Feature Article

Optimizing and evaluating the MEDPILOT search rngine. Boosting medical information retrieval by using a morpheme thesaurus

Waldemar Dzeyk[1] **Kornél Markó**[2]

1) German National Library of Medicine, Cologne, Germany
2) Averbis GmbH, Freiburg, Germany
Contacts: waldemar.dzeyk@zbmed.de
marko@averbis.de

Abstract
This article describes the implementation and evaluation of a computational linguistic approach to improve the quality of the MEDPILOT medical search engine, maintained by the German National Library of Medicine. At the core of the system lies a new type of multilingual dictionary, in which entries are equivalence classes of morphemes, i.e. semantically minimal units. Documents, as well as user queries, are mapped to these language-independent (conceptual) classes on which retrieval operations are performed. Early results of the evaluation have shown that the used language technology has many advantages in medical information retrieval. In combination with up-to-date search software of the linguistic approach leads to more and better results (i.e. relevant hits) for phenomena such as synonyms, translations and linguistic variants (inflection, derivation, word-composition, etc.). Additionally, a normalization of laymen and expert queries can be achieved. The formal structure of user queries as well as the information needs of MEDPILOT users were analyzed in detail. Results and consequences for the development of user centered design and usability improvements of the search engine will be discussed.

Key words: information storage and retrieval; medical search engines; matural language processing; evaluation; usability.

Introduction and background

The German National Library of Medicine (ZB MED) is the largest European medical library. It maintains numerous bibliographic databases, which are accessible online via the MEDPILOT-Portal (1). MEDPILOT is a cooperation project of the ZB MED (2) and DIMDI (German Institute of Medical Documentation and Information) (3). With a single request, users of MEDPILOT can simultaneously research a wide range of medical sources and databases by a meta search system.

In this context, the challenges of medical information retrieval (4) in scientific databases are manifold. Firstly, the retrieval system stores unstructured or semi-structured text (such as title, authors, abstract, etc.), and hence, does not

"know" anything about the content. Secondly, the information need of searchers is vague and cannot be formally expressed. Search engines are designed to find those relevant documents of a collection, which *somehow* best fit to a particular user query – as selective as possible. Obviously, these circumstances can hardly be considered helpful in the process of the formal evaluation of such systems.

An additional complexity emerges from the multilingual dimension of information retrieval applied to the medical domain. While clinical documents are typically written in the physicians' native language, searches in scientific databases require sophisticated knowledge of (expert-level) English medical terminology, which most non-English speaking physicians do not have. Hence, some sort of bridging between synonymous or, at least, closely related terms from different languages has to be realized to make use of the information these databases hold.

Furthermore, the user population of medical document retrieval systems and their search strategies are really diverse. Not only physicians, but also nurses, medical insurance companies and patients are increasingly obtaining access to these resources, with the Web adding an even more scattered crowd of searchers. Hence, mappings between different jargons and sublanguages are inevitable to serve the needs of such a heterogeneous searcher community. The simplicity of the content representation of the documents, as well as automatically performed intra- and interlingual lexical mappings or transformations of equivalent expressions become crucial issues for an adequate methodology of medical information retrieval.

In a current project (5), a commercial retrieval solution has been integrated into the MEDPILOT search portal. It is provided by Averbis GmbH (6), a spin-off of the Freiburg University Hospital (Germany). The company offers search and classification solutions particularly adapted to health care needs. The Averbis core technology is specifically designed for the consideration of linguistic variants (such as inflection, derivation, word-composition, etc.) and focuses on the transfer from expert to common language as well as multilingual search in currently seven European languages (7, 8).

Further objectives of this project are, firstly, to analyse the searchers' needs in order to customize the new search engine, and, secondly, to provide a gold standard for measuring the retrieval performance of MEDPILOT before and after the integration of the new search platform. In the second half of 2007 an expert team at ZB MED carried out an evaluation on the efficiency of the new search system, for which preliminary results are presented in this article.

Methods

Technical background and implementation

In a common free-text information retrieval environment (such as the current MEDPILOT implementation), the search for a particular document is based on a (exact) pattern matching operation between the query term(s) and the document terms. Therefore, a query term such as *leukocytes* retrieves all those documents in which this query term occurs literally. On the other hand, documents containing the singular form *leukocyte*, the adjective *leukocytic*, or the compound term *leukemia* cannot be found.

