Medical Informatics Europe 2002
G Surján et al. (eds)
IOS Press 2002

Text Retrieval Based on Medical Subwords

Martin HONECK [a], Udo HAHN [b], Rüdiger KLAR [a], Stefan SCHULZ [a]

[a] *Freiburg University Hospital, Department of Medical Informatics*
[b] *Freiburg University Hospital, Text Knowledge Engineering Lab*

Abstract. In biomedical documents, there is ample evidence for complex morphological structures in specialized terms. While inflection is relatively easy to deal with, productive morphological processes such as derivation and single-word composition constitute a major challenge. Considering the problem from an information retrieval perspective, we split morphologically complex words into biomedically significant, morpheme-like subwords and match subwords the query terms and document terms are composed of. This way, morphologically motivated word form alterations can be eliminated from the retrieval procedure. Based on a series of retrieval experiments, we have gathered evidence that subword-based indexing and retrieval – for the German biomedical sublanguage, at least – outperforms conventional string matching approaches.

1. Introduction

Both medical professionals and the general information-seeking public need easy-to-use query interfaces in order to retrieve health-related contents. Simple string matching procedures in information retrieval (IR) systems usually fail to account for morphological variants of a search term (e.g, [*abdomen, abdominal*]; [*lung, lungs*]; [*foot, feet*]) so that their recall performance decreases [1, 3, 4]. The efforts required for extracting word stems from inflected, derived or composed words vary between languages and domains. Whereas the English language is known for the limited number of inflection patterns, others, such as German, are much more diverse. Therefore, English general-purpose stemming algorithms [5, 9] already operational in many IR applications have no counterparts in these morphologically more diverse languages. When derivation (e.g., *hepat⊕ic*) and composition (e.g., *hepat⊕o⊕cellul⊕ar*) phenomena have to be considered, too, even for the English language only restricted, domain-specific algorithms exist up to now [8, 7].

Not only is the German language known for excessive single-word nominal compounding, but also its medical sublanguage, in particular, characterized by a mix of Latin and Greek roots. Besides fairly standardized noun compounds, which already form a common part of biomedical terminologies, a myriad of *ad hoc* compounds (and derivational forms) are formed on the fly. These cannot be fully anticipated when a retrieval query is formulated. For this reason, the enumeration of morphological variants in a semi-automatically generated lexicon, such as proposed for French [14], turns out to be infeasible for the German language. Our approach to deal with this challenge is based on the segmentation of complex biomedical terms into *subwords*. We have experimental evidence that subwords significantly improve the performance of text retrieval in German-language biomedical document collections.

2. Subword Segmentation

From a linguistic perspective, *morphemes* are defined as the smallest content-bearing (*stems*) or grammatically relevant (*affixes*) units. From an IR perspective, however, our notion of *subwords* features more coarse-grained morphological units in order to align with semantic equivalents (e.g. [*diaphys* (rather than *dia⊕phys*); *shaft*], or even [*ascorb; vitamin C*]). In this approach, both query and document terms are split into medically non-decomposable subword units using two components, viz. a *subword thesaurus* and a *morphological analyzer* (for details, cf. [12, 13]).

Subword thesaurus. The German-language part[1] underlying this study is currently composed of 5,275 entries. Those entries considered semantically equivalent are grouped using a common identifier (equivalence class identifier, called *ECI*). The ECI allows semantic matching, especially between foreign-language translates and source language terms (e.g., [*kidney, nephr*]) and morphological variants (so-called allomorphs, e.g., [*hepar, hepat*] or [*foot, feet*]).

The **morphological analyzer** implements a simple morphological word model based on a regular language, and uses heuristic rules in order to select the most plausible reading among alternative segmentations. Prior to segmentation, a language-specific *orthographic* normalization step is performed[2]. As a result of segmentation, two output formats can be generated:

(i) *morphological normalization*: The morphologically segmented input text.

(ii) *morphosemantic normalization*: A "pseudo-text" in which each meaningful subword is substituted by its *ECI* code.

Both output formats are compatible with standard full-text indexing and retrieval systems.

3. Retrieval Environment

In a preliminary evaluation study [2] we used an off-the-shelf text retrieval environment, the *AltaVista*™ search engine, in order to demonstrate how a publicly available tool fits our indexing approach. Such a "black box" system, however, is quite problematic in an experimental test setting[3]. We, therefore, implemented our own search engine using the Python[4] script language. It crawls text/HTML files, produces an inverted file index, and assigns weights to terms and documents based on term frequency and inverse document frequency. Query processing relies on Salton's vector space model [11] using the cosine measure for determining the similarity between a query and a document. All terms (except stop words) from the document collection are organized in an inverted term index accessible for retrieval. The search engine then produces a ranked output of documents. Proximity between search terms is used as an additional ranking criterion.

---

[1] The subword thesaurus currently under construction includes English as well as Portuguese.

