MORPHOSAURUS in ImageCLEF 2006: The effect of subwords on biomedical IR

Philipp Daumke, Jan Paetzold, Kornél Markó
University Hospital Freiburg
Philipp.Daumke@klinikum.uni-freiburg.de

Abstract

We here describe the subword approach we used in the 2006 ImageCLEF Medical Image Retrieval task. It is based on the assupmtion that neither fully inflected nor automatically stemmed words constitute the appropriate granularity for lexicalized content description. We therefore introduce subwords as morphologically meaningful word units. Subwords are organized in language specific lexica that were partly manually and partly automatically generated and currently cover six European languages. They are linked together via a multilingual thesaurus. The use of subwords instead of full words significantly reduces the number of lexical entries that are needed to sufficiently cover a specific language and domain. A further benefit of the approach is its independence from the underlying retrieval system, thus making it usable by any search engine. In this year's test runs we combined MORPHOSAURUS with the open-source search engine *Lucene* and achieved precision gains of up to 25% over the baseline for a monolingual setting and promising results in a multilingual scenario.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing[Dictionaries, Thesauruses]; H.3.3 Information Search and Retrieval[Query formulation, Retrieval models, Search process]; H.3.4 Systems and Software; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Algorithms, Experimentation, Languages

Keywords

Biomedical Information Retrieval, Cross Language Information Retrieval, Stemming, Subwords, MORPHOSAURUS

1 Introduction

The conventional view on human language is word-centered, at least for written language where words are clearly delimited by spaces. It builds on the hypothesis that words are the basic building blocks of phrases and sentences. In syntactic theories, words constitute the terminal symbols. Therefore, it appears straightforward to break down natural langugae to the word level. However, looking at the sense of natural language expressions, evidence can be found that semantic atomicity frequently does not coincide with the word level, which bears methodical challenges even for pretended 'simple' tasks such as tokenization of natural language input [3]. As an example, considering the English noun phrase *“high blood pressure”*, the word limits reflect quite well the semantic composition, whereas this is not the case in its literal translations *“verhoogde bloeddruk”* (Dutch), *“högt blodtryck”* (Swedish) or *“Bluthochdruck”* (German). Especially in technical sublanguages such as the medical one, atomic senses are encountered at different levels of fragmentation or granularity. An atomic sense may correspond to word stems (e.g., *“hepat”* referring to *“liver”*), prefixes (e.g., *“anti-”*, *“hyper-”*), suffixes (e.g., *“-logy”*, *“-itis”*), larger word fragments (*“hypophys”*), words (*“spleen”*, *“liver”*) or even multi-word terms (*“yellow fever”*). The possible combinations of these word-forming elements are immense and ad-hoc term formation is common.

Extracting atomic sense units from texts as a basis for the semantic interpretation of natural language is therefore an important goal in various areas of research dealing with natural language processing, such as information retrieval, text mining or speech recognition. Especially in morphologically rich languages (German, Finnish, Swedish, etc.), where classical rule-based stemmers such as the Porter Stemmer[9] are not particular effective, a deeper morphological analysis is widely acknowledged to improve the retrieval performance in IR systems [4], [8], [1]. However, the lack of accuracy of current unsupervised approaches for morphological analysis hampers the automatic or semi-automatic creation of an underlying knowledge base.

We developed such lexical resource and introduced the notion of subwords [5], i.e., self-contained, semantically minimal units. Language-specific subwords are linked by intralingual as well as interlingual synonymy and grouped together in terms of concept-like language independent equivalence classes.

For ImageCLEFmed 2006, we prove the positive effect of using subwords on the monolingual and multilingual biomedical text retrieval tasks.

2 Morpho-Semantic Indexing

Morpho-semantic indexing translates source documents and queries into an interlingual representation in which their content is represented by language-independent semantic descriptors. This procedure is based upon the MorphoSaurus document pre-processing engine, which consists of orthographic normalization rules, a morphological component for word segmentation, language-specific subword lexicons for each of the natural languages to be analyzed, as well as a language-independent thesaurus.

