

Learning Indexing Patterns from One Language for the Benefit of Others

Udo Hahn¹ Kornél Markó^{1,2} Stefan Schulz²

¹Text Knowledge Engineering Lab, Freiburg University
Werthmannplatz 1, D-79085 Freiburg, Germany
<http://www.coling.uni-freiburg.de/>

²Department of Medical Informatics, Freiburg University Hospital
Stefan-Meier-Str. 26, D-79104 Freiburg, Germany

Abstract

Using language technology for text analysis and light-weight ontologies as a content-mediating level, we acquire indexing patterns from vast amounts of indexing data for English-language medical documents. This is achieved by statistically relating interlingual representations of these documents (based on text token bigrams) to their associated index terms. From these ‘English’ indexing patterns, we then induce the associated index terms for German and Portuguese documents when their interlingual representations match those of English documents. Thus, we learn from past English indexing experience and transfer it in an unsupervised way to non-English texts, without ever having seen concrete indexing data for languages other than English.

Introduction

Manual indexing or classification requires skilled human experts to perform a routine task, *viz.* to assign index terms or classification codes (usually, taken from a controlled vocabulary) to journal or newspaper articles, technical reports, etc. – once they have at least partially read and understood them. Limiting the choice of allowed descriptors to those conceptually organized in a thesaurus (the *Medical Subject Headings* (MESH 2001; Aronson *et al.* 2000) or the *Unified Medical Language System* (UMLS 2003; Aronson 2001), e.g., in the medical domain) creates additional benefits in that the description space is semantically structured in terms of more general/specific or related subject areas. Consequently, search capabilities become more powerful, e.g., through query expansion that incorporates synonyms, taxonomically more specific terms, etc. Large bibliographic services such as PUBMED (<http://www.ncbi.nlm.nih.gov/PubMed/>), which is maintained by the National Library of Medicine (NLM) and hosts MEDLINE, the largest bibliographic medical online database, still rely on the human indexers’ performance, as far as the content description of documents is concerned.

The manual assignment of index terms from a very large set of descriptors is not only a laborious and often tedious task, but it is also quite expensive. In the nineties, the NLM spent over two million dollars and employed 44 full-time equivalent indexers each year just for that task (Hersh *et*

al. 1994), facing yearly updates of less than 100,000 bibliographic units. In 2002, more than 350,000 new records were added to MEDLINE. Therefore, costs tend to increase at an even higher rate when indexing is performed by humans only.

MEDLINE includes English as well as non-English documents, though the indexing is only in English. After many years of hard work, more than 12 million bibliographic units have been indexed and classified using the *English* MESH as a controlled vocabulary. A few terminology mappers exist from English to some non-English languages, but their coverage is far from complete (compared with the English MESH). Because the role of the physicians’ native languages is much more dominant than in other scientific disciplines, the focus on English as the medical content description language creates a serious bottleneck for tentative users of PUBMED in non-English-speaking countries.

In order to reuse this bulk of scholarly work for languages other than English, we started a project which benefits from that experience in the following way: Assuming that the English indexing of medical documents is a highly valued asset, we determine lexical patterns in the abstracts and relate them to their associated index terms. This mapping is affected by language technology tools, which incorporate light-weight ontologies for interlingual content representation. Once we (can) map lexical items from an arbitrary non-English language (e.g., German or Portuguese) to their English lexical correlates, we may then reuse the English indexing patterns for the non-English languages, once a mediating interlingua is available. Thus, we learn from past English indexing experience and transfer it in an unsupervised way to non-English texts, without ever having seen concrete indexing data for languages other than English.

Document Analysis

We take a lexicalized, bag-of-words-style approach to document analysis. Our main focus is on morphological analysis, because morphological processes alter the appearance of words but leave their core meaning, by and large, intact. Such morphological variants can generally be described as concatenations of basic sublexical forms (e.g., stems) with additional substrings (affixes). We distinguish three kinds of morphological processes, *viz.* inflection (e.g., adding the

plural “*es*” in “*leuk⊕o⊕cyt⊕es*”,¹ derivation (e.g., attaching the derivation suffix “*ic*” in “*leuk⊕o⊕cyt⊕ic*”), and composition (e.g., in “*leuk⊕em⊕ia*”). Morphological analysis is concerned with the reverse processing of morphological variants, i.e., deflection (or lemmatization), dederivation and decomposition. The goal is to map all occurring morphological variants to some canonical base form(s) — e.g., ‘*leuk*’ or ‘*cyt*’ in the examples from above.

