

Bootstrapping Dictionaries for Cross-Language Information Retrieval

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

Olena Medelyan
Freiburg University Hospital, Germany
Department of Medical Informatics
Stefan-Meier-Str. 26, D-79104 Freiburg
medelyan@coling.uni-freiburg.de

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

Udo Hahn
Jena University, Germany
Language and Information Engineering Lab
Fürstengraben 30, D-07743 Jena
udo.hahn@coling.uni-jena.de

ABSTRACT

The bottleneck for dictionary-based cross-language information retrieval is the lack of comprehensive dictionaries, in particular for many different languages. We here introduce a methodology by which multilingual dictionaries (for Spanish and Swedish) emerge automatically from simple seed lexicons. These seed lexicons are automatically generated, by cognate mapping, from (previously manually constructed) Portuguese and German as well as English sources. Lexical and semantic hypotheses are then validated and new ones iteratively generated by making use of co-occurrence patterns of hypothesized translation synonyms in parallel corpora. We evaluate these newly derived dictionaries on a large medical document collection within a cross-language retrieval setting.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Dictionaries, Thesauruses;
H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Algorithms

Keywords

Cross-Language Information Retrieval, Lexical Acquisition

1. INTRODUCTION

Cross-language information retrieval can broadly be divided into dictionary-based and corpus-based approaches [12] to find relevant documents from different languages. Although dictionaries provide explicit lexical links within and between the natural languages

involved, for large-scale retrieval purposes at least, manually built dictionaries often lack sufficient coverage. Therefore, we propose a mechanism by which comprehensive dictionaries can be automatically set up relying on simple techniques and ready-to-use resources.

Our methodology employs seed lexicons and parallel corpora, thus unifying once competing efforts. The seed lexicons for Spanish and Swedish are automatically generated from (previously manually constructed) Portuguese and German as well as English lexicons. This step relies on cognate mapping, i.e., string-pattern-based transformations of orthographically very similar lexical forms from the source language into the target language (cf. Section 3). This seed is then thrown onto parallel corpora in order to filter out valid lexical and semantic hypotheses. For this step, we focus on co-occurrence patterns of hypothesized translation equivalents in the parallel corpora. Subsequently, valid cognates contribute to further dictionary upgrades by iteratively incorporating non-cognates into the lexical assimilation process (cf. Section 4). To estimate the value of dictionaries derived this way we test their performance on a large medical document collection within a cross-language retrieval setting (cf. Sections 5 and 6).

At the core of this approach lies the MORPHOSAURUS¹ text processing engine (an acronym for MORPHEME THE SAURUS). Its indexing technique is particularly sensitive towards cross-language morpho-semantic regularities (cf. Section 2). The system is centered around a new type of dictionary, in which the entries are subwords, i.e., semantically minimal, morpheme-style units. Language-specific subwords are linked by intralingual as well as interlingual synonymy and grouped in terms of concept-like equivalence classes at the layer of a language-independent interlingua. Our claim that this interlingual approach is useful for the purpose of cross-lingual text retrieval and categorization has already been experimentally supported [7, 11].

2. MORPHO-SEMANTIC INDEXING

Our work starts from the assumption that neither fully inflected nor automatically stemmed words – such as common in many text retrieval systems – constitute the appropriate granularity level for lexicalized content description. Especially in scientific sublanguages,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '05, August 15–19, 2005, Salvador, Brazil.

Copyright 2005 ACM 1-59593-034-5/05/0008 ...\$5.00.

¹<http://www.morphosaurus.net>

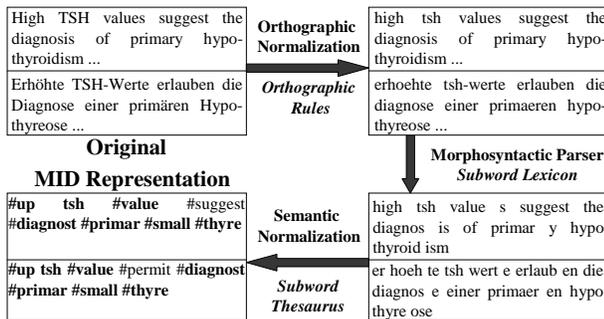


Figure 1: Morpho-Semantic Indexing Pipeline

we observe a high frequency of domain-specific suffixes (e.g., ‘-itis’, ‘-ectomy’ in the medical domain) and numerous occurrences of complex word forms such as ‘*pseudo⊕hypo⊕para⊕thyroid⊕ism*’ or ‘*gluco⊕corticoid⊕s*’.² To properly account for these particularities of ‘medical’ morphology, we introduced the notion of subwords [20] i.e., self-contained, semantically minimal units and motivated their existence by their usefulness for document retrieval rather than by purely linguistic arguments.