2.1 Subwords as Document Description Units

In many NLP applications evidence can be found 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 and technical sublanguages, semantically atomic, i.e., non-decomposable entities are chained in complex word forms such as in *’pseudo⊕hypo⊕para⊕thyroid⊕ism’*, *’gluco⊕corticoid⊕s’* or *’pancreat⊕itis’*. Domain-specific suffixes (e.g., *’-itis’*), and single-word compounding are even more accentuated in morphologically richer languages than English, such as German, Finnish, Dutch, Hungarian or Swedish. We refer to these self-contained, semantically minimal units as *subwords* and motivate their existence by their usefulness for document retrieval rather than by linguistic arguments.

The minimality criterion is difficult to define in a general way, and so far is based more on empirical examinations rather than on formal criteria. Considering, e.g., the text token ’diaphysis’, a linguistically plausible morpheme-style segmentation would possibly lead to *’dia⊕phys⊕is’*. From a medical perspective, a 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’ segmentations of semantically unrelated tokens such as *’dia⊕lyt⊕ic’* and *’phys⊕io⊕logy’* also lead to morpheme-style subwords *’dia’* and *’phys’*, and, thus, unwarrantedly match segmentations such as *’dia⊕phys⊕is’*, too. The (semantic) self-containedness of the chosen subword is often supported by the existence of a synonym, e.g., *’shaft’* for *’diaphys’*.

Subwords are assembled in language specific lexicons and a multilingual thesaurus, which together contain subword entries, special subword attributes and semantic relations between subwords. The lexicons and the thesaurus are both constructed manually, with the following considerations in mind:

- Subwords are collected, together with their attributes such as language (English, German), subword type (stem, prefix, suffix, invariant), etc. Each lexicon entry is then assigned to a unique identifier representing one synonymy class, the MorphoSaurus identifier (MID).
- Intralingual synonyms and interlingual translations of subwords are grouped together by the same equivalence class.
- Semantic links between synonymy classes are added. We subscribe to a shallow approach in which semantic relations are restricted to:
  1. a paradigmatic relation *has-sense*, which relates one ambiguous class to its specific readings, e.g.: {head}→({kopf,zephal,cephal}OR{leader, boss}).
  2. a syntagmatic relation *expands-to*, which consists of predefined segmentations in case of utterly short subwords such as {myalg}→{muskel,muscle}⊕{schmerz,pain}.

We refrain from introducing hierarchical relations between MIDs, because such links can be acquired from domain-specific vocabularies, e.g., the Medical Subject Headings (MeSH [8], cf. also [9] for the mapping of MIDs to appropriate MeSH terms).

2.2 Morpho-Semantic Normalization

Figure 1 depicts how source documents (top left) are converted into an interlingual representation by a three-step procedure. The first step deals with orthographic normalization (cf. Figure 1, top right). A preprocessor reduces all capitalized characters from the input documents to lower-case characters and, additionally, performs language-specific character substitutions in order to ease the matching of (parts of) text tokens and entries in the lexicons.

The next step in the pipeline is concerned with morphological segmentation. The system decomposes the orthographically normalized input stream into a sequence of sublexical items, the content-bearing ones correspond to subwords as listed in the lexicon (cf. Figure 1, bottom right). The segmentation proceeds as follows: Each document token t of length n defined as a sequence of characters $c_1, c_2, ..., c_n$ is processed, in parallel, by a forward and backward matching process. The forward matching process starts at the positions 1 and k=n and decrements k iteratively by one unless the sequence $c_1, c_2, ..., c_k$ is found in the subword lexicon. Alternatively, the backward matching process starts at the positions k=1 and n and increments k iteratively by one unless the sequence $c_k, c_{k+1}, ..., c_n$ is found in the lexicon. Substrings recognized this way are entered into a chart. Unless the remaining sequences are empty, $c_{k+1}, c_{k+2}, ..., c_n$, as well as $c_1, c_2, ..., c_{k-1}$ are tested recursively in the same manner, by forward and backward matching, respectively.