Medical terminology, on which we focus here, is further characterized by a typical mix of Latin and Greek roots and the corresponding host language, e.g., as evidenced by “neuroencephalomyelopathy”, “glucocorticoid”, “pseudohypoparathyroidism”. This may also surface in some languages, e.g., German, as mild orthographic variations such as with *Karzinom* and the equivalent *Carcinom*. Obviously, dealing with such phenomena is crucial for medical information retrieval, in particular, for a system that maps free text to a controlled indexing vocabulary such as MESH.

Interlingual Morpho-Semantic Normalization

In order to deal with such morphological problems in an IR scenario, we developed a document pre-processing engine which takes plain English, German and Portuguese texts as input and transforms them in three steps, viz. orthographic, morphological and semantic normalization (cf. Figure 1).

During orthographic normalization all capitalized characters from input documents are reduced to lower-case characters. Additionally, language-specific character substitutions are made (e.g., for German ‘*ß*’ → ‘*ss*’, ‘*ä*’ → ‘*ae*’, ‘*ö*’ → ‘*oe*’, ‘*ü*’ → ‘*ue*’ and for Portuguese ‘*ç*’ → ‘*c*’, ‘*ú*’ → ‘*u*’, ‘*õ*’ → ‘*o*’). This step eases the matching of (parts of) text tokens and entries in the lexicons.

For each source language, we provide a *subword lexicon* as background knowledge for morphological normalization. Each one contains morpheme-like entries (22,000 subwords for German and English, 14,000 for Portuguese). Based on such a lexicon, the system segments the orthographically normalized input stream into a sequence of semantically plausible subwords. The results are then checked for morphological plausibility using a finite-state automaton in order to reject invalid segmentations (e.g., ones 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.

For semantic normalization, we extended the subword lexicon by a *subword thesaurus*, in which intra- and interlingual semantic equivalence classes are defined, e.g., the one composed of English “*kidney*”, German “*niere*”, and Latin “*nephr*”. Each semantically relevant sublexical unit produced by the morphological segmentation is replaced by its corresponding *morpho-semantic class identifier* (MID), which represents all subwords assigned to one particular class. The result is a morpho-semantically normalized document in a language-independent, interlingual representation.

¹‘ \oplus ’ denotes the string concatenation operator.

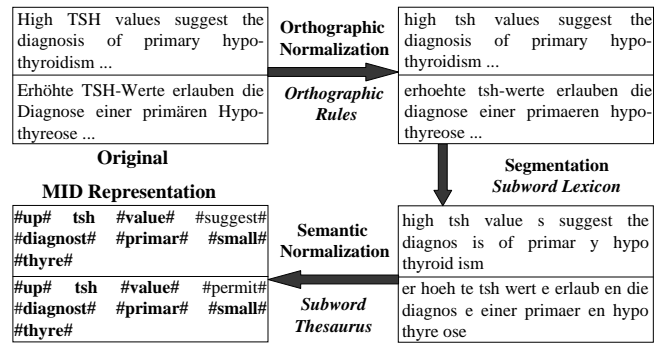


Figure 1: Morpho-Semantic Normalization Pipeline

In Figure 1, bold-faced MIDs co-occur in both document fragments. A comparison of the original natural language documents at the beginning of the pipeline and their interlingual representation at the very end already reveals the degree of (hidden) similarity uncovered by the overlapping MIDs.

During the construction of the subword thesaurus, we faced many of the notorious problems one encounters in the field of cross-language information retrieval, viz. ambiguity when translating terms, intralingual synonymy and homonymy. In our approach, multiple entries connected to a single MID express intralingual synonymy and interlingual translation. Ambiguous terms are linked to their respective equivalence classes by a special thesaurus link, producing a sequence of their related MIDs.

In practice, after three person years of construction on the medical subword lexicons and the thesaurus, their growth has almost reached the end at the level of 58,000 subwords connected to 28,500 equivalence classes, in total.

