Subword Segmentation — Leveling out Morphological Variations for Medical Document Retrieval

Udo Hahn [a] Martin Honeck [b] Michael Piotrowski [a] Stefan Schulz [b]

[a]Freiburg University, Text Knowledge Engineering Lab (`http://www.coling.uni-freiburg.de`)

[b]Freiburg University Hospital, Department of Medical Informatics (`http://www.imbi.uni-freiburg.de/medinf`)

Abstract

Many lexical items from medical sublanguages exhibit a complex morphological structure that is hard to account for by simple string matching (e.g., truncation). While inflection is usually easy to deal with, productive morphological processes in terms of derivation and (single-word) composition constitute a real challenge. We here propose an approach in which morphologically complex word forms are segmented into medically significant subwords. After segmentation, both query terms and document terms are submitted to the matching procedure. This way, problems arising from morphologically motivated word form alterations can be eliminated from the retrieval procedure. We provide empirical data which reveals that subword-based indexing and retrieval performs significantly better than conventional string matching approaches.

INTRODUCTION

The Internet, intranets, as well as the electronic patient record expose health professionals to increasingly larger amounts of computer-readable text, while WWW-based portals provide consumers and patients with ever-growing volumes of health-related information. The full utilization of these resources, however, is currently hampered by inadequate retrieval facilities. Often, relevant documents which contain morphological variants of a search term are not retrieved so that the recall performance of IR systems decreases [2, 8, 9]. A query term such as *"Leukocyte"* retrieves only those documents in which this term occurs literally, but fails to cover morphological variants, e.g., *"leukocytic"*.

Such morphological variants can generally be described as concatenations of a basic lexical forms (stems) with additional substrings (affixes). We distinguish three kinds of morphological processes, *viz.* inflection (e.g., adding plural *s* in *"Leukocyte⊕s"*,[1] derivation (e.g., attaching the derivation suffix *ic* in *"leukocyt⊕ic"*), and composition (e.g., in *"Leuk⊕em⊕ia"*).

The efforts required for performing morphological analysis vary between languages and application domains. Whereas the English language is known for the limited number of inflection patterns, others, e.g., German, Dutch or Russian are much more diverse. Therefore, English general-purpose stemming algorithms available for IR applications [11, 14] have no counterparts in these morphologically richer languages. When derivation and composition phenomena have to be considered, too, even for the English language only restricted, domain-specific algorithms yet exist.

This is particularly true for the medical domain. While one may argue that single-word compounds are quite rare in English (which is not the case in the medical domain either, cf. [18]), this is certainly not true for the German language and related ones known for excessive single-word nominal compounding. Besides fairly standardized noun compounds, which already form a common part of the medical terminology, a myriad of *ad hoc* compounds are formed on the fly which cannot be anticipated when formulating a retrieval query, though they appear in relevant documents. Hence, morphological analysis is mandatory for optimal retrieval results.

Unlike other sublanguages, medical terminology is also characterized by a typical mix of Latin and Greek roots with the corresponding host language (e.g., *"zerebrovaskulär"* in German or *"proctosigmoidoscopy"* in English), often referred to as *neoclassical compounding*. While this is not even a side issue for general-purpose morphological analyzers, the need to deal with such phenomena is crucial for any attempt to cope adequately with medical free texts in an IR setting (cf. also [22]).

In this paper, we propose an approach to document retrieval where query and document terms are segmented into basic, medically plausible subword units. This approach provides a homogeneous treatment for deflection (stripping off inflectional suffixes), dederivation (stripping off derivational suffixes), and decomposition (separating a complex single-word compound into its constituent word stems).

MODEL FOR MORPHOLOGICAL ANALYSIS

Two basic approaches to deal with morphological variation can be thought of. In the first, at least the deriva-

[1] '⊕' denotes the string concatenation operator.

tional and compositional forms have to be explicitly spelt out for each item. This causes the size of dictionaries to grow considerably by the sheer number of different terms. Also, given the speed of terminological change and growth, the goal of enumerating all morphological varieties can always only be approximated, while *ad hoc* compounds cannot be accounted for at all.

We propose an alternative approach that avoids these scaling problems and keeps up with continuous terminological dynamics by exploiting basic linguistic regularities through a morphological analyzer. When we subscribe to the subword model, corresponding dictionaries or thesauri are expected to be several orders of magnitude smaller than phrasal or fully lexicalized dictionaries. This parsimony must, however, be traded against the reduced level of semantic discrimination of the subword units as compared with the associated compounds/derivates.