The segmentation results that are stored in the chart are checked for morphological plausibility using a finite-state automaton in order to reject invalid segmentations (e.g., those 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, the original word is restituted. This method proved 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 then 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 descriptor in that

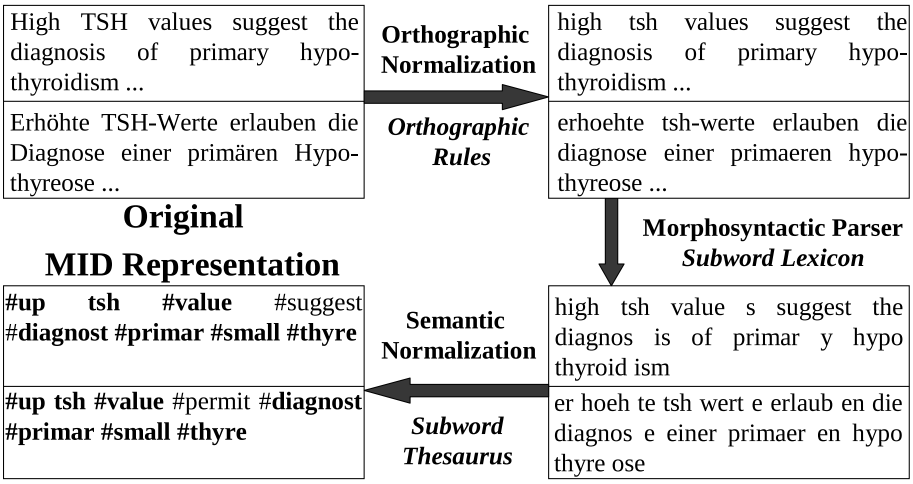
High TSH values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypo-thyreose ...
Orthographic Normalization
Orthographic Rules
high tsh values suggest the diagnosis of primary hypo-thyroidism ...
erhoehte tsh-werte erlauben die diagnose einer 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: Morpho-Semantic Indexing Pipeline

target representation. 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. Ambiguous classes, i.e., those related by a *has-sense* link to two or more classes, are disambiguated by an in-house developed disambiguation tool [6]. The final result is a morpho-semantically normalized document in a language-independent, interlingual representation (cf. Figure 1, bottom left).

2.3 Subword Lexicons and Subword Thesaurus

The process of lexicon construction is a challenging task which requires in-depth knowledge of biomedical terminology. During the development workflow, the effect of lexical changes are immediately fed back to the developers using word lists to test both the segmentation and the assignment of MIDs. Furthermore, a collection of parallel texts (e.g., abstracts of medical publications in English and German) are used to detect errors in the assignment of MIDs. In order to impose common policies on the lexicon builders, we developed a maintenance manual which contains over 30 rules for tasks such as:

- 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., *'neuros⊕is'* instead of *'neur⊕osis'*)
- Data-driven decisions, such as to add *'-otomy'* as a synonym of *'-tomy'* in order to block erroneous segmentations (e.g., *'nephrotomy'* into *'nephr⊕oto⊕my'*)
- The decision to exclude short stems from segmentation (such as *'ov'* in *'ovum'*) in order to block false segmentations; short stems are then treated as invariants so that they cannot be subject to a segmentation - therefore, possible combinations have to be considered as separate lexical units
- The decision to locate the appropriate level of semantic abstraction when defining equivalence classes, e.g., by grouping *hyper-, high, elevated* into the same equivalence 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'*.

2.4 Lexicon Statistics

The manual construction of our lexical resources was done over the last five years with a changing amount of man power. The English, German and Portuguese subword lexicons were created fully

Table 1: Total number of subwords in different languages

$Lang$	$Subwords_{all}$	$Subwords_{auto}$	$EqClasses$	eT	hS
EN	22,706	-	16,668	261	438
GE	24,178	-	16,713	306	448
PT	14,997	-	10,523	286	343
SP	13,060	7,795	9,096	215	254
FR	8,751	3,735	6,006	122	250
SE	13,557	5,946	7,908	182	478
All	97,249	17,476	21,679	497	1,369

Table 2: ($Subwords_{all}$ - number of all subwords, $Subwords_{auto}$ - number of automatically acquired subwords, $EqClasses$ - number of equivalence classes, eT - number of $expands - To$-relations, hS - number of $has - Sense$-relations)

manually, while for French, Spanish and Swedish, additional machine learning techniques were applied in order to bootstrap the lexical resources for these languages[7].