Learning Indexing Patterns

In the following section, we describe a statistical, a heuristic and a hybrid approach to automatically identifying English MESH entries as document descriptors for English, as well as German and Portuguese documents, given sets of *a priori* assigned index terms to English documents. In the MESH (2001), descriptors are organized in a conceptual hierarchy. In its 2002 version (which we use), over 20,500 so-called main headings occur as potential descriptors.²

The particularity of our approach of assigning descriptors to documents is based on an interlingua representation for the documents as well as the indexing vocabulary. Using this representation (generated by the normalization procedures described the previous section) is advantageous in that we are able to train our indexing system on texts written in one language (English) and test it on documents of another, up until now unseen, languages. This approach allows the processing of documents in any language covered by the lexicons.

We start with a sample of 35,000 medical abstracts ($word_1 \dots word_m$ in Figure 2, step A) taken from the set of approximately 350,000 new entries stored in MEDLINE in

²Publication Types and special descriptors such as Subheadings, Chemical Supplementary Terms, Age Groups, Check Tags and Geographics are not considered in our experiments.

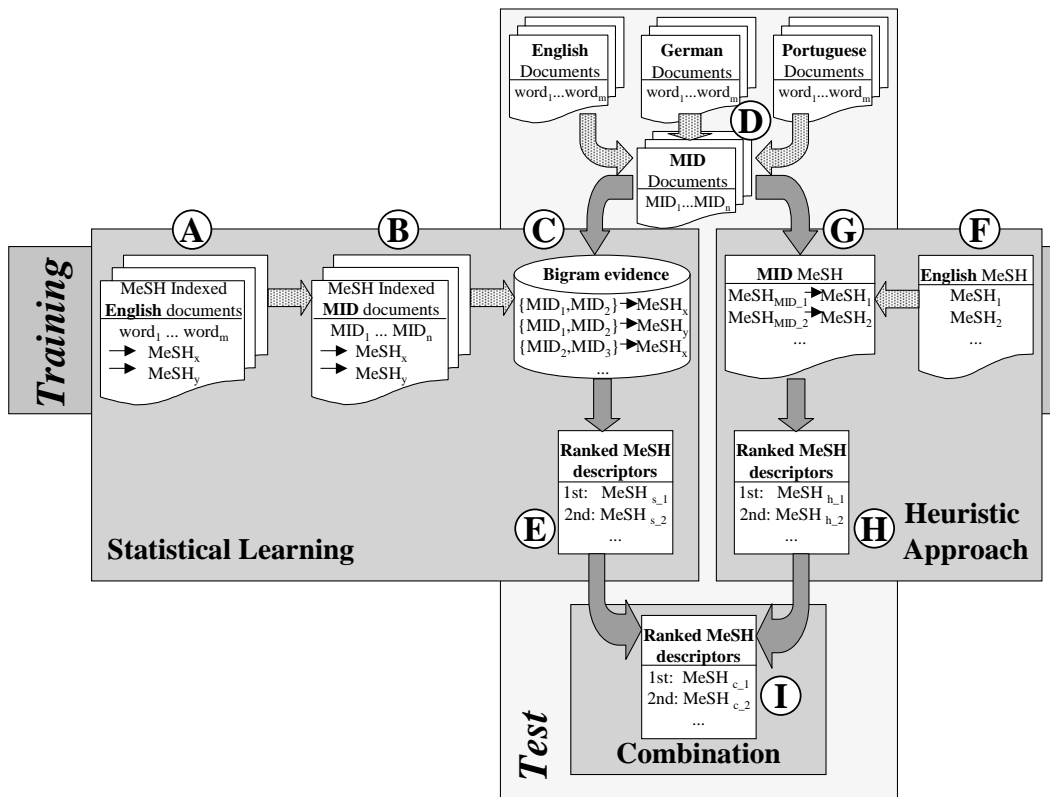


Figure 2: Architecture of the Combined Indexing System

$$w(MeSH_i | MID_1, \dots, MID_n) = \log \prod_{j=1}^{n-1} \begin{cases} \frac{P(MID_j, MID_{j+1} | MeSH_i)}{P(MID_j, MID_{j+1})} & , \text{ if defined} \\ 1 & , \text{ otherwise} \end{cases}$$

Figure 3: Estimating the Conditional Weighting Value of a MESH Descriptor Given n Morpho-Semantic Class Identifiers (MIDs)

2002, to which English MESH main headings have already been assigned manually ($MeSH_x$ and $MeSH_y$ in Figure 2, step A). After morphological normalization, each natural language document is represented by a sequence of morpho-semantic class identifiers ($MID_1 \dots MID_n$ in Figure 2, step B), while still preserving the original MESH index terms assigned to it. Based on that representation we employ a Bayesian approach which ignores the *a priori* probabilities of the descriptors. Thus, statistical evidence for class identifier (MID) bigrams is basically computed by counting their frequency of co-occurrence in the training corpus with individual (manually supplied) MESH entries (Figure 2, step C). That is how indexing patterns are learned.