In contrast, the Averbis technology is based upon a unique semantic analysis of texts. Relevant sections - no matter whether parts of words, whole words or phrases - are identified using a new type of dictionary, in which the entries are subwords, i.e., semantically minimal, morpheme-style units. Language specific subwords, which have the same meaning, are linked within and across languages and grouped in terms of concept-like equivalence classes at the layer of a language-independent interlingua.

This methodology started from the assumption that neither fully inflected nor automatically stemmed words – such as is common in many text retrieval systems – constitute the appropriate granularity level for lexicalized content description. Especially in the medical domain, a high frequency of domain-specific suffixes (e.g. "-itis", "-ectomia") and numerous occurrences of complex word forms such as "pseudo | hypo | para | thyroid | ism" or "gluco | corticoid | s" have to be considered. To properly account for these particularities, the notion of subwords, i.e. self-contained, semantically minimal units had been introduced. Thus, completely invisible to the user, the documents are reduced to their essential semantic elements.

By combining linguistic-semantic analysis and state-of-the-art retrieval technologies, documents are normalized and disposed of linguistic variations. Thus, it no longer makes any difference whether searching for "myocarditis", "inflammation of the heart muscle" or German "Herzmuskelentzündung" (*Figure 1*). The described retrieval solution provides every relevant match.

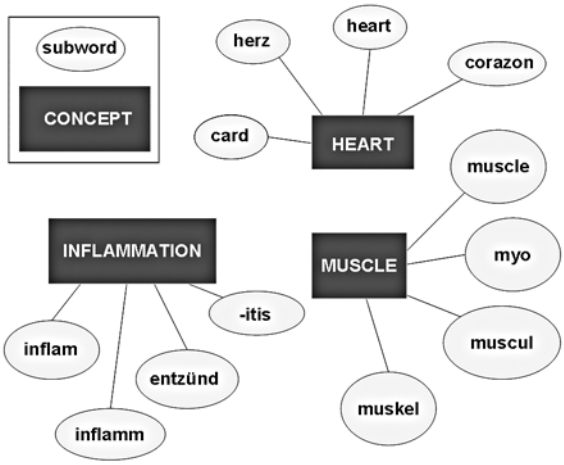

Fig. 1. *Multilingual language repository.*

In the current stage of the project, three different bibliographic resources have been made searchable with the new technology: Medline, CCMED (Current Contents in Medicine), and the online public access catalog (OPAC) of the German National Library of Medicine, resulting in a total of 15.7 million bibliographic items.

Evaluation methods

The evaluation process comprised three stages:

- a content analysis of user queries to explore the users' information needs and to discover the scope of special contents such as acronyms, linguistic variations, the use of classical chemical terms, pharmaceutical trade names or brands;
- development and evaluation of a set of test collections to compare the old and the new search system regarding the relevance of hits (9);
- optimizing the usability of the search engine user interface. Usability plays an important role for satisfaction and users trust on a website (10, 11).

Content analysis of the MEDPILOT logfile

There were two main reasons for conducting a content analysis. Firstly, for the examination of the relevance of search engine hits, there is a need to construct realistic search queries (and test collections of them). The best way to do this is to find out what kind of medical information, real users of MEDPILOT are looking for. Secondly, it is very important to know about the structure and the content of doctors and students' information needs for the optimization of the search engine user interface.

Consequently, the logs of MEDPILOT queries were analyzed in detail (12). We extracted queries covering seven months (142,922 queries) and drew a random sample of 10,000 queries. Afterwards, we developed a category-system with 24 classes which was constructed by an interplay of deductive and inductive procedures (13). On the one hand, it was based upon what is known regarding medical information seeking (14) and on the other, the content of the material was analyzed. A validation of the category-system was made by two evaluators who examined 150 of the queries. For this task, an intercoder reliability

of 88% was achieved, which can be seen as a sufficient correlation (15). Then, each of the 10,000 queries was assigned to one or more categories by a domain expert.

The content analysis was carried out with the following questions in mind:

- What kind of content are users of MEDPILOT interested in?
- How complex are the search queries? How many words are used for searching medical content?
- To what extent were Boolean operators or field search used?
- To what extent were medical acronyms/ abbreviations used in queries?
- What were the typical misspellings and how many errors were made?
- To what extent do users search for classical chemistry, biochemical content, drugs or brands?