Overall, the lexicons contain 97,249 entries[1] (see Table 2), corresponding to 21,679 equivalence classes. In terms of relations between equivalence classes, there are currently 497 distinct *expands-To* and 1,369 distinct *has-Sense* relations defined in the thesaurus. When medical corpora are indexed by the morpho-semantic indexing routine, an average number of 1.62 MIDs per word is obtained, considering all languages.

3 Experiments

3.1 Lucene Search Engine

In this year's ImageCLEFmed we combined the MorphoSaurus-System with the open-source search engine Lucene,[2] a high performance, scalable, cross-platform retrieval system. It offers a high degree of flexibility by implementing a well-scaling indexing approach and has desirable I/O characteristics for both merging and searching. Lucene supports batch indexing and incremental indexing and a wide range of query features, including full Boolean queries. Its rich query language includes more than ten different query operators and allows multi-field search. Lucene handles adjacency queries and searches multiple indexes at once merging the results to give a meaningful relevance score based on TF-IDF [10]. Its ranking model achieves results that can even outperform advanced vector retrieval systems [11].

3.2 Indexing and Query Preparation Process

In the preparation phase all annotations of the whole dataset containing textual information from Casimage, MIR, PEIR, and PathoPIC datasets were extracted. For all image annotations, we created four Lucene fields which can be queried separately. The first field contained headlines, keywords and additional concise XML tags in the original representation (*original_t*). The second field contained all other free text information that was extracted from the dataset (*original_d*). The third and fourth field contained the Morphosaurus representation of the corresponding first and second field (*mid_t* and *mid_d*). In order to compare the subword approach with traditional automated stemming routines we added two further fields (*stem_t* and *stem_d*) containing the annotations processed by the Porter Stemmer[3]. In addition to these six query fields, a language flag was added for each image annotation (*language*).

[1] Just for comparison, the size of WordNet [2] 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/`). Linguistically speaking, the entries are basic forms of verbs, nouns, adjectives and adverbs.

[2] `http://jakarta.apache.org/lucene/docs/index.html`

[3] We used the stemmer available on `http://www.snowball.tartarus.org` (last visited on July 2006).

In the topics collection, ”Show me”, ”Zeige mir” and ”Montre-moi des” and language specific stopwords[4] were removed. The queries were subsequently transformed into the Morphosaurus interlingua resulting in 30 *“original/mid”* query pairs, each represented in all three languages.

3.3 Monolingual Scenario

One of our goals was to show the effectiveness of the subword approach for monolingual biomedical text retrieval. We therefore decided to create a scenario which makes use of the english subset only. Four different runs were prepared:

1. *Orig-En-En*: As a baseline we took the original english query and searched in the corresponding two Lucene fields which contained the original English annotations (*original_t*, *original_d*, *language* : *en*).
2. Stem-En-En: The Orig-En-En queries were processed by the Porter Stemmer and sent to the corresponding (*stem_t*, *stem_d*)-fields of Lucene.
3. *Mids-En-En*: In this run we took the morpho-semantic normalized form of each English query and searched in the two fields that contained the MID representation of the document annotations (*mid_t*, *mid_d*, *language* : *en*).
4. *Both-En-En*: Here, we combined both the original query and the morpho-semantic normalized query in a simple disjunct fashion. The original words were queried in the original fields (*original_t* and *original_d*, *language* : *en*) and the normalized queries in the MID fields (*mid_t*, *mid_d*, *language* : *en*).

Example Box 1 lists the query syntax of query one in all four test scenarios.