Subwords are assembled in a multilingual dictionary and thesaurus, which contain their entries, special attributes and semantic relations between them, according to the following considerations:

- Subwords are listed, together with their attributes such as language (English, German, Portuguese) and subword type (stem, prefix, suffix, invariant). Each lexicon entry is assigned one or more morpho-semantic identifier(s) representing the corresponding synonymy class, the MORPHOSAURUS identifier (MID). 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 a single paradigmatic relation *has-meaning*, which relates one ambiguous class to its specific readings,³ and a syntagmatic relation *expands-to*, which consists of pre-defined segmentations in case of utterly short subwords.⁴

Compared to relationally richer, e.g., WORDNET-based, interlinguas as applied for cross-language information retrieval [6, 17], we, obviously, use a rather limited set of semantic relations and pursue a more restrictive approach to synonymy. In particular, we restrict ourselves to the specific language usage in the context of the medical domain. We also refrain from introducing hierarchical relations between MIDs, because such links can be acquired from domain-specific vocabularies, e.g., the Medical Subject Headings (MESH [18, 11]).

Figure 1 depicts how source documents (top-left) are converted into an interlingual representation by a three-step morpho-semantic indexing procedure. First, each input word is orthographically normalized in terms of lower case characters and according to language-specific rules for the transcription of diacritics (top-right). Next, words are segmented into sequences of subwords or left unchanged when no subwords can be decomposed (bottom-right). The segmentation results are checked for morphological plausibility using

²⊕ denotes the concatenation operator.

³For instance, {*head*} ⇒ {*zephal,kopf,caput,cephal,cabec,cefal*} OR {*leader,boss,lider,chef*}

⁴For instance, {*myalg*} ⇒ {*muscle,muskel,muscul*} ⊕ {*schmerz,pain,dor*}

a finite-state automaton which rejects invalid segmentations (e.g., segmentations without stems or ones beginning with a suffix). Finally, each meaning-bearing subword is replaced by a language-independent semantic identifier, its MID, thus producing the interlingual output representation of the system (bottom-left). In Fig. 1, MIDs which co-occur in both document fragments appear in bold face. A comparison of the original input documents (top-left) and their interlingual representation (bottom-left) already reveals the degree of (hidden) similarity uncovered by the overlapping MIDs.

The manual construction of a trilingual dictionary and the thesaurus has consumed three and a half person years. The combined subword dictionary contains 59,288 entries,⁵ with 22,041 for English, 22,385 for German, and 14,862 for Portuguese. In an effort to further expand the language coverage of MORPHOSAURUS by Spanish and Swedish, we wanted to reuse these already available resources for Portuguese, English, and German in order to speed up and to ease the lexicon acquisition process. This procedure can be divided into three separate steps. First, cognate pairs for similar languages such as Portuguese-Spanish, English-Swedish and German-Swedish are generated. Second, the generated lexical hypotheses are checked for validity considering simple corpus statistics. Next, we use the list of already ‘approved’ cognates in order to augment, step by step, the target dictionaries by processing parallel corpora in terms of co-occurrence patterns of hypothesized translation equivalents which are *not* cognate pairs.

3. GENERATION OF COGNATE PAIRS

Table 1 lists the resources we used for the generation of cognates:

- Manually established PORTuguese, ENGLISH and GERman subword dictionaries, as described in the previous section.
- Manually created lists of SPANish and SWEDish affixes. They were assembled by medical linguists based on introspection and heuristic support from commercial dictionaries.
- Medical corpora for all languages involved, which were automatically compiled exploiting heterogeneous WWW sources.
- Word frequency lists generated from these corpora.

Languages	Seed Lexicon		Corpus	
	Stems	Affixes	Types	Tokens
POR	14,004	858	133,146	13,400,491
GER	21,705	680	17,151	161,952
ENG	21,501	540	11,349	56,317
SPA	-	824	82,431	3,979,051
SWE	-	633	47,823	957,904

Table 1: Resources Used for the Generation of Cognates

3.1 Spanish and Swedish Subword Candidates

It is well known that typologically related languages reveal similarities both at the lexical and grammatical level. With these considerations in mind, we pursued the idea that already available resources from one language might be reused to build up corresponding resources for another typologically related one. The language

⁵Just 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.