[2] Mapping German umlauts *ä*, *ö*, and *ü* to *ae*, *oe*, and *ue*, respectively; *ca* to *ka* and other character conversions.

[3] Especially the adjacency (proximity) criterion deserves attention, particularly in those cases when subwords are extracted from the same text token. Otherwise, no distinction could be made between a document containing *appendectomy* and *thyroiditis* and one containing *appendicitis* and *thyroidectomy*.

[4] www.python.org

4. Experiments

We tried to assess whether morphological segmentation had a positive impact on text retrieval using a standard evaluation approach developed in the IR community [10]. As a *document collection* for our experiments we chose the CD-ROM edition of a German language handbook of clinical medicine [6][5]. Two *user query collections* were acquired:

Expert queries: 63 medical students were presented a random selection of multiple choice

(MC) questions covering clinical medicine. Then we asked them to formulate free-form natural language queries intended to help find the correct answer. So we ended up with 630 queries, from which 25 ones were randomly chosen for our experiments.

Laymen queries: The operators of the German-language medical search engine *Dr. Antonius*[6] provided us with a set of 38.600 logged queries. A random sample (n=400) was classified by a medical expert whether containing medical sublanguage or layman expressions. From 125 queries without medical sublanguage (layman expressions) 27 ones were randomly chosen for our study.

Table 1: Evaluation Results – Precision/Recall Table
Bold face indicates statistically significant differences between (*plain* vs. *segm*, *plain* vs. *norm*)

	Precision (%)								
Recall	*plain*	*segm*	*norm*	*plain*	*segm*	*norm*	*plain*	*segm*	*norm*
(%)	expert queries n=25			layman queries n=27			all queries n=52		
0	60.8	67.3	64.7	59.1	**80.3**	**81.0**	60.0	**74.0**	**73.2**
10	59.8	60.3	60.3	52.2	64.0	61.6	55.8	62.3	61.0
20	48.6	50.8	50.3	36.2	**53.6**	**52.9**	42.1	52.3	51.7
30	37.3	46.5	45.7	31.9	**45.1**	**44.5**	34.5	**45.8**	**45.1**
40	29.0	37.3	32.0	31.4	**41.0**	**40.7**	30.3	39.2	36.5
50	26.5	34.2	28.3	30.7	36.8	36.8	28.7	**35.6**	32.7
60	20.1	24.7	20.3	29.6	34.4	35.3	25.0	29.7	28.1
70	11.1	19.9	15.7	25.8	28.5	29.2	18.7	24.4	22.7
80	9.1	14.2	10.3	18.5	24.6	25.3	14.0	**19.6**	18.1
90	4.7	**9.2**	8.3	14.8	19.7	**20.5**	9.9	**14.7**	**14.6**
100	4.4	**8.3**	7.6	3.7	**11.5**	**12.7**	4.0	**10.0**	**10.2**
11pt avrg.	24.1	33.9	31.2	29.5	40.0	40.0	26.9	37.0	35.8

The relevance judgments were done by three domain experts, identifying relevant documents in the whole test collection for each of the queries in either set. This very time-consuming task explains the low number of queries. We conducted the following experiments, using an experimental version of the subword thesaurus:

- **Test 1: Token Search** ("*plain*"). Only orthographic normalization (cf. section 2) precedes indexing and the submission of the query for retrieval. The search was run on the index covering the entire document collection (182,306 index terms). This scenario serves as the baseline for determining the benefits of our approach.
- **Test 2: Morphological Segmentation** ("*segm*"). After orthographic normalization (cf. section 2) document and query words were split into subwords. This resulted in a decrease of the size of the index, with 39,315 index terms remaining.

[5] It contains 5,517 articles (about 2.4 million text tokens) on a broad range of clinical knowledge using biomedical terminology.
[6] http://www.dr-antonius.de/

- **Test 3: Morphological Segmentation and Synonym Expansion.** (*"norm"*). In addition to Test 2, all subwords in queries and documents were substituted by the related synonym class identifier thus enabling synonym matching (cf. section 2).

The assessment of the experimental results is based on the aggregation of 25 expert queries respectively 27 layman queries. We calculated the average interpolated precision values at fixed recall levels (we chose a continuous increment of 10%) based on the consideration of the top 200 documents retrieved by the search engine, and as a global measure, the average precision of all eleven recall points (11pt average).