Query 1: “Show me images of the oral cavity including teeth and gum tissue”

Example 1 – Run: Orig-En-En (Baseline), Query: 1
(+language:en +original_t:images) (+language:en +original_t:oral) (+language:en +original_t:cavity) (+language:en +original_t:including) (+language:en +original_t:teeth) (+language:en +original_t:gum) (+language:en +original_t:tissue) (+language:en +original_d:images) (+language:en +original_d:oral) (+language:en +original_d:cavity) (+language:en +original_d:including) (+language:en +original_d:teeth) (+language:en +original_d:gum) (+language:en +original_d:tissue)

Example 2 – Run: Stem-En-En, Query: 1
(+language:en +stem_t:imag) (+language:en +stem_t:oral) (+language:en +stem_t:caviti) (+language:en +stem_t:includ) (+language:en +stem_t:teeth) (+ language:en +stem_t:gum) (+language:en +stem_t:tissu) (+language:en +stem_d:imag) (+language:en +stem_d:oral) (+language:en +stem_d:caviti) (+language:en +stem_d:includ) (+language:en +stem_d:teeth) (+language:en +stem_d:gum) (+language:en +stem_d:tissu)
Example 3 – Run: Mids-En-En, Query: 1
(+language:en +mid_t:#imag) (+language:en +mid_t:#stom) (+language:en +mid_t:#excav) (+language:en +mid_t:#enfold) (+language:en +mid_t:#tusk) (+language:en +mid_t:#gum) (+language:en mid_t:#histio) (+language:en +mid_d:#imag) (+language:en +mid_d:#stom) (+language:en +mid_d:#excav) (+language:en +mid_d:#enfold) (+language:en +mid_d:#tusk) (+language:en +mid_d:#gum)(+language:en +mid_d:#histio)

[4] The Snowball Stemmer incorporates stop word lists containing 172 English, 232 German and 155 French entries.

Example 4 – Run: OrigMids-En-En, Query: 1
(+language:en +original_t:images) (+language:en +original_t:oral) (+language:en +original_t:cavity) (+language:en +original_t:including) (+language:en +original_t:teeth) (+language:en +original_t:gum) (+language:en +original_t:tissue) (+language:en +original_d:images) (+language:en +original_d:oral) (+language:en +original_d:cavity) (+language:en +original_d:including) (+language:en +original_d:teeth) (+language:en +original_d:gum) (+language:en +original_d:tissue) (+language:en +mid_t:#imag) (+language:en +mid_t:#stom) (+language:en +mid_t:#excav) (+language:en +mid_t:#enfold) (+language:en +mid_t:#tusk) (+language:en +mid_t:#gum) (+language:en mid_t:#histio) (+language:en +mid_d:#imag) (+language:en +mid_d:#stom) (+language:en +mid_d:#excav) (+language:en +mid_d:#enfold) (+language:en +mid_d:#tusk) (+language:en +mid_d:#gum)(+language:en +mid_d:#histio)

Example Box 1: Query syntax of the first query for Orig-En-En, Stem-En-En, Mids-En-En and Both-En-En. Instead of using nested query terms we prefer the more flexible but longer disjunctive normal form which consists of disjuncts, each of which is a conjunction of one or more query terms. Disjuncts (OR) do not have to be marked separately, conjunctions (AND) can be expressed by the + symbol.

3.4 Multilingual Scenario / Textual

In the multilingual scenario, only the morpho-semantic representation of the topics in English, German and French were used to search in both MID fields (*mid_t* and *mid_d*). No restrictions to the language flag of the document collections were made. The resulting runs were named *Mids-En-All*, *Mids-De-All* and *Mids-Fr-All*, respectively. As a baseline, the original queries in all languages were used to separately search in the language corresponding original document fields (i.e., English queries in English documents, German queries in German documents and French queries in French documents). The baseline is refered to as *Orig-All-All.* As a second baseline, we also consider the monolingual baseline *Orig-En-En* as a solid reference value, taking into account that the English subset represents a comprehensive part of the annotations (almost 93% of all annotations are available in English).