44 Rules:	POR	SPA
ss → s	fracass	fracas
lh → j	mulher	mujer
+ca → za	cabeça	cabeza
19 Rules:	GER	SWE
ei → e	bein	ben
+aa+ → a	saal	sal
+u+ → ö	brust	bröst
6 Rules:	ENG	SWE
c → k	cramp	kramp
ph → f	phosphor	fosfor
ce → s	iceland	island

Table 2: Some String Substitution Rules

pairs we draw on in our study are Portuguese/Spanish and English/Swedish as well as German/Swedish. From the Portuguese (alternatively, English and German) dictionary, identical and similarly spelled Spanish (Swedish) subword candidates are generated. As an example, the Portuguese word stem ‘*estomag*’ [‘*stomach*’] is identical with its Spanish cognate, while ‘*mulher*’ [‘*woman*’] (Portuguese) is similar to ‘*mujer*’ (Spanish). Similar subword candidates are generated by applying a set of string substitution rules, some of which are listed in Table 2. In total, 44 rules for the Portuguese-Spanish pair, 19 rules for German-Swedish and 6 for English-Swedish had been formulated by medical linguists based on introspection, also using various dictionaries for heuristic guidance. Some of these substitution patterns cannot be applied to starting or ending sequences of characters in the source subword. This constraint is captured by a wildcard (‘+’ in Table 2), which stands for at least one arbitrary character.

Based on these string substitution rules and the already available (Portuguese, English, German) dictionaries, for each entry (excluding affixes) from these sources, all possible Spanish and Swedish variant strings were generated. This led, on the average, to 8.8 Spanish variants per Portuguese subword (ranging from 2.7 for high-frequent four-character words to 355.2 for low-frequent 17-character words). Since the rule set is much smaller for German-Swedish and English-Swedish, their average is far less than for Portuguese-Spanish (cf. Table 3).

All generated Spanish and Swedish variants were subsequently compared with the target language word frequency list, previously compiled from the Spanish and Swedish text corpora. Wherever a (purely formal) prefix string match (in the case of stems) or an exact match (for invariants) occurred, the matching string was listed as a potential Spanish (Swedish) cognate of the Portuguese (alternatively, English and German) subword it originated from. Whenever several substitution alternatives for a source subword had to be considered that particular alternative was chosen which had the most similar lexical distribution in the corpora considered.

Language Pair	String Variants			
	#Variants	4-chars	17-chars	over-all
POR-SPA	123,235	2.7	355.2	8.8
GER-SWE	145,423	2.7	14.6	6.7
ENG-SWE	68,803	1.8	15.3	3.2

Table 3: Variant Generation: For each language pair (first column), the total number of variants is depicted in the second column. Columns three to five show variant averages per length.

Language Pair	Source Lexicon	Selected Cognates	Linked MIDs
POR-SPA	14,004	8,644	6,036
GER-SWE	21,705	4,249	3,308
ENG-SWE	21,501	4,140	3,208
Combined Evidence		6,086	4,157

Table 4: Selected Cognates

Similarity was measured as follows: Let S be the source lexical item, C_S the source language corpus containing n tokens and V_1, V_2, \dots, V_p the hypotheses generated from S that match the target language corpus C_T , containing m tokens. With $f(x, y)$ denoting the frequency of a word x in a corpus y , that particular V_j ($1 \leq j \leq p$) was chosen for which

$$\left| \frac{f(S, C_S)}{n} - \frac{f(V_j, C_T)}{m} \right|$$

was minimal. All other candidates were discarded.

As a result, we obtained a list of tentative Spanish (Swedish) subwords each linked by the associated MID to their grounding source cognate in the Portuguese (alternatively, English and German) dictionary. We refer to these lists of cognate candidates as CC_{SPA} for Spanish and CC_{SWE} for Swedish.

In numbers, starting from 14,004 Portuguese, 21,705 German and 21,501 English subwords (cf. Table 1), a total of 123,235 Spanish subword variants were created using the string substitution rules. For Swedish, 145,423 variants were derived from German, and, additionally, 68,803 from English (cf. Table 3). Matching these variants against the Spanish and Swedish corpora and allowing for a maximum of one candidate per source subword (by, possibly, applying the similarity criterion from above), we identified 8,644 tentative Spanish and (combining English and German evidence) 6,086 tentative Swedish cognates (cf. Table 4). Spanish candidates are linked to a total of 6,036 MIDs from their Portuguese correlates (hence, 2,608 synonym relationships were also hypothesized), whilst Swedish candidates are associated with 4,157 MIDs from their German and English correlates (cf. Table 4).

3.2 Validation Using Parallel Corpora

We took advantage of the availability of large parallel corpora in the biomedical domain in order to identify *false friends*, i.e., similar words in different languages with different meanings. In our experiments, we found, e.g., the Spanish subword candidate *‘*crianz*’ for the Portuguese ‘*crianc*’ [‘*child*’] (the normalized stem of ‘*criança*’). The correct translation of Portuguese ‘*crianc*’ to Spanish, however, would have been ‘*nin*’ (the stem of ‘*niño*’), whilst the Spanish ‘*crianz*’ refers to ‘*criac*’ [‘*breed*’] (stem of ‘*criação*’ in Portuguese).