In standard linguistic approaches, *morphemes* are chosen as nondecomposable entities and defined as the smallest content-bearing (*stem*) or grammatically relevant (*affixes*) units. *Subwords* differ from morphemes only, if the meaning of a combination of linguistically significant morphemes is (almost) equal to that of another nondecomposable medical synonym. This way, subwords preserve a sublanguage-specific composite meaning that would get lost, if they were split up into their linguistically legitimate constituent (morpheme) parts. Hence, we trade linguistic atomicity against medical plausibility considerations and assume that the latter are beneficial for boosting the system's retrieval performance. As an example, a medically reasonable minimal segmentation of *'diaphysis'* into *'diaphys⊕is'* will be preferred over a linguistically motivated one (*'dia⊕phys⊕is'*), because in the former case *'diaphys'* can be mapped to the quasi-synonym stem *'shaft'*. Such a mapping would not be possible with the overly unspecific morphemes *'dia'* and *'phys'*, which occur in numerous other contexts as well. Hence, a decrease of the precision of the retrieval system would be highly likely due to over-stemming.

We then distinguish between the following decomposition classes:

Subwords like {*'gastr', 'hepat', 'nier', 'leuk', 'diaphys'*, ...} are the primary content carriers in a word. They can be prefixed, linked by infixes, and suffixed. As a particularity, *short words*, generally with four characters or less, like *'ion', 'gene', 'ovum'*, are classified separately applying stricter rules (e.g., they cannot be composed at all). We intentionally exclude their very short stems (e.g., *'ion', 'gen', 'ov'*) from being listed in the subword dictionary in order to avoid a large number of ambiguities. The same applies to acronyms such as *'AIDS', 'ECG'*, which are also classified as nondecomposable entities.

Prefixes like {*'a-', 'de-', 'ver-', 'anti-'*, ...} precede a subword.

Infixes (e.g., *'-o-'* in *"gastr⊕o⊕intestinal"*) are used as a (phonologically motivated) 'glue' between subwords.

Derivational suffixes such as {*'-io-', '-ion-', '-itis-', '-tomie-'*, ...} usually follow a subword.

Inflectional suffixes like {*'-e', '-en', '-s', '-idis', '-ae', '-oris'*, ...} appear at the very end of a composite word form following the subwords or derivational suffixes.

The German-language *subword dictionary* underlying this study is composed of 4,630 subwords (and short words), 344 proper names, and an *affix list* composed of 139 prefixes, 8 infixes and 154 suffixes, making up 5,275 entries in total. As a further enhancement, we enriched the subword dictionary with the simple semantic relation EQ which links subwords that stand in a semantic *equivalence* relation to each other. This extension is particularly directed at foreign-language translates (mostly Greek or Latin terms) of source language terms, e.g., German *'nier'* EQ Latin *'ren'* (EQ English *'kidney'*). For details of the construction of the subword dictionary, cf. [19].

The morphological segmentation engine builds *all* possible morphological segmentations for an input word using the above-mentioned resources and concatenation regularities. Ambiguous morphological segmentations of an input word are ranked according to several preference criteria, including *longest match from the left, minimal number of stems per word*, and *minimal number of consecutive affixes*, as well as a semantic weight factor assigned to all subwords and affixes.

RETRIEVAL EXPERIMENTS

The *document collection* for our experiments consists of the CD-ROM edition of a standard handbook of clinical medicine, the *"MSD - Manual der Diagnostik und Therapie"*, a close though not fully parallel translation of *"The Merck Manual of Diagnosis and Therapy"*. It contains 5,517 articles (about 2.4 million text tokens) on a broad range of clinical medical knowledge. Since we envisage the routine application of our approach in a nonexperimental, highly standardized system framework, we chose the *AltaVista™Search Engine 3.0* as our testbed.[2] All terms from the doc-

[2] The *AltaVista™Search Engine 3.0* (http://solutions.altavista.com/downloads/downloads.html) is a widely distributed, easy to install off-the-shelf IR system. The system manual is not fully conclusive about the details of index term processing but the following criteria are mentioned. *"The relevancy of a document is determined by*

ument collection are assembled in an inverted term *index* accessible for retrieval. The search engine then produces a ranked output of documents (in our experiments, we set the cut-off value to the top 200 documents retrieved).