In the test phase, we aim at identifying MESH terms as valid descriptors for unseen normalized documents (cf. Figure 2, step D), and we rank these terms by their weighting values (w as defined in Figure 3). Given a document which contains n class identifiers (MIDs), the conditional weighting value for $MeSH_i$, a particular MESH main heading, is computed with the product of the computed conditional probabilities P of the MID bigrams in the text that co-occur

with the descriptor $MeSH_i$ in the training set, divided by the *a priori* probability of the corresponding text bigrams in the training collection, if both probabilities can be observed. The denominator of the weighting function takes into account the fact that infrequent terms have a greater explanatory power for a given entity when faced with large quantities of data, therefore increasing the weighting value for that entity. If no bigram that is currently being processed appears in the training data, or if it is not associated with the current descriptor $MeSH_i$, it remains neutral (multiplication by 1).

We treat MID bigrams in an unordered way. They are defined as a set of MIDs that co-occur within a document window of two text tokens, regardless of the original sequence of words that license the set of MIDs. This is due to the fact that in German and English, the MID order changes when genitives or prepositions come into play, as with "femoral neck fracture" vs. "fractured neck of femur" corresponding to ["#femur#", "#nuchal#", "#fractur#"] vs. ["#fractur#", "#nuchal#", "#femur#"]. Finally, all identified MESH descriptors, $MeSH_{s,1}$, $MeSH_{s,2} \dots$, are ranked according to their weighting value (Figure 2, step E).

MESHing Document Representations

The indexing patterns identified with the methodology from the previous section are the backbone for the retrieval performance of our system. In the course of retrieval experiments, however, we also found that some heuristic add-ons were helpful to further increase the system's performance. In this section, we describe some of these heuristics and a hybrid approach, which combines the learned indexing patterns with these heuristics to automatically identify MESH entries as document descriptors.

In the heuristic approach, we just rely on the MESH Thesaurus and a collection of (non-indexed) documents. Based on suitable criteria, a fully automatic MESH indexing of the documents is computed. Unlike the learning method, no prior indexing of documents is necessary. In the training phase, all English MESH main headings, $MeSH_1$, $MeSH_2$, etc., (cf. Figure 2, step F) undergo the morpho-semantic normalization procedure. Thus, all words from the main headings which are covered by the English sub-word lexicon are substituted by their corresponding unique MIDs resulting in morpho-semantically normalized representations, $MeSH_{MID,1}$, $MeSH_{MID,2}$, etc., which are linked to the original MESH descriptors (Figure 2, step G).

In the test phase, English, German and Portuguese documents, defined by a sequence $word_1 \dots word_m$, are processed by the morphological engine which generates for each document an identifier sequence, $MID_1 \dots MID_n$, at the interlingua level (Figure 2, step D). Afterwards, heuristic rules (some of them already proposed by NLM's indexing initiative (Aronson *et al.* 2000)) are applied to the normalized test documents. In essence, this means that each MESH descriptor, whose normalized representation contains at least one of the MIDs in the document, is retrieved. Next, each normalized MESH descriptor is assessed against the normalized text by computing diverse factors. We confine ourselves to the most important metrics:

- **Longest Match Factor:** On the level of MIDs, individual MESH descriptors, which appear as single entries, can also appear as part of other MESH entries. For example, the German technical term "*Bauchschmerzen*" ("*abdominal pain*") that appears in a text and is normalized to the MIDs "*#abdom#*" and "*#pain#*" is, amongst others, associated with the MESH entries "Abdominal Pain" (*[#abdom#*, "*#pain#*"]), "Abdomen" (*[#abdom#*) and "Pain" (*[#pain#*]). If two or more normalized MESH descriptors can be merged into one longer MESH descriptor, the latter is preferred over its constituents.
- **Phrase Factor:** The number of different MIDs in a phrase that match the MIDs in a normalized descriptor is called *MID number*. In addition, we consider the phrase interval of a normalized descriptor as the span between the first and the last MID associated with this descriptor in a phrase. Then, the *phrase factor* is defined as the ratio of MID number and phrase interval. The Portuguese phrase "*o fígado do paciente foi transplantado*" ("*the patient's liver was transplanted*"), e.g., will be transformed into [*#hepat#*, "*#patient#*", "*#transplant#*"]. Given the normalized descriptor for "liver transplantation" ("*#hepat#*",

#transplant#"), the corresponding MID number is 2, and the phrase interval amounts to 3. Therefore, the phrase factor equals $2/3$.