In order to more adequately deal with such problems we incorporated the *Metathesaurus* of the *Unified Medical Language System* (UMLS) [19]. It contains about one million concepts (from different biomedical terminologies) and over two million terms from various languages. (Quasi-) synonyms and translations can easily be identified as they share the same concept identifier. Terms are phrased as single words but may also surface as (very) complex noun phrases. The latter are especially prevalent in vocabularies for disease encoding and literature indexing. In our framework, the *Metathesaurus* can thus be considered as a valuable parallel corpus. Examples for typical English-Spanish alignments it contains are “*Cell Growth*” aligned with “*Crecimiento Celular*”, or “*Heart transplant, with or without recipient cardiectomy*” aligned with “*Trasplante cardiaco, con o sin cardiectomia en el receptor*”.

Language Pair	Hypotheses	Valid
POR-SPA	8,644	3,230 (37.4%)
GER/ENG-SWE	6,086	1,565 (25.7%)

Table 5: Cognates Matching the UMLS Alignments

We use English as the pivot language for our experiments, since it has the broadest lexical coverage in the UMLS. The linkage to other languages is considerably poorer, both in qualitative as well as quantitative terms. The size of the corpora derived from the linkages of the English UMLS to other languages amounts to 60,526 alignments for English-Spanish and 10,953 alignments for English-Swedish.⁶ In order to determine the false friends in the set of the already generated cognate candidate pairs, CC_{SPA} and CC_{SWE} , the parallel corpora of the aligned UMLS expressions were then morpho-semantically processed as described in Section 2. Whenever the same equivalence class identifier (MID) occurred on both sides after this simultaneous bilingual processing, the appropriate Spanish (Swedish, alternatively) subword entry that led to this particular MID is taken to be a valid entry. We think that this approach is reasonable, since it is quite unlikely that a false friend occurs within the same translation context.

Those cognates which never matched in this validation procedure were rejected from the candidate set. As a result, 37% of the Spanish and 26% of the Swedish hypotheses are kept (cf. Table 5). These then served as the seed dictionary (in the following, $L(0)$) for acquiring additional entries, which are *not* cognates to elements of any of the source dictionaries.

4. INCREMENTAL SUBWORD LEARNING USING PARALLEL CORPORA

The parallel corpora derived from the UMLS and the dictionaries with approved cognates both serve as starting points for an iterative continuation of the lexical acquisition process, as described in Algorithm 1. In order to illustrate this process, assume the Swedish subword ‘*blod*’ is identified as being a true cognate to the English subword ‘*blood*’ (and, therefore, is included in $L(0)$). Then, the yet unknown Swedish word ‘*Blodtryck*’, which has the English translation ‘*blood pressure*’ in the UMLS Metathesaurus, gets segmented into [ST:*blod*|UK:~|SF:~|UK:~], with ST being a marker for a stem, SF for a suffix and UK for an unknown sequence, thus satisfying the condition in line 12 of the algorithm. At the same time, the morpho-semantic normalization of ‘*blood pressure*’ leads to the sequence of MIDs [#blood #tense], whilst the normalization of ‘*Blodtryck*’ leads to [#blood], since ‘*tryck*’ is not yet part of the Swedish dictionary. By comparing these two representations, the condition in line 13 of the algorithm is satisfied, since there is exactly one more MID resulting from English, which cannot be found in the Swedish normalization result. The invalid segmentation is then reconstructed (‘ $t\oplus r\oplus yck$ ’) by eliminating those substrings that led to a matching MID (‘*blod*’) in the aligned unit (‘*Blodtryck*’) (line 15). The supernumerary MID resulting from the English normalization is assigned to that remaining string (line 17 in the algorithm). After processing all UMLS alignments, this new entry is then incorporated in the Swedish dictionary as a stem, resulting in the dictionary $L(1)$ (line 26).

In the next run, in which all UMLS alignments are processed once again, this newly derived dictionary entry may serve for extracting, e.g., the Swedish word ‘*luft*’ with its identifier #aero from the UMLS entry ‘*Air Pressure*’ (English, indexed by [#aero #tense])

⁶We focused only on the so-called *preferred entries*.

linked to ‘*Luftryck*’ (Swedish). When no new entries can be generated using this method (quiescence), the algorithm stops.

Table 6 depicts the growth steps of the target dictionaries for the entire bootstrapping process (new entries in comparison to each previous step, Δ , are in brackets). In the first run, for Spanish, 3,587 new subwords are added to the dictionary which leads to a size of 6,817, including those lexical entries already generated by the cognate identification routines (cf. Table 5). For Swedish, only 759 new subwords were generated in the first step. Remarkably, these entries lead to the acquisition of 1,361 new subwords in the next step. After 14 cycles, learning activities calm down with 7,154 subwords generated for Spanish, while after 9 runs 4,148 dictionary entries for Swedish are acquired.