The *user query collection* was acquired as follows: 63 medical students (between the 3rd and 5th study year) were presented a random selection of multiple choice questions from the nationally standardized year 5 examination questionnaire for medical students in Germany.[3] Then we asked them to formulate free-form natural language queries intended to help finding the correct answer to the MC question. Ten topics were assigned to each student at random. So we ended up with 630 queries, from which 25 were randomly chosen for our experiments. The relevance judgments came from three 6th-year medical students, identifying relevant documents in the whole test collection for each of the 25 queries.[4] We conducted the following experiments:

Test 1: Token Search. No term processing precedes indexing and the submission of the query for retrieval. The search was run on the index covering the entire MSD document collection (182,306 index terms). This scenario serves as the baseline for determining the benefits of our approach.

Test 2: Token Search with Stemming. Text tokens in the documents and in the queries were submitted to the operation of the (language-specific) stemmer included as an add-on feature in the *AltaVista™Search Engine*.

Test 3: Morphological Segmentation. Text tokens in the documents and in the queries were submitted to the subword-based retrieval approach described above. Morphological segmentation resulted in a decrease of the size of the index, with 39,315 index terms remaining. This amounts to a reduction rate of 78% compared with the original number of index terms in the MSD.

Test 4: Morphological Segmentation and Synonym Expansion. The simple subword model is augmented by introducing the EQ semantic relation between suitable subwords into the retrieval procedure. In the documents, as well as in the queries, each known word form was substituted by an alphabetic code identifying the corresponding thesaurus class.[5]

	Test 1	Test 2	Test 3	Test 4
Recall(%)	Precision(%) over 25 queries of top r=200 retr. documents			
0	53.6	43.3	69.4	66.9
10	51.7	42.1	65.5	60.5
20	45.4	37.6	61.4	54.9
30	34.9	33.3	55.4	51.6
40	29.5	30.5	51.4	46.7
50	27.8	29.7	49.7	44.1
60	26.2	27.1	40.7	39.2
70	18.1	19.7	32.6	31.7
80	15.2	17.4	26.3	22.4
90	5.6	5.4	20.1	11.4
100	5.4	5.3	16.3	11.0
3pt avrg	29.5	28.2	45.8	40.5
11pt avrg	28.5	26.5	44.4	40.0

Figure 1: Evaluation Results – Precision/Recall Table

The assessment of the experimental results is based on the aggregation of all 25 selected 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 *AltaVista™Search Engine*. The corresponding P/R values for all four test scenarios are summarized in Figure 1 and visualized in Figure 2.

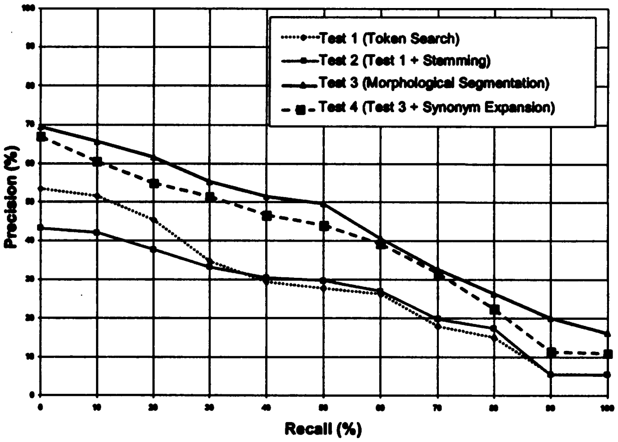

Figure 2: Evaluation Results – Precision/Recall Graph

the frequency of words, the position of words in documents, whether the words appear in the document title, whether the complete phrase exists for multi-word query, and the proximity of words to each other within documents." [p. 6, Software Product Description 052000]. Additional term processing tools (spelling correction, phrase recognition, thesaurus, stemming, etc.) were disabled except for Test 2.

[3] In order to match the content of the document collection, only questions referring to *clinical* disciplines were selected.

[4] Due to the high workload implied by matching 25 queries with 5,517 documents, we were not able to hire medical doctors for the rating, nor could we assess the inter-rater reliability. In order to avoid biases, knowledge about our indexing and retrieval method was not disclosed to the raters.