3.5 Multilingual Scenario / Mixed

In the mixed multilingual scenario we combined the retrieval results of the multilingual textual scenario with the GIFT visual retrieval results. In addition, a test run with the best textual run *Both-En-En* in combination with the GIFT results was carried out, as we expected this run to be the best of all runs. As the textual retrieval results were expected to perform better than the visual retrieval results, we ranked those documents at top which were found in both retrieval systems under the first 50 hits, followed by only textual retrieval results. Hits only found in the visual retrieval system were discarded. The runs are named *Orig-All-All-Comb*, *Both-En-En-Comb Mids-En-All-Comb*, *Mids-De-All-Comb* and *Mids-Fr-All-Comb.*

4 Results

The average precision values at all eleven standard recall points (0.0, 0.1, 0.2, ..., 1.0) are depicted in Figure 3 for the monolingual and in Figure 5 for the multilingual scenario.[5] We also depict the precision values between P5 and P100 in Figure 2 (monolingual) and Figure 4 (multilingual scenario). Interesting from a realistic retrieval perspective, at least to our view, is the average

[5] We here present results from a detailed re-evaluation after the results were officially published. They slightly diverge from the original results.

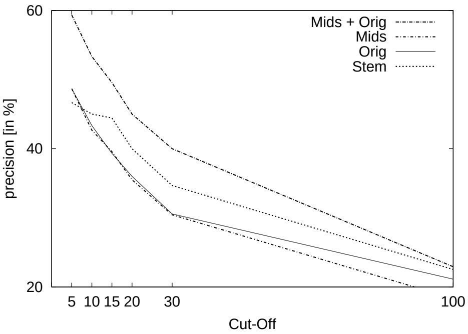

Fig. 2: Precision Monolingual

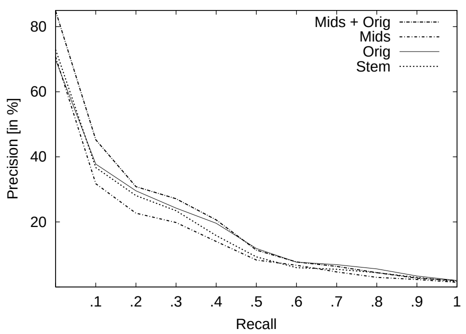

Fig. 3: IRCL Monolingual

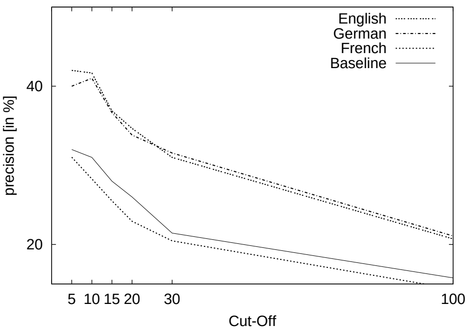

Fig. 4: Prec Multilingual / Textual

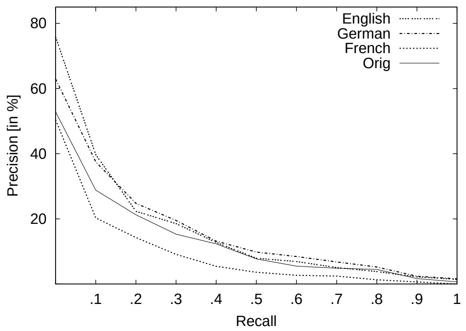

Fig. 5: IRCL Multilingual / Textual

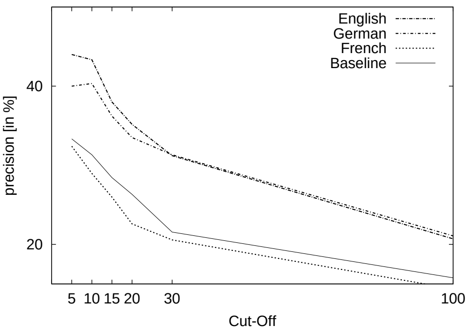

Fig. 6: Prec Multilingual / Mixed

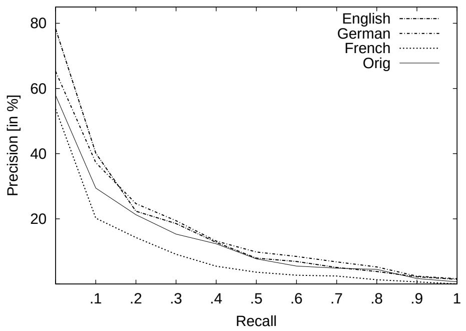

Fig. 7: IRCL Multilingual / Mixed

Fig. 8: Precision/Recall Graphs for the monolingual (first), multilingual/textual (second) and multilingual/mixed scenarios

	MonoText						
Scenario	Orig-En-En	Stem-En-En		Mids-En-En		Both-En-En	
map	0.1625	0.1482	91%	0.1297	80%	0.1792	110%
top2avg	0.3778	0.3669	97%	0.3175	84%	0.4525	120%
P5	0.4867	0.4667	96%	0.4867	100%	0.5933	122%