Dictionary	Spanish (Δ)	Swedish (Δ)
L(0)	3,230	1,565
L(1)	6,817 (3,587)	2,324 (759)
L(2)	7,001 (184)	3,685 (1,361)
L(3)	7,094 (93)	4,013 (328)
L(4)	7,108 (14)	4,119 (106)
L(5)	7,109 (1)	4,136 (17)
L(6)	7,110 (1)	4,142 (6)
L(7)	7,111 (1)	4,147 (5)
L(8)	7,114 (3)	4,148 (1)
L(9)	7,126 (12)	4,148 (0)
...	...	-
L(14)	7,154 (28)	-

Table 6: Dictionary Growth Steps

5. CLIR EXPERIMENTAL FRAMEWORK

Our experiments were run on the OHSUMED corpus [8], 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). The following is a typical query: “*effectiveness of etidronate in treating hypercalcemia of malignancy*”.

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. Such a consideration addresses a realistic scenario, since unlike in sciences with English as a *lingua franca*, medical doctors adhere to their native languages in their academic education and everyday practice. Hence, 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 resources such as the UMLS Metathesaurus introduced in Section 3.2. Relying on the quality of the translation, this procedure then reduces the cross-language

```

1: MSI: morpho-semantic indexing procedure from Section 2 (maps sequences of words to sequences of MIDs and remainders)
2: current ← 0
3: quiescence ← false
4: while not quiescence do
5:   the lexicon for MSI is set to  $L(\textit{current})$ 
6:   the list of new_entries is empty
7:   for all  $AU_i, i \in [1, n]$  (UMLS alignment units) do
8:      $AU_S \leftarrow$  source language part of  $AU_i$ 
9:      $AU_T \leftarrow$  target language part of  $AU_i$ 
10:     $MID_S \leftarrow MSI(AU_S)$ 
11:     $MID_T \leftarrow MSI(AU_T)$ 
12:    if for exactly one word there is an invalid segmentation (checked by the FSA) in  $MID_T$  then
13:      if there is exactly one more MID in  $MID_S$  than in  $MID_T$  then
14:         $mid \leftarrow$  supernumerary MID from  $MID_S$ 
15:         $entry \leftarrow$  restore the invalid segment and remove substrings that led to a matching MID in  $MID_S$  and  $MID_T$ ;
16:        strip off potential suffixes from  $entry$ , if the remaining substring is longer than 4 (thus, avoiding too short entries);
17:        add  $entry$  together with the associated  $mid$  to new_entries
18:      end if
19:    end if
20:  end for
21:  if new_entries is empty then
22:    quiescence ← true
23:  else
24:    current ← current + 1
25:    copy  $L(\textit{current} - 1)$  to  $L(\textit{current})$ 
26:    add all entries from new_entries to the lexicon  $L(\textit{current})$ 
27:  end if
28: end while

```

Algorithm 1: Bootstrapping Algorithm for Lexical Acquisition

retrieval problem to a monolingual one. We take this approach as a frame of reference for the interlingua-based cross-language approach. Both of them 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 [14] and language-specific stop word lists⁷ so that the system runs on (original) English documents with (original) English queries.

The (human or machine) translation of native-language queries (in the following called QTR) into the target language of the document collection to be searched is a standard experimental procedure in the cross-language retrieval community [3]. In our experiments, the original English queries were first translated into Portuguese, German, Spanish, and Swedish by medical experts (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.⁸ 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, while Swedish is not supported at all). 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, 62,687 for Spanish-English, and 7,619 for English-Spanish.⁹ If no English correspondence could be found, the terms were left untranslated (this, finally, happened to 7% of the German, 5% of the Portuguese, 5% of the Spanish query terms and to 85% of the Swedish terms). Just as in the baseline condition, the stop words were removed from both the documents and the automatically translated queries.

⁷We 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, 329 Spanish, and 114 Swedish entries.

⁸http://www.google.de/language_tools, last visited on January, 2005

⁹In contradistinction to the UMLS-derived parallel corpora described in Section 3.2, we here only consider word-to-word translations.

As an alternative to QTR, we probed the morpho-semantic indexing (MSI) 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 (cf. Figure 1).

For an unbiased evaluation, we ran several experiments with LUCENE,¹⁰ a freely available open-source search engine which combines Boolean searching with a sophisticated ranking model based on TF-IDF. Beside its ranking formula, which achieves results that even can outperform advanced vector retrieval systems [21], this search engine has another advantage: it supports a rich query language, like multi-field search, including more than ten different query operators. In our experiments we make use of proximity search, which allows to find words within a specified window size. For example, given the query ‘*talar fracture~3*’, LUCENE finds documents containing the words ‘*talar*’ and ‘*fracture*’ within three words distance to each other and allows word swaps (e.g., “*fracture of the talar bone*”, “*talar bone fracture*”). In previous experiments, we found evidence that this feature increases the retrieval performance in any scenario, including the baseline condition [7]. 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.¹¹

6. EXPERIMENTAL RESULTS

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

- **BASELINE:** The baseline of our experiments is given by the OHSUMED corpus both in terms of its Porter-stemmed English queries, as well as its Porter-stemmed (English) document collection.