[5] This experiment should also be run using a 'standard thesaurus' for the medical domain. We did not perform such an experiment for the following reasons: First, a stemming algorithm accounting for the properties of German medical language (including Greco-Latin derivates) was not readily available. Second, the clinical thesaurus most commonly used, the German MeSH [3], proved to provide insufficient coverage, since from 77 words, acronyms or noun phrases contained in the 25 selected queries, only 40 were identified in the thesaurus — even after manual reduction of all query terms to their respective base form. This remarkable result coincides both with our previous findings [18], as well as those from Hersh et al.'s study on cross-language medical information retrieval [6].

For our baseline, *Test 1*, the direct match between query terms and document terms, precision is already poor at low recall points ($R \leq 30$), ranging in an interval from 54% to 35%. At high recall points ($R \geq 70$), precision drops from 18% to 5%. Adding the (German) stemming procedure of the *AltaVista™Search Engine* in *Test 2* (surprisingly) increases noise in the system, since precision values drop by a factor 10% for real low recall values, while for high ones the precision curve almost overlaps with that for searches without stemming, showing no significant improvement.

The subword approach in *Test 3* clearly outperforms the results achieved for *Test 1* and *Test 2*. For low recall values the gain in precision ranges from 14% to 21%, while for high recall values the gain is still in the range of 11% to 15%. Adding equivalent terms slightly decreases the performance of our basic approach, roughly on the order of 5%. This indicates that truly equivalent terms are hard to determine, even in the medical domain. Since the addition of equivalent terms produced no advantage over simple segmentation into subwords, we cannot recommend their inclusion into the search process on the basis of our data.

In order to estimate the statistical significance of this result, we compared relevant test pairs for each fixed recall level, using the two-tailed sign test (for a description and its applicability for the interpretation of P/R graphs, cf. [17]), and obtained the results summarized in Figure 3.

Recall Level	Test2 vs. Test1	Test3 vs. Test1	Test4 vs. Test1	Test2 vs. Test3	Test4 vs. Test3
0%	< 0.05	n.s.	n.s..	< 0.005	n.s.
10%	n.s.	< 0.05	n.s.	< 0.005	n.s.
20%	< 0.05	< 0.05	n.s.	< 0.05	n.s.
30%	n.s.	< 0.05	n.s.	< 0.05	n.s.
40%	n.s.	< 0.05	< 0.05	< 0.05	n.s.
50%	n.s.	< 0.005	< 0.05	n.s.	n.s.
60%	n.s.	n.s.	n.s.	n.s.	n.s.
70%	n.s.	n.s.	n.s.	< 0.05	n.s.
80%	n.s.	n.s.	n.s.	n.s.	n.s.
90%	n.s.	< 0.005	< 0.05	< 0.005	n.s.
100%	n.s.	< 0.005	< 0.05	< 0.005	n.s.

Figure 3: P-values for Relevant Test Pairs at Fixed Recall Levels

Generalizing the interpretation of our data in the light of these findings, we recognize a substantial increase of retrieval performance when query and text tokens are segmented according to the principles of the subword model. The benefit we achieve is not dependent on whether we aim at maximizing precision or recall. No benefit at all was found for the built-in German language stemmer of the search engine – at least in our domain. Surprisingly, the resolution of synonyms did not increase the performance either. We ascribe this to the fact that the terminology used by the students in the queries was nearly identical to the one occurring in the documents of the test collection. We expect a different result in a scenario where terminology mismatches between queries and documents occur (e.g., common-sense queries posed by people without deep medical expertise). Currently, we run a second evaluation using original queries of non-expert users of a health-specific WWW site as input.

RELATED WORK

Effectiveness of simple stemming [11, 14] for document retrieval has been discussed controversially [5, 7, 10]. The key issue seems to be the presence of a dictionary whose positive impact on document retrieval has been described by [10, 9, 20].