P20	0.3600	0.4000	111%	0.3550	99%	0.4500	125%

Table 3: Standard Precision/Recall Table for the Monolingual Scenario

	Multilingual / Textual								
Scenario	Orig-All-All	Orig-En-En		Mids-En-All		Mids-De-All		Mids-Fr-All	
map	0.1068	0.1625	152%	0.1366	128%	0.1439	135%	0.0734	69%
top2avg	0.2881	0.3778	131%	0.3970	138%	0.3751	130%	0.2028	70%
P5	0.3200	0.4867	152%	0.4200	131%	0.4000	125%	0.3103	97%
P20	0.2600	0.3600	138%	0.3467	133%	0.3383	130%	0.2293	88%

Table 4: Standard Precision/Recall Table for the Multilingual Scenario (Textual)

Scenario	Multilingual / Mixed								
	Orig-All-All	Both-En-En		Mids-En-All		Mids-De-All		Mids-Fr-All	
map	0.1079	0.1791	166%	0.1383	128%	0.1441	134%	0.0746	69%
top2avg	0.2942	0.4516	154%	0.4015	136%	0.3726	127%	0.2020	69%
P5	0.3333	0.6000	180%	0.4400	132%	0.4000	120%	0.3241	97%
P20	0.2633	0.4500	171%	0.3517	134%	0.3350	127%	0.2259	86%

Table 5: Standard Precision/Recall Table for the Multilingual Scenario (Mixed). *Comb*-suffixes of the scenario identifiers were stripped for lack of space.

gain on the top two recall points as well as the P5 and P20 values. Together with the overall *map* value, these figures can be found in Table 3 for the monolingual and in Table 4 for the multilingual scenario.

The first observation in the **monolingual scenario** was that the simple stemming approach *(Stem-En-En)* does not perform better than the baseline *Orig-En-En*, with values between 91% (*map*) and 111% (*P20*) of the baseline. Also, *Mids-En-En* performs not as good as the baseline *Orig-En-En* regarding *map* (0.13 vs.016) and top2avg (0.32 vs. 0.38) and achieves same figures as *Orig-En-En* for the P5 (0.49) and P20 (0.36) precision values.

The *Both-En-En* run outperforms all other runs in all relevant values. Regarding the P5 and P20 values of both *Both-En-En* and the baseline run *Orig-En-En* (P5: 0.60 vs. 0.48, P20: 0.45 vs. 0.36), *Both-En-En* exceeds the baseline by up to 25%. Top2avg (0.45 vs. 0.38) is 20% higher and the map-value (0.18 vs. 0.16) is still 10% increased. The P5 value of *Both-En-En* is second best of all (automatic/textual) runs at ImageCLEFmed 2006.