¹⁰<http://jakarta.apache.org/lucene/docs/index.html>, last visited on January 2005

¹¹ Otherwise, a document containing ‘*append⊕ectomy*’ and ‘*thyroid⊕itis*’, and another one containing ‘*append⊕ic⊕itis*’ and ‘*thyroid⊕ectomy*’ become indistinguishable after segmentation.

- **QTR:** In this condition set, German, Portuguese, Spanish, and Swedish 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 OHSUMED document collection.
- **MSI:** This condition stands for the automatic transformation of the German, Portuguese, Spanish, and Swedish queries 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 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 of the precision values at all eleven standard recall points (0.0, 0.1, 0.2, ..., 1.0). These values are depicted in Figure 2 for all scenarios considered. 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 five and top 20 ranked documents.

As shown in Table 7 (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. 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 drops by 0.08 points (0.12). This means that the MSI approach achieved 95% of the baseline performance (quite a high score given CLIR standards), whereas the QTR approach scored far lower (60%), resulting in a 35 percentage points difference between the two approaches. This 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). Certainly, this result is not due to any particularities of the Portuguese language, but rather, to the uneven investment of effort in the different lexica (the size of the Portuguese subword dictionary has 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 certainly a significant win given that the Spanish dictionary was built in a fully automatic way.

The Swedish results, however, are far less shiny. MSI achieves 27% of the baseline for the 11pt measurement. Disappointing as this data might be, this can be explained by a lack of data and resources. The Swedish dictionary has less than 60% the size of the Spanish one (cf. Table 6). Moreover, the similarity between Swedish and English or German is much lesser than between Spanish and Portuguese. Data sparseness is even more painful for the QTR condition because we were unable to find a Web tool for the translation of Swedish to English and relied on the English-Swedish UMLS entries only. However, re-considering this outcome, the advantages of the subword approach become even more evident. The Swedish subword dictionary for MSI was solely generated by the automatic morpho-syntactic transformation of the Swedish UMLS entries. On the other hand, the QTR scenario was completely based on the UMLS without any transformation. Since

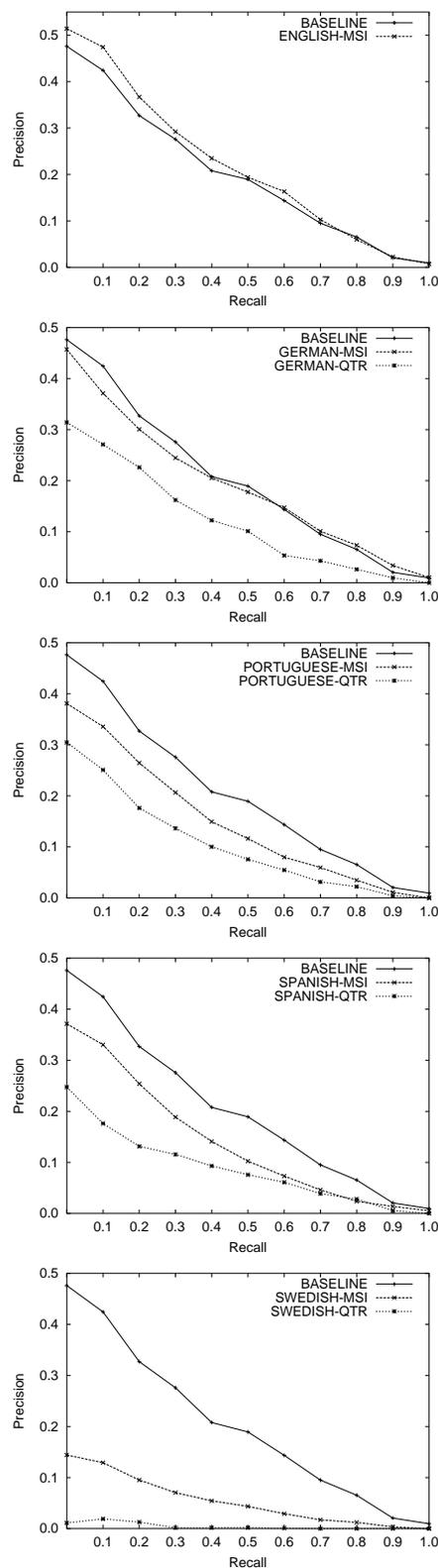


Figure 2: Precision/Recall Graphs for the English-English baseline, (first) German-English (second), Portuguese-English (third), Spanish-English (fourth) and Swedish-English (fifth)