- **Entry Factor:** The *entry factor* is the MID number divided by the number of MIDs of the associated descriptor. For example, the German noun phrase "*noduläre Hyperplasie*" ("*nodular hyperplasia*") is normalized to [*#nodul#*, "*#above#*", "*#plast#*"] and the MESH descriptor "Focal Nodular Hyperplasia" to [*#focal#*, "*#nodul#*", "*#above#*", "*#plast#*"]. The corresponding entry factor is $3/4$.
- **Title Factor:** A descriptor found in the title will be ranked higher than others.

Finally, all possible descriptors are ordered according to a weighted average of the above (and other) metrics ($MeSH_{h,1}$, $MeSH_{h,2} \dots$ in Figure 2, step H).

We also pool the results of the statistical learning of indexing patterns and the heuristic add-ons in order to find out whether a combined effort performs better than any of the two in isolation. Thereupon, we merged both approaches in the following way. Firstly, all descriptors that are ranked in the top thirty by *both* of the methods are put at the top of the result list ($MeSH_{c,1}$, $MeSH_{c,2} \dots$ in Figure 2, step I). After the first k positions ($30 \geq k$) have been filled that way, the remaining positions are incrementally generated by the following rule: The two entries at the top of the output of the statistical approach are alternately incorporated into the final result, followed by one entry of the heuristic approach, until both lists (maximum length: 100 terms) are exhausted. Previous experiments have shown that this empirically motivated procedure leads to much more favorable results than a formal one, e.g., by multiplying the weights from the different weighting functions.

Evaluation Scenario

For our evaluation, we selected abstracts from English, German and Portuguese medical journals. English and Portuguese texts were taken from PUBMED, the online interface of the MEDLINE database. German articles came from SpringerLink (<http://link.springer-ny.com/>), an online library for medicine which contains dozens of medical journals. We chose those journals which are linked to MEDLINE. Manually assigned MESH identifiers were also extracted from PUBMED.

We then randomly assembled text collections for the training phase (35,000 English abstracts for the statistical learning of indexing patterns from the English corpus) and the test phase (4,000 abstracts from the English/German corpus and due to limited availability, only 800 abstracts from the Portuguese corpus).

The data acquired during the training phase were then used for the indexing of English, German and Portuguese documents. The abstracts were automatically processed according to the methods described in the previous sections and the indexing results were evaluated against the manually supplied MESH main headings. This data serves as the *de facto* gold standard for our experiments (similar to the study

of the indexing initiative of the NLM (Aronson *et al.* 1999). Unfortunately, human indexings in MEDLINE are not really consistent. Funk & Reid (1983) measured 48.2% interrater agreement with regard to manually assigned MESH main-heading for English abstracts (only 41.5% for German abstracts). Obviously, such inconsistencies in the test collection will also affect the validity of our evaluation results when we take this data as the gold standard.

Evaluation Results

Table 1 depicts the precision and recall values for the chosen test scenarios. For each of the three languages we considered the top 5, 10 and 40 ranked descriptors. These cut-off layers were previously introduced by Aronson *et al.* (1999) and we adopt them in order to more accurately compare the results.

Cut Off	Method	English		German		Portuguese	
		Prec	Rec	Prec	Rec	Prec	Rec
5	Heuristic	32.7	18.5	25.9	16.9	21.6	17.5
	Statistical	34.3	18.6	28.2	18.1	18.7	14.1
	Combined	41.7	23.2	33.2	21.5	25.7	20.0
10	Heuristic	23.0	25.2	17.7	22.8	14.3	22.3
	Statistical	24.9	26.3	19.9	25.0	13.1	19.2
	Combined	30.2	32.8	23.3	29.7	17.6	26.8
40	Heuristic	8.6	36.4	6.5	32.9	5.6	33.9
	Statistical	10.4	42.5	8.4	41.7	5.8	33.9
	Combined	12.8	52.9	9.6	47.7	7.1	42.0

Table 1: Precision/Recall (Prec/Rec) Table for Different Languages and Indexing Methods at Different Cut-off Points

When focusing on the combined approach and considering the top five descriptors, our procedure already retrieves 23% of all relevant MESH terms for English (22% for German, 20% for Portuguese) at a precision rate of 42% (33% and 26%, respectively). Looking at the top 40 of the system-generated descriptors, precision for English drops to 13% (10% for German, 7% for Portuguese), whilst recall increases to 53% (48% and 42%, respectively).