Development of test collections

Averbis has installed a test system with the linguistic search engine which was evaluated by ZB MED project members, who are familiar with both medical and biological terminology. Due to the fact that no one can estimate how many of the million items of Medline, for instance, are relevant to a special query (recall), we had to choose a pragmatic way to evaluate our data. For this purpose, two indicators of the performance of information retrieval systems were evaluated: the quantity and the quality (precision) of hits. Each test collection that has been used to compare the output of the old with the new system consisted of 50 queries and reflected a special linguistic aspect which we had examined, which are, amongst others:

1. misspellings;
2. acronyms / abbreviations;
3. synonyms;
4. word-compositions;
5. translations;
6. layman – expert queries.

We compared the quantity (number of hits) and quality (number of relevant hits out of the first 20) of results after querying the Medline and CCMED databases, each within the old MEDPILOT search engine and the new Averbis system. In the future we will construct one test collection which we can use as a representative sample for a wide range of linguistic phenomena.

Optimizing usability

The task of optimizing the usability of the Averbis user interface has not been performed yet. Over the next months, doctors and students will participate as subjects for usability-tests. As a result, we will modify the medical search interface for better, user-friendly navigation.

Early evaluation results

About a third (35.9%) of the search queries in MEDPILOT consist of one single word. The other 30% of queries contain only two words; 16% of the queries consist of three terms and 6.6% comprise four words. Five words were still used in 3.7% of the search queries and 7.8% of the queries contain more than five words. To summarize, one and two-word queries are the most frequent strategies for searching medical content (nearly 66% of the queries). The examination of the log files revealed that field searching and Boolean operators – others than AND – were rarely used. Acronyms/abbreviations were used in 5.36% while misspellings were discovered in 4.58% of the queries.

Figure 2 illustrates the type of medical content which users of MEDPILOT searched for. The top categories of medical information needs were diseases, *syndromes, symptoms,* (30.85%), *methods of treatment, therapy and diagnostics* (28.45%) and content about *social medicine, statistics, studies, and epidemiology* (15.58%). One reason for carrying out a content analysis was to investigate to what extent the queries comprised classical chemistry concepts. By studying the results, we noticed that only 2.1% of the queries were

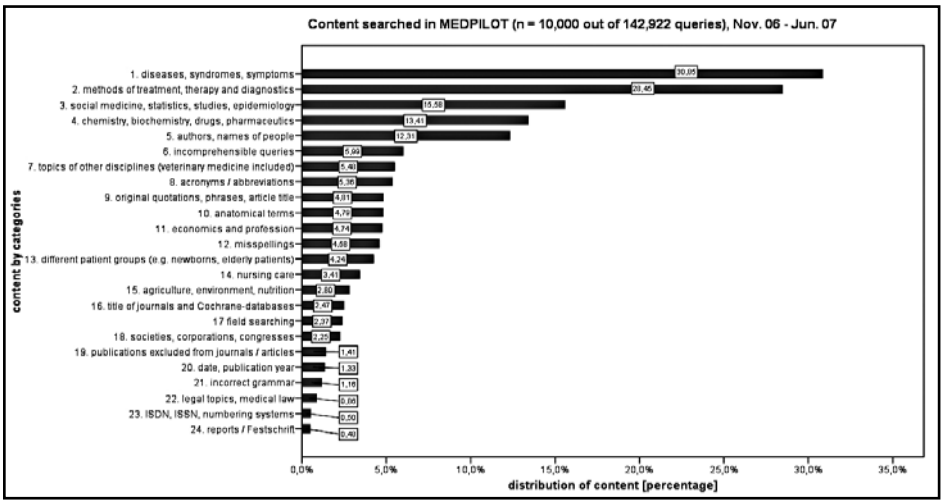

Fig. 2. *Content searched in the medical search engine MEDPILOT.*

classical chemistry terms. The other information needs in category 4 were biochemistry, drugs, pharmaceutics and brands.

Preliminary results indicate that the Averbis system has remarkable advantages both in quantity and quality of the results. A first estimation for the databases Medline and CCMED revealed that in the old system the average number of hits is 13.5 times smaller than in the new one. At the same time, when comparing the quality of hits (precision for the top 20 results) users can expect an average of 20-30% more relevant hits through the new Averbis system. On average, 60% of the hits of the top 20 results were relevant. At first sight, improvements can still be achieved by better recognizing misspellings and acronyms /abbreviations. Detailed results will be published soon.