	English		German		Portuguese		Spanish		Swedish	
	BASE	MSI	QTR	MSI	QTR	MSI	QTR	MSI	QTR	MSI
11pt	.203	.221 (108.9)	.121 (59.5)	.193 (95.0)	.105 (51.7)	.149 (77.6)	.081 (39.9)	.141 (69.0)	.004 (2.2)	.054 (26.8)
top 2pt	.450	.494 (109.8)	.293 (65.0)	.414 (92.0)	.278 (61.7)	.358 (79.6)	.190 (42.2)	.351 (77.9)	.015 (3.3)	.137 (30.3)
top 5	.757	.730 (96.5)	.460 (60.8)	.553 (73.0)	.440 (58.1)	.509 (67.3)	.357 (47.1)	.517 (68.3)	.013 (1.7)	.110 (14.5)
top 20	.603	.607 (100.5)	.374 (62.0)	.476 (79.0)	.353 (58.5)	.412 (68.3)	.273 (45.2)	.421 (69.8)	.007 (1.1)	.081 (13.5)

Table 7: Standard Precision/Recall Table (% of Baseline in Brackets)

Swedish (just as German) is a highly agglutinative language, translating subwords rather than complex single-word compounds yields a much higher coverage. Hence, considering 2% of the baseline for Swedish QTR, even 27% constitute a solid figure of merit.

Interesting from a realistic retrieval perspective is the average gain on the top two recall points. In Table 7 (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, whilst Swedish yields 30%.

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 7, third and fourth row). MSI exceeds QTR by 14-15 percentage points for German, 9-10 percentage points for Portuguese, even 21-25 percentage points for Spanish, and 13 percentage points for Swedish, considering the top 5, as well as top 20 ranked documents, respectively.

7. DISCUSSION

The success of dictionary-based CLIR depends on the coverage of the dictionary, tools for conflating morphological variants, phrase and proper name recognition, as well as word sense disambiguation [13]. For medical terminology, as well as for other sublanguages, non-specialized multilingual lexicons (based on WORDNET) or commercial machine translation systems offer limited support only [6, 16]. We optimize the lexical coverage by limiting the dictionary 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 and Swedish (medical) terminology and cannot be sufficiently treated by dictionary-free techniques. This partially explains the poor results for German in the SAPHIRE medical text retrieval system which uses the UMLS Metathesaurus for semantic indexing [9]. The UMLS, together with WORDNET, is also the lexical basis of the approach pursued by the (medical) MUCHMORE project [22]. Here, concept mapping occurs after various steps of linguistic pre-processing, including lemmatization.

Eichmann et al. [3] 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 fulfill.

We consider automatic lexicon acquisition techniques to be a key issue for any sort of dictionary-based efforts in IR, CLIR in partic-

ular. Most approaches to multilingual lexical acquisition employ statistical methods, such as context vector comparison [5, 15, 23, 2] or mutual information statistics [4] and require a seed lexicon of trusted translations. Koehn and Knight [10] derived a seed lexicon from German-English cognates which were selected by using string similarity criteria.

The second issue concerns the processing of suitable corpora. Whilst Widdows et al. [23] deal with parallel German-English corpora to enrich an existing multilingual lexicon (also taken from the UMLS Metathesaurus), Fung and Yee [5], Rapp [15] and Déjean et al. [2] propose methods that require only weaker comparable corpora (cf. also Fung [4] for a linguistically motivated distinction between both types of corpora). Furthermore, Déjean et al. [2] incorporate hierarchical information from an external thesaurus [18] for combining different evidence for lexical acquisition. Cheng et al. [1] as well as Zhang and Vines [24] propose co-occurrence-based methods to automatically extract word translations from mixed-language texts which are dynamically made available through common Web search engines.

Our work differs from these precursors in many ways. First of all, we propose a basically symbolic method for acquiring translations of subwords, instead of using statistics. This is made possible by the availability of relatively large and well aligned parallel corpora, as provided by the UMLS Metathesaurus. Finally, rather than acquiring bilateral word translations, our focus lies on assigning subwords to interlingual semantic identifiers.

8. CONCLUSIONS

We have shown that a significant amount of Portuguese, English and German subwords from the medical domain can be mapped to Spanish and Swedish cognates by simple string transformations. With these seeds, we further enlarge the Spanish and Swedish cognate dictionaries by subwords which are *not* cognates. For the latter task, we used an aligned corpus, the UMLS Metathesaurus, and identified those non-cognates via bootstrapping.

