 (**100.5**)	.3741 (62.0)	.4764 (**79.0**)	.3528 (58.5)	.4118 (**68.3**)	.2726 (45.2)	.4212 (**69.8**)

Table 1: Standard Precision/Recall Table (% of Baseline in Brackets)

5 Related Work

After more than a decade of active research, cross-language information retrieval (CLIR) has made considerable achievements (Grefenstette, 1998; Gey et al., 2002). From a methodological point of view, the field is divided into dictionary-based *vs.* corpus-based approaches (Oard and Diekema, 1998). Since corpus-based approaches depend on the availability of large parallel corpora, which are mostly out of reach for technical sublanguages, most efforts in CLIR are centered around either query translation or document translation (Rosemblat et al., 2003). McCarley (1999) reports on a translation model, which incorporates both query and document translation and outperforms either translation direction. A more recent strategy for machine translation based CLIR is the use of commercial software for query processing (Savoy, 2003), which usually provides only poor support of technical sublanguages. For medical terminology and other sublanguages, non-specialized multilingual lexicons (e.g., based on WordNet) also offer limited support only (Gonzalo et al., 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 the lexical coverage by limiting the lexicon to semantically relevant subwords of the medical domain. This also helps us in dealing with morphological variation, including single-word decomposition. The latter is a very common phenomenon, especially in German medical terminology (Schulz and Hahn, 2000) and cannot be sufficiently treated by dictionary-free techniques (Savoy, 2002). This might explain the poor results for German in the SAPHIRE retrieval system which uses the UMLS Metathesaurus for semantic indexing (Hersh and 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 home-grown document and query collection, as well as baselines diverging from ours.

Eichmann et al. (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 baseline for Spanish and 61% for French. With the vector space engine they employ, their overall 11pt performance (0.24) is slightly above the one for the search engine we use (0.20). This, however, does not compromise our results since our experiments are aimed at comparing the performance of two different CLIR methods and not at comparing different search engine architectures. Moreover, the search engine we employ is more in line with current clinical and Web retrieval engines and the requirements they have to fulfil.

6 Conclusions

We presented an interlingua approach to cross-language retrieval on a medical document collection. It is based on subword clusters, i.e., equivalence classes of subwords which capture intra- and interlingual synonymy.[8] Compared with a direct-translation approach in which queries are translated by online translators, the light-weight interlingua approach fared well. We achieved a remarkable benefit for German document retrieval in terms of 95% reaching the English baseline. The results for Portuguese and Spanish are weaker (78% and 69%, respectively), but this can be attributed to the current underspecification of both lexicons we assume.

7 Acknowledgements

This work was funded by Deutsche Forschungsgemeinschaft (DFG), grant Klar 640/5-2, and the EU Network of Excellence *Semantic Interoperability and Data Mining in Biomedicine* (NoE 507505).

[8] Please, check the MORPHOSAURUS system at `http://www.morphosaurus.net`, our implemetation of the interlingua-based MSI approach.

References

D. Eichmann, M. Ruiz, and P. Srinivasan. 1998. Cross-language information retrieval with the UMLS Metathesaurus. In *SIGIR'98 – Proceedings of the 21st Annual International ACM SIGIR Conference*, pages 72–80. Melbourne, Australia, August 24-28, 1998.

F. Gey, N. Kando, and C. Peters. 2002. Cross-language information retrieval: A research roadmap. *SIGIR Forum*, 36(1):72–80.

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

G. Grefenstette, editor. 1998. *Cross-Language Information Retrieval*. Kluwer.

W. Hersh and L. Donohoe. 1998. SAPHIRE International: A tool for cross-language information retrieval. In *AMIA'98 – Proceedings of the 1998 Annual Fall Symposium.*, pages 673–677. Orlando, FL, November 7-11, 1998.

W. Hersh, C. Buckley, T. Leone, and D. Hickam. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94 – Proceedings of the 17th Annual International ACM SIGIR Conference*, pages 192–201. Dublin, Ireland, 3-6 July 1994.

K. Markó, U. Hahn, S. Schulz, P. Daumke, and P. Nohama. 2004. Interlingual indexing across different languages. In *RIAO 2004 – Conference Proceedings*, pages 82–99. Avignon, France, 26-28 April 2004.

K. Markó, S. Schulz, and U. Hahn. 2005a. Unsupervised multilingual word sense disambiguation via an interlingua. In *AAAI'05 – Proceedings of the 20th National Conference on Artificial Intelligence*, pages 1075–1080. Pittsburgh, PA, USA, July 9-13, 2005.

K. Markó, S. Schulz, A. Medelyan, and U. Hahn. 2005b. Bootstrapping dictionaries for cross-language information retrieval. In *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference*. Salvador, Brazil, August 15-19, 2005.

J. McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the ACL*, pages 208–214. College Park, MD, USA, 20-26 June 1999.

MeSH. 2004. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

D. Oard and A. Diekema. 1998. Cross-language information retrieval. In M. E. Williams, editor, *Annual Review of Information Science and Technology (ARIST), Vol. 33: 1998*, pages 223–256.

M. Pacak, L. Norton, and G. Dunham. 1980. Morphosemantic analysis of *-itis* forms in medical language. *Methods of Information in Medicine*, 19(2):99–105.

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

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

G. Rosemblat, D. Gemoets, A. Browne, and T. Tse. 2003. Machine translation-supported cross-language information retrieval for a consumer health resource. In *AMIA'03 – Proceedings of the 2003 Annual Symposium.*, pages 564–568. Washington, D.C., November 8-12, 2003.

M. Ruiz, A. Diekema, and P. Sheridan. 1999. CINDOR conceptual interlingua document retrieval: TREC-8 evaluation. In *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 597–606. Gaithersburg, MD, November 17-19, 1999.

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

J. Savoy. 2003. Report of CLEF-2003 multilingual tracks. In *Working Notes for the 2003 CLEF Workshop*. Trondheim, Norway, 21-22 August.

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

S. Schulz, K. Markó, E. Sbrissia, P. Nohama, and U. Hahn. 2004. Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *COLING 2004 – 20th International Conference on Computational Linguistics*, pages 813–819. Geneva, Switzerland, August 23-27, 2004.

S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR 2003 – Proceedings of the 26th Annual International ACM SIGIR Conference*, pages 41–47. Toronto, Canada, July 28 - August 1, 2003.

UMLS. 2004. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

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