Figure 4 summarizes the resulting precision/recall values for the different languages for the top 5, 10, 20, etc. up to the top 100 proposed descriptors using the combined indexing procedure. The crossing of the curves indicate that, on the average, ten MESH descriptors were manually assigned to the abstracts of the test collection. With recall values ranging between 27% and 33% for each of the languages for the top ten assigned descriptors, results seem to be not so shiny. However, when we compare these values to the average agreement of *human* indexers (for English 4.82 MESH descriptors out of ten, for German only 4.15 (Funk & Reid 1983)) our system derives 1.54 less descriptors on the average for English (68.1%) and only 1.18 less for German (71.6%) in a *fully automatic* indexing environment.

Considering the human indexing agreement for English as baseline for all of the three languages we achieve 68% for English, 62% for German and 56% for Portuguese. Keeping in mind that the processing of German and Portuguese

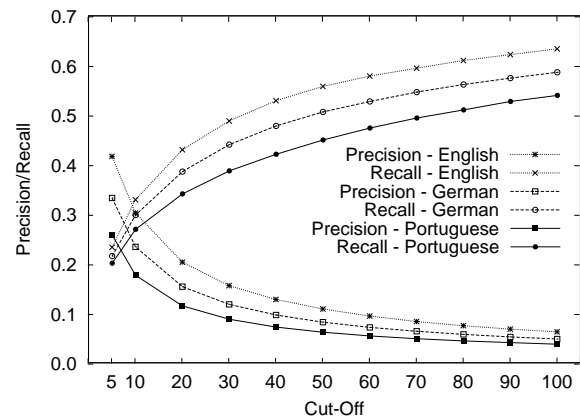


Figure 4: Precision/Recall Graphs for the English, German and Portuguese Indexing Task at Different Cut-off Points (Combined Indexing Method)

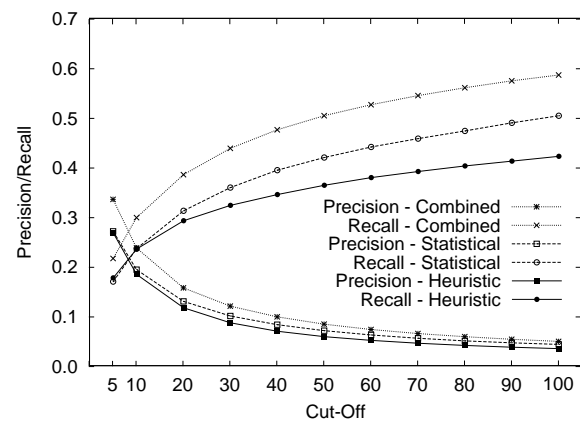


Figure 5: Precision/Recall Graphs for the Different Indexing Procedures at Different Cut-off points Averaged Over Three Languages (English, German, and Portuguese)

abstracts was not part of the training phase, these results are still promising.

We also examined the different contributions of the two basic approaches we pursue. With the exception of Portuguese, in isolation, the statistical learning approach always significantly outperforms the heuristic one, for all languages at all cut-off points with respect to recall and precision. The learned indexing patterns are, therefore, the driving force for the performance of the system. Pooling both approaches, however, yields additional, mostly substantial benefits.

To summarize, Figure 5 shows the resulting precision/recall value pairs for the different indexing methods for the top 5, 10, 20, etc. up to the top 100 proposed descriptors averaged over the three languages we dealt with.

Related Work

The (monolingual) MESH mapping methods proposed by NLM's indexing initiative (Aronson *et al.* 2000; 1999) reach

29% recall (top 5) and 61% recall (top 40)³ on a small test corpus comprised of 200 MEDLINE abstracts (our data is 23% and 53% recall, respectively). In previous studies, when we focused on a smaller subset of MEDLINE documents covering clinical disciplines only, we observed similar results. The loss of performance in this study indicates that the lexicon has to be further adapted to non-clinical topics, e.g., public health, since we now take a much more comprehensive sample of MEDLINE into account.