The earliest approach which deals with medical terminology by way of morphological analysis is due to Pratt and Pacak [15]. Their approach transformed semantically equivalent adjectival and nominal forms by employing simple suffix trees and transformation rules for recoding morphologically reduced forms. Transformations succeed if a recoded form is matched with an entry in the dictionary, *viz.* the SNOP nomenclature. Follow-up studies [13, 12] which focused on the suffixes *'-itis'*, *'-ectomy'* and *'-plasty'* not only determined a preferred normalized form for several morphological variants but also computed paraphrase and other semantic relations which can be made explicit by breaking up compounds into their constituent parts. The distributional patterns suggested by Pacak and Norton are based on the four axes of SNOP/SNOMED. In a similar vein, Dujols et al. [4] treated *'-osis'* forms only. These restrictions were somewhat weakened in the work of Wolff [22], both in terms of a larger coverage of Greco-Latin suffixes, as well as more general compositional patterns of neo-classical compounding. In an attempt to formulate the principles of medical word segmentation in a formally rigid, almost language-independent framework, Wingert chose an automaton-based specification for morphological analysis in terms of augmented transition networks [21]. He arrived at a set of 255 cascading rules to capture the combinatorial regularities of different morpheme classes and, similar to Pratt & Pacak, referred to the entries of the SNOP nomenclature.

For almost one decade, research in morphology ceased in the Medical Language Processing (MLP) community. Just recently, interest in this topic was revived by work employing much more sophisticated linguistic and conceptual knowledge. Baud et al. [1] use finite-state technology for the decomposition of complex terms into semantic units they refer to as *morphosemantemes*. A lot of the power of their approach derives from the fact that conceptual correlates of these morphosemantemes no longer refer to flat SNOMED-style categories but are rather formulated in GRAIL, a highly expressive deductive terminological knowledge representation language within the GALEN framework [16]. In order to isolate a morphosemanteme, composite concepts are dissected to their medically plausible conceptual core, using terminological knowledge derived from GRAIL. Since GRAIL's coverage of the medical domain is fairly limited, this dependence might constitute a crucial factor for hampering routine usage due to scaling problems of the underlying knowledge base. We diverge from previous approaches in that our model of morphological analysis covers the entire range of clinical medical terminology. On the other hand, we do not attempt any form of semantic interpretation (neither SNOMED-, nor GRAIL-style) during dederivation and decomposition.

CONCLUSION

In this paper, we argued that natural languages with a rich morphology — in terms of derivation and (single-word) composition — face serious performance problems with the direct query-term-to-text-word matching paradigm that underlies the vast majority of standard document retrieval systems. Therefore, we proposed an approach – especially adapted to the medical domain – in which morphologically complex word forms, which appear in both query 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 with a common search engine on a large collection of medical documents. Our experiments lent (mostly statistically significant) support to the subword hypothesis.

References

[1] R. Baud, C. Lovis, A. Rassinoux, and J. Scherrer. Morpho-semantic parsing of medical expressions. In *Proceedings of the 1998 AMIA Annual Fall Symposium*, pages 760–764, 1998.

[2] 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.

[3] DIMDI. *Medical Subject Headings. German Version.* Deutsches Institut für Medizinische Dokumentation und Information, 2000.

[4] P. Dujols, P. Aubas, C. Baylon, and F. Grémy. Morphosemantic analysis and translation of medical compound terms. *Methods of Information in Medicine*, 30(1):30–35, 1991.

[5] D. Harman. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1):7–15, 1991.

[6] W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the UMLS metathesaurus. In *Proceedings of the 1998 AMIA Fall Symposium*, pages 344–348, 2000.

[7] D. Hull. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84, 1996.

[8] 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.

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

[10] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th ACM SIGIR Conference*, pages 191–203, 1993.

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

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

[13] 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.

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

[15] A. Pratt and M. Pacak. Identification and transformation of terminal morphemes in medical English. *Methods of Information in Medicine*, 8(2):84–90, 1969.

[16] A. Rector, S. Bechhofer, C. Goble, I. Horrocks, W. Nowlan, and W. Solomon. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, 9:139–171, 1997.

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

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

[19] S. Schulz, M. Honeck, and U. Hahn. Indexing medical WWW documents by morphemes. In *Proceedings of the 10th World Congress on Medical Informatics – MedInfo 2001*, 2001.

[20] E. Tzoukermann, J. Klavans, and C. Jacquemin. Effective use of natural language processing techniques for automatic conflation of multi-word terms: The role of derivational morphology, part of speech tagging, and shallow parsing. In *Proceedings of the 20th ACM SIGIR Conference*, pages 148–155, 1997.

[21] F. Wingert. Morphologic analysis of compound words. *Methods of Information in Medicine*, 24(3):155–162, 1985.

[22] S. Wolff. The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. *Methods of Information in Medicine*, 23(4):195–203, 1984.