Regarding the **multilingual runs**, the first observation was that the multilingual baseline *Orig-All-All* was notedly lower than the monolingual baseline (the monolingual results of *Orig-En-En* are added in Table 4 for comparison). Obviously, due to the comprehensive part of English annotations (93% of the annotations are available in English) querying the English subset achieves better results than querying the German and French subset. Regarding the multilingual runs, we observe similar results for the German and the English test scenarios with P5 values between 0.40 (*Mids-De-All*) and 0.42 (*Mids-En-All*) and top2avg values between 0.38 (*Mids-De-All*) and 0.40 (*Mids-En-All*). Both runs exceeded the multilingual baseline regarding all relevant values by about 25% to 38%.

Considering the low number of German and French annotations compared to English, the *Mids-De-All* run can nearly be considered a fully *translingual* run (German queries on English documents). It is particular promising that this run performed as good as the English run (*Mids-En-All*) and the monolingual baseline (*Orig-En-En*) and clearly better than the multilingual baseline (*Orig-All-All*).

The French run *Mids-Fr-All* performed relatively poor compared to the others. This reflects the fact that we only just started to build the French Subword lexicon.

The **mixed multilingual runs** (Table 5) performed in average not better than the multilingual textual runs. Obviously, better merging algorithms to better exploit the synergies between the textual and the visual runs are needed and are due to future work.

While the P5, P20 and top2avg values of our best runs are quite high compared with other groups, the overall map value is mean. This is due to an overbalance of certain unspecific MIDs in our subword approach that cause a decrease of precision values beginning roughly at the cut-off point 100. However, as the P100-P1000 values are of only limited value for a user in real life scenario, these findings are not too worrying. Still the increase of the overall map value is an important goal of our future work.

5 Conclusion and Future Work

We introduced a subword-based approach for biomedical text retrieval that addresses many of the general and domain specific challenges in the current CLIR research. In our monolingual test runs, we showed that a combination of original and morpho-semantic normalized queries remarkably boosts precision up to 25% (P5, P20), compared to the baseline. The best of our runs was second best of all (automatic/textual) test runs regarding the P5 value. In the multilingual runs we achieve similar results for the English and German test runs. In all scenarios, the P5, P20 and top2avg values are distinctly higher than the multilingual baseline.

A detailed analysis of the test runs is now due in order to determine which part of our system contributed to which extend to the precision gain. Also, an increase of the overall *map*-value is aimed at, even though we consider P5 and P20 as the most important values in terms of user-friendliness. A medium-term goal is to show the usefulness of subwords in other (technical) domains.

References

[1] Martin Braschler and Bärbel Ripplinger. How effective is stemming and decompounding for german text retrieval? *Information Retrieval*, 7(3-4):291–316, 2004.

[2] Christiane Fellbaum, editor. WordNet*: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[3] Gregory Grefenstette and Pasi Tapanainen. What is a word, what is a sentence? problems of tokenization. In *Proceedings of the 3rd International Conference on Computational Lexicography*, pages 79–87, 1994.

[4] Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: an International Journal*, 41(3):523–547, May 2005.

[5] Kornél Markó, Stefan Schulz, and Udo Hahn. Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4):537–545, 2005.

[6] Kornél Markó, Stefan Schulz, and Udo Hahn. Unsupervised multilingual word sense disambiguation via an interlingua. In *AAAI 2005 – Proceedings of the 20th National Conference on Artificial Intelligence & IAAI'05 – Proceedings of the 17th Innovative Applications of Artificial Intelligence Conference*, pages 1075–1080. Pittsburgh, Pennsylvania, USA, July 9-13, 2004. Menlo Park, CA; Cambridge, MA: AAAI Press & MIT Press, 2005.

[7] Kornél Markó, Stefan Schulz, Alyona Medelyan, and Udo Hahn. Bootstrapping dictionaries for cross-language information retrieval. In *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 528–535. Salvador, Brazil, August 15-19, 2005. New York, NY: ACM, 2005.

[8] Ari Pirkola. Morphological typology of languages for IR. *Journal of Documentation*, 57(3):330–348, 2001.

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

[10] Gerald Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[11] Stefanie Tellex, Boris Katz, Jimmy J. Lin, Aaron Fernandes, and Gregory Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR 2003 – Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–47. Toronto, Canada, July 28 - August 1, 2003. ACM, 2003.