Discussion

First results of this project illustrate that the new technique of using a domain specific morpheme thesaurus in a search engine, remarkably boosts the performance when searching for medical content. Limitations of the approach are, firstly, false word segmentations which can not be prevented entirely: For example, on a formal level, the segmentation of the word "nephrotomy" can lead to "nephr-o-tomy" with the concepts *kidney* and *incision* (with "o" being a syntactical infix), but also to "nephr-oto-my", denoting *kidney*, *ear* and *muscle*. Such erroneous interpretations can only be avoided by the manual revision of the morpheme-thesaurus. To some extent, false segmentations can also be counterbalanced by the (relevance-based) ranking mechanism of the underlying search engine. Secondly, ambiguity of words can lead to irrelevant search results, as indicated by, e.g. the word "patient" which has completely different meanings when occurring as a noun or an adjective (a limitation that practically affects any retrieval system). In future, context-sensitive disambiguation methods (16) are incorporated in the language processing engine in order to further enhance the overall performance of MEDPILOT's search behavior.

By analyzing the queries of physicians and medical students, we recognized that the complexity of the used search terms is rather poor. It is almost impossible (at least for search engines) to predict the users' information needs precisely when using only one word to search for specific medical information. As a consequence, we have to think about technical support for users so that they will be better enabled to specify their needs. One solution to this problem might be the integration of a MeSH-term navigation for browsing or to automatically suggest query-associated search terms (e.g. anatomical regions or treatments

for a query denoting a disease). The latter would be the approach of our choice.

Conclusion

The primary goal of this project is to satisfy the users' needs and wishes when they are using a medical search engine. In this context, the Averbis system adds a substantial value to the MEDPILOT search portal. However, as a consequence of this study, additional strategies have to be developed and implemented in order to assist the users to specify their information needs precisely. This can be regarded as a prerequisite for search engines to continuously deliver high quality results. The forthcoming process of evaluation will also comprise a validation of the results by real users, therefore we will start usability tests in the near future.

Acknowledgements

Our thanks go to the team members of the ZB MED Maarit Stoor and Stefanie Paschke as well as to Natascha Dahmen (student assistant), Ulrich Korwitz, the director of the ZB MED and to the software engineers at Averbis. This project is supported financially by the German program *Pact for Research and Innovation*, an initiative of the Federal Government's Campaign for Innovation and Growth. Duration of the project is June 2007 - December 2008.

References

1. Medical search engine MEDPILOT. Available from: http://www.medpilot.de
2. German National Library of Medicine (ZB MED). Available from: http://www.zbmed.de
3. German Institute of Medical Documentation and Information (DIMDI): http://www.dimdi.de
4. Hersh WR. Information retrieval. A health and piomedical perspective (2nd edition). New York: Springer. 2002.
5. Project "Optimisation of the MEDPILOT search: multilingualism and normalization of linguistic variations": Available from: http://www.zbmed.de/morphosaurus.html
6. Averbis GmbH. Available from: http://www.averbis.de
7. Markó K, Schulz S, Hahn U. MorphoSaurus - Design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. Methods of Information in Medicine. 2005; 44(4):537-545.
8. Markó K, Daumke P, Schulz S, Klar R, Hahn U. Large-scale evaluation of a medical cross-language information retrieval system. Proceedings of the 12th World Congress on Medical Informatics, Brisbane, Australia. 2007; 392-396.
9. Lewandowski D. Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen? In: Die Macht der Suchmaschinen. / The Power of Search Engines. Machill M, Beiler M, editors. Köln: Herbert von Halem Verlag. 2007; 243-258.
10. Flavián C, Guinalíu M, Guerra R. The role played by perceived usability, satisfaction and consumer trust on website loyalty. Information & Management. 2006; 43:1-14.
11. Dzeyk W. Vertrauen in Internetangebote. Saarbrücken: VDM Verlag. 2007.
12. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. A day in the life of PubMed: analysis of a typical day's query log. J Am Med Inform Assoc. 2007;14:212-220.
13. Rustemeyer R. Praktisch-methodische schritte der inhaltsanalyse. Eine einführung am beispiel der analyse von interviewtexten. Münster: Aschendorff. 1992.
14. Davies K. The information-seeking behaviour of doctors: a review of evidence. Health Information and Libraries Journal. 2007; 24:78-94.
15. Wirtz M, Caspar F. Beurteilerübereinstimmung und beurteilerreliabilität. Methoden zur bestimmung und verbesserung der zuverlässigkeit von einschätzungen mittels kategoriensystemen und ratingskalen. Göttingen: Hogrefe. 2002.
16. Markó K, Schulz S, Hahn U. Unsupervised multilingual word sense disambiguation via an interlingua. Proceedings of the 20th National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania. 2005; 1075-1080.