In what concerns the generality of our approach, we rely on large aligned thesauri. Fortunately, large-coverage multilingual thesauri

Thesaurus	# Languages	Subject
Eurovoc	13	European Communities
GEMET	19	activities: science,
UNESCO	3	politics, law, culture,
OECD	4	economics, etc.
Eurodicautom	12	technical terminology
Europ. Education	18	education, teaching,
Europ. Schools	13	individual development
Treasury Browser		research, etc.
AGROVOC	6	agriculture
Astronomy Thes.	5	astronomy

Table 8: Overview of Selected Multilingual Resources (http://sky.fit.qut.edu.au/middletm/cont_voc.html, last visited on January 2005)

are already available for many relevant domains (cf. Table 8), both in terms of the number of languages being covered and the number of alignment units available (e.g., about 5 million for Eurodicautom). Hence, this approach bears high potential for CLIR tasks.

In order to assess the value of what we have done, we tested the usefulness of the newly derived dictionaries on a medical document collection. We achieve a remarkable benefit for German documents by reaching 95% of the English baseline for German. The results for Portuguese are weaker (78%) but this can be attributed to the current underspecification of the Portuguese dictionary. Remarkably, based on fully automatically generated dictionaries, the interlingua approach yields 69% for Spanish, while only modest 27% result for Swedish. Still, the limits we observe can be attributed to the lack of data and resources (corpora, Web translation engines) rather than to the methodology we propose.

9. ACKNOWLEDGMENTS

This work was partly supported by Deutsche Forschungsgemeinschaft (DFG), grant KL 640/5-1 and the European Network of Excellence "Semantic Mining" (NoE 507505).

10. REFERENCES

- [1] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien. Translating unknown queries with web corpora for cross-language information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–153, 2004.
- [2] H. Déjean, É. Gaussier, and F. Sadat. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th Intl. Conf. on Computational Linguistics*, pages 218–224, 2002.
- [3] D. Eichmann, M. E. Ruiz, and P. Srinivasan. Cross-language information retrieval with the UMLS Metathesaurus. In *Proceedings of the 21st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 72–80, 1998.
- [4] P. Fung. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, pages 1–17, 1998.
- [5] P. Fung and L.Y. Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, pages 414–420, 1998.
- [6] J. Gonzalo, F. Verdejo, and I. Chugur. Using EUROWORDNET in a concept-based approach to cross-language text retrieval. *Applied Artificial Intelligence*, 13(7):647–678, 1999.
- [7] U. Hahn, K. Markó, M. Poprat, S. Schulz, J. Wermter, and P. Nohama. Crossing languages in text retrieval via an interlingua. In *RIAO 2004 – Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, pages 100–115, 2004.
- [8] W. R. Hersh, C. Buckley, T. J. Leone, and D. H. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–201, 1994.
- [9] W. R. Hersh and L. C. Donohoe. SAPHIRE International: A tool for cross-language information retrieval. In *Proceedings of the AMIA Annual Fall Symposium*, pages 673–677, 1998.
- [10] P. Koehn and K. Knight. Learning a translation lexicon from monolingual corpora. In *Unsupervised Lexical Acquisition. Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9–16, 2002.
- [11] K. Markó, U. Hahn, S. Schulz, P. Daumke, and P. Nohama. Interlingual indexing across different languages. In *RIAO 2004 – Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, pages 82–99, 2004.
- [12] D. W. Oard and A. R. Diekema. Cross-language information retrieval. In M. E. Williams, editor, *Annual Review of Information Science and Technology (ARIST)*, Vol. 33: 1998, pages 223–256. Medford, NJ: Information Today, 1998.
- [13] A. Pirkola, T. Hedlund, H. Keskustalo, and K. Järvelin. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4(3/4):209–230, 2001.
- [14] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [15] R. Rapp. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, 1999.
- [16] M. Rogati and Y. Yang. Resource selection for domain-specific cross-lingual IR. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, 2004.
- [17] M. Ruiz, A. Diekema, and P. Sheridan. CINDOR conceptual interlingua document retrieval: TREC-8 evaluation. In *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 597–606, 1999.
- [18] MESH. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine, 2004.
- [19] UMLS. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2004.
- [20] S. Schulz, M. Honeck, and U. Hahn. Biomedical text retrieval in languages with a complex morphology. In *Proceedings of the ACL/NAACL 2002 Workshop on 'Natural Language Processing in the Biomedical Domain'*, pages 61–68, 2002.
- [21] S. Tellex, B. Katz, J. J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–47, 2003.
- [22] M. Volk, B. Ripplinger, S. Vintar, P. Buitelaar, D. Raileanu, and B. Sacaleanu. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67(1/3):79–112, 2002.
- [23] D. Widdows, B. Dorow, and C.-K. Chan. Using parallel corpora to enrich multilingual lexical resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 240–245, 2002.
- [24] Y. Zhang and P. Vines. Using the web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169, 2004.