In Eichmann et al.'s system (Eichmann, Ruiz, & Srinivasan 1998) and in the SAPHIRE International server (Hersh & Donohoe 1998) for cross-language IR, multilingual user queries are transformed into entries of the UMLS Metathesaurus in order to retrieve suitable documents (cf. also Aronson (2001)). These mapping approaches have much in common with our heuristic indexing procedure which is, however, clearly outperformed by the statistical learning method.

Sebastiani (2002) points out that any content-based indexing method that incorporates some machine learning algorithm (e.g., probabilistic or decision tree classifiers) does better than methods without any learning device. Various experiments carried out on the REUTERS-21578 corpus, the most widely used benchmark collection in automated indexing, showed that the combination of different indexing methods seems to generally perform the best. These considerations are also backed up by our results.

Conclusions

We have investigated the use of large quantities of indexing knowledge as available for English medical documents and its transfer to non-English documents for which no such data is available. Our approach is based upon the supply of manually created subword lexicons and a conceptual interlingua linked to these language-specific lexicons. After automatic lexical analysis, indexing patterns are automatically learned from interlingual text representations. The subsequent assignment of English document descriptors to English, German and Portuguese texts is also achieved without any manual intervention. Based on evidence from our experiments, we may cautiously conclude that our approach is a reasonable way to go (still keeping in mind the fragile gold standard on which our results rest).

The methodology we use for learning indexing patterns is a standard Bayesian approach. The methodological news we have to offer is two-fold. First of all, the combined effort of statistical and heuristic methods beats a purely statistical approach. Secondly, pattern learning is not only based on the literal word (n-gram) level, but incorporates the level of light-weight ontologies, as well. The latter provides us with a conceptual interlingua which is crucial for the language-mapping we focus on.

Concerning the generality of our approach, medical lexicons for any natural language are certainly much more constrained in terms of lexico-semantic ambiguities than the general language lexicon (as witnessed by the ambiguity rates, e.g., in the WORDNET lexical database (Fellbaum

³Including Check Tags and Age Groups, which are easier to identify.

1998)). At the same time, the need for developing a home-grown medical lexicon becomes apparent after an inspection of WORDNET. Although some medical terminology can be found there, it is by no means adequately covered. While the development of the newly developed subword thesaurus has almost been completed for English and German, the Portuguese version still needs further curation.

Acknowledgments

This work was funded by *Deutsche Forschungsgemeinschaft*, grant Klar 640/5-1. We thank our Brazilian colleagues, Percy Nohama & Roosewelt Leite de Andrade, for the acquisition and maintenance of the Portuguese lexicon.

References

- Aronson, A. R.; Bodenreider, O.; Chang, H. F.; Humphrey, S. M.; Mork, J. G.; Nelson, S. J.; Rindflesch, T. C.; and Wilbur, W. 1999. The indexing initiative. In *A Report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications*.
- Aronson, A. R.; Bodenreider, O.; Chang, H. F.; Humphrey, S. M.; Mork, J. G.; Nelson, S. J.; Rindflesch, T. C.; and Wilbur, W. J. 2000. The NLM indexing initiative. In *AMIA 2000 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, 17–21. Los Angeles, CA, November 4-8, 2000. Hanley & Belfus.
- Aronson, A. R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The METAMAP program. In *AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, 17–21. Washington, D.C., November 3-7, 2001. Hanley & Belfus.
- Eichmann, D.; Ruiz, M. E.; and Srinivasan, P. 1998. Cross-language information retrieval with the UMLS metathesaurus. In *SIGIR'98 – Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 72–80. Melbourne, Australia, August 24-28, 1998. ACM.
- Fellbaum, C., ed. 1998. *WORDNET: An Electronic Lexical Database*. MIT Press.
- Funk, M. E., and Reid, C. A. 1983. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association* 71(2):176–183.
- Hersh, W. R., and Donohoe, L. C. 1998. SAPHIRE International: A tool for cross-language information retrieval. In *AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium*, 673–677. Orlando, FL, November 7-11, 1998. Hanley & Belfus.
- Hersh, W. R.; Hickman, D. H.; Haynes, B.; and McKibbin, K. A. 1994. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association* 1(1):51–60.
- MESH. 2001. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.
- UMLS. 2003. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.
- Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47.