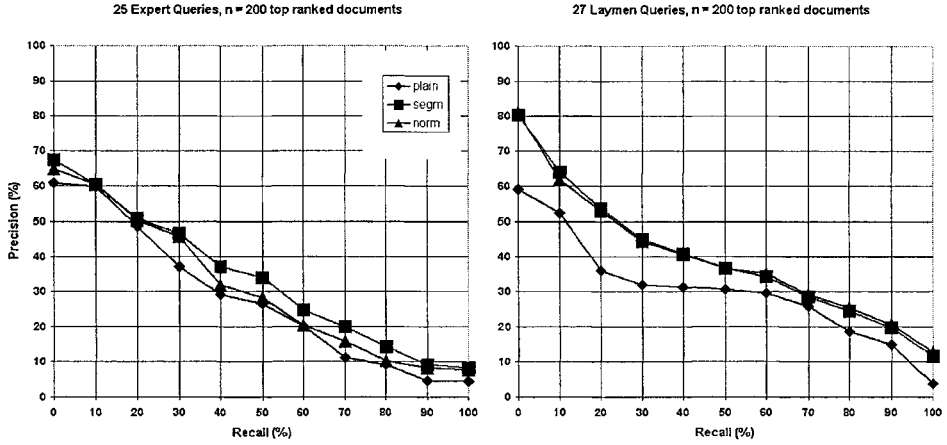

Figure 1: Precision / Recall Graph for Layman and Expert Queries

The corresponding precision / recall values for all test scenarios are summarized in Tab. 1, the results for the layman and expert scenarios additionally visualized in Figure 1. Using the two-tailed Wilcoxon test we identified a significant ($\alpha < 5\%$) difference between *plain* and *segm* at the recall levels 0%, 30%, 50%, 80%, 90% and 100%, between *plain* and *norm* at 0%, 30%, 90% and 100%.

In addition we found that

- The comparative advantage of the segmentation is mostly due to the laymen queries.
- Eight out of 27 layman queries (29.6 %) as well as 2 of 25 expert queries (8%) yielded no results using the *plain* method. However in all of these cases, *segm* and *norm* returned relevant results.
- The *AltaVista™* search engine produces higher recall values (as 11pt average) than our experimental crawler [*in brackets*] (*plain*: 28.5 [*26.9*], *segm*: 44.4 [*36.5*], *norm*: 40.0 [*30.7*], only measured for expert queries in a preliminary study [2])

Generalizing the interpretation of our data in the light of these findings, we recognize a mild increase of retrieval performance when query and text tokens are segmented according to the principles of the subword model. The resolution of synonyms, to our surprise, did not increase the performance. We ascribe this to the fact that the terminology used in the queries was nearly identical to the one occurring in the documents of the test collection, and the noise created by false matches was higher. As a direct consequence for our work we will check the synonym classes of the thesaurus as a source of reasons for possible over-generalization as the cause of precision loss.

5. Conclusion

In this paper, we assessed a new approach to biomedical document indexing and retrieval in which morphologically complex word forms, which appear in both queries and documents, are segmented into domain-relevant subwords and subsequently submitted to the matching procedure. This way, the impact of word form alterations can be eliminated from the retrieval procedure. We evaluated our hypothesis on a large collection of medical documents annotated for a total of 52 test queries, which were formulated by medical experts and laymen. Our experiments lend support to the hypothesis that a document index built of subwords performs better than a conventional index. With respect to semantic matching our expectations were not met, so far, even for the laymen queries, where we had expected a more accentuated vocabulary mismatch, and, as a consequence, a significant improvement by synonym mapping. Here, obviously the additional noise counterbalanced the gain of retrieving synonyms. In the future, we will make a more restrictive use of synonym mapping in our ongoing thesaurus development.

References

[1] Y. Choueka. RESPONSA: An operational full-text retrieval system with linguistic components for large corpora. In A. Zampolli, editor, Computational Lexicology and Lexicography., pages 181–217. Pisa: Giardini Press, 1992.

[2] U.Hahn, M. Honeck, M. Piotrowski, and S. Schulz. Subword segmentation: Leveling out morphological variations for medical document retrieval. In Proceedings of the 2001 AMIA Fall Symposium, pages 229–234, 2001.

[3] H. Jäppinen and J. Niemistö. Inflections and compounds: Some linguistic problems for automatic indexing. In Proceedings of the RIAO'88 Conference, pages 333–342, 1988.

[4] W. Kraaij and R. Pohlmann. Viewing stemming as recall enhancement. In Proceedings of the 19th ACM SIGIR Conference, pages 40–48, 1996.

[5] J. Lovins. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(1/2):22–31, 1968.

[6] MSD-Manual der Diagnostik und Therapie, 5. Auflage. CD-ROM, 1993.

[7] L. Norton and M. Pacak. Morphosemantic analysis of compound word forms denoting surgical procedures. Methods of Information in Medicine, 22(1):29–36, 1983.

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

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

[10] C. Rijsbergen. Information Retrieval. London: Butterworths, 1979.

[11] G. Salton. Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer. Reading, MA: Addison-Wesley, 1989.

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

[13] S. Schulz, M. Honeck, and Udo Hahn. Indexing medical WWW documents by morphemes. In MEDINFO'01 – Proceedings of the 10th World Congress on Medical Informatics, pages 266–270, 2001.

[14] P. Zweigenbaum, SJ Darmoni, and N Grabar. The contribution of morphological knowledge to French MeSH mapping for information retrieval. In Proceedings of the 2001 AMIA Fall Symposium, 796–800, 2001.