Subword-based Semantic Retrieval of Clinical and Bibliographic Documents

P. Daumke[1,2]; S. Schulz[2]; M. L. Müller[3]; W. Dzeyk[4]; L. Prinzen[5]; E. J. Pacheco[6,7]; P. Secco Cancian[6]; P. Nohama[6,7]; K. Markó[1,2]

[1]AVERBIS GmbH, Freiburg, Germany;
[2]Freiburg University Hospital, Medical Informatics Group, Freiburg, Germany;
[3]Freiburg University Hospital, Department of Dermatology, Freiburg, Germany;
[4]German National Library of Medicine, Cologne, Germany;
[5]Bertelsmann Stiftung, Gütersloh, Germany;
[6]Paraná Catholic University, Health Informatics Laboratory, Curitiba, Brazil;
[7]Universidade Tecnológica Federal do Paraná, CPGEI, Curitiba, Brazil

Keywords

Thesauri, semantic information retrieval, document indexing, cross-language information retrieval, bibliographic databases, consumer health informatics, electronic health record

Summary

Objectives: The increasing amount of electronically available documents in bibliographic databases and the clinical documentation requires user-friendly techniques for content retrieval.

Methods: A domain-specific approach on semantic text indexing for document retrieval is presented. It is based on a subword thesaurus and maps the content of texts in different European languages to a common interlingual representation, which supports the search across multilingual document collections.

Results: Three use cases are presented where the semantic retrieval method has been implemented: a bibliographic database, a department EHR system, and a consumer-oriented Web portal.

Conclusions: It could be shown that a semantic indexing and retrieval approach, the performance of which had already been empirically assessed in prior studies, proved useful in different prototypical and routine scenarios and was well accepted by several user groups.

Correspondence to:
Philipp Daumke
AVERBIS GmbH
c/o IMBI
University Medical Center Freiburg
Stefan-Meier-Str. 26
79104 Freiburg
Germany
E-mail: daumke@averbis.de

Methods Inf Med 2010; 49: 141–147
doi: 10.3414/ME9303
prepublished: February 22, 2010

1. Introduction

Today more than half of all information produced worldwide is digitally available [1]. This is also true for the health domain with vast amounts of machine-readable texts in clinical information systems and e-publications. This is one of the reasons that justify the development of tailored information retrieval (IR) systems. Many research groups have been engaged in the adaptation of such systems to the health care domain [2, 3].

A complicating factor is that, especially in the Germanic languages, medical terms are often characterized by complex forms of composition, derivation and inflection, as well as by the constant generation of new acronyms, abbreviations and proper names. Moreover, spelling and syntax rules are not always respected. Due to these peculiarities, current information retrieval approaches that are usually based on simple comparison of entire words are inappropriate because they produce results that are incomplete, inaccurate, or outside the desired scope [4].

Considering the need of information recovery from documents other than clinical narratives, such as health-related articles, manuals and guidelines, the importance of English as source language should be taken into account. Non-native speakers, however, often have difficulties in formulating appropriate English queries in a document retrieval context [5]. This highlights the importance of multi- and cross-language IR techniques where the query language is different from the language of the documents.

For cross-language document retrieval, three main strategies may be distinguished [6, 7]:

- Query translation: only the query terms are translated, so the information used in the document indexes does not need to be modified when new languages are incorporated to the system. The computational load of this technique is the lowest [8].
- Document translation: May outperform the query translation as one term often occurs with morphological variations and synonyms in the same document. Its biggest drawback is the need to translate the totality of the documents [9].
- In interlingua techniques both documents and queries are translated to a common representation. This requires a lexicon of language-independent iden-

tifiers with mappings from all languages to be supported.

In the following, we present an interlingua-based retrieval system called MorphoSaurus[a], that takes into account the peculiarities of various Germanic languages and allows both intra- and cross-language search in biomedical data. In Section 3, two real-world applications based on the MorphoSaurus system are described.

2. The MorphoSaurus System

2.1 Subwords as Atomic Meaning Identifiers

The MorphoSaurus system [4, 5, 10, 11] maps the content of domain-specific texts to a concept-like interlingua. This entails a simplification and standardization of document source and user queries in order to facilitate the retrieval of documents in multilingual collections.

MorphoSaurus uses subwords as lexical units. A subword is defined as a minimum lexical unit for a meaningful term in a domain. This premise sets that the meaning cannot be anymore split. Thus, we can consider the term "*hepat* + *itis*" as a composition of two subwords, "*hepat*" and "*itis*" because its meaning results from the meaning of its constituents. A counterexample is "*hypophysis*", the meaning of which can definitively not be derived from "*hypo*" + "*physis*". Subwords therefore tend to be less granular than linguistic morphemes, but shorter than words.

In MorphoSaurus each subword entry has attributes such as language (currently *English* (*en*), *German* (*de*), *French* (*fr*), *Spanish* (*es*), *Portuguese* (*pt*), *Swedish* (*se*), and *Italian* (*it*)) and type (stem (ST), prefix (PF), suffix (SF), invariant (IV)). Subwords are contained in language-specific subword lexicons.

2.2 Thesaurus Structure

The semantic layer of MorphoSaurus is represented by equivalence classes, identified by so-called MIDs (MorphoSaurus identifiers). Each lexical entry is associated with exactly one equivalence class. Equivalence classes group lexical variants, synonyms, and translations. All MIDs together form the language-independent *thesaurus*.

The MorphoSaurus thesaurus uses two types of relations for linking equivalence classes, viz. "*has_word_part*" and "*has_sense*" (cf. ▶Fig. 1).

- The syntagmatic relation "*has_word_part*" links one MID to at least two other MIDs in order to "hide" semantic compositions. It is generally used for component terms that cannot be properly split by the segmentation routine, e.g. due to missing characters (e.g. in the word "*urinalysis*");
- The paradigmatic relation "*has_sense*" relates ambiguous MID to at least two other MIDs. This type of relationship is used to relate it to its possible meanings.

2.3 Peculiarities of MorphoSaurus' Lexical Resources

Lexicon builders' decisions of subword delimitation are driven not only by formal linguistic criteria, but also by the proper function of segmentation. This is especially relevant with long and composed words where different valid segmentations are possible. For example, "*nephrotomy*" may be segmented into nephr$^{[en,ST]}$ (#kidney) + o$^{[en;sp;pt]IN}$ + tomy$^{[en]PS}$ (#incision), but also in nephr$^{[en]ST}$ + oto$^{[en]ST}$ (#ear) + my$^{[en]ST}$ (#muscle). Only costly knowledge and language processing routines (which are not available, in general) would be expected to ensure that not the second (erroneous) segmentation is preferred. A pragmatic solution is to include additional synonymous lexeme variants. In our example, that means that the sense #kidney is not only represented by nephr$^{[en]ST}$ but also by nephro$^{[en]ST}$ (as well as by nefr$^{[sp;pt]ST}$ and nefro$^{[sp;pt]ST}$).

The delineation of subwords and semantic classes (MIDs) is a task that requires considerable knowledge of the domain terminology and can therefore not be fully automated [4]. For the lexica and thesaurus construction we have invested (Nov/2008) seven years of work, initially focusing on English and German and then adding Portuguese, Spanish, Swedish, and French. In this process we capitalized on the fact that there were substantial similarities between

[a] The MorphoSaurus system is commercially available as "Averbis Search Platform", cf. www.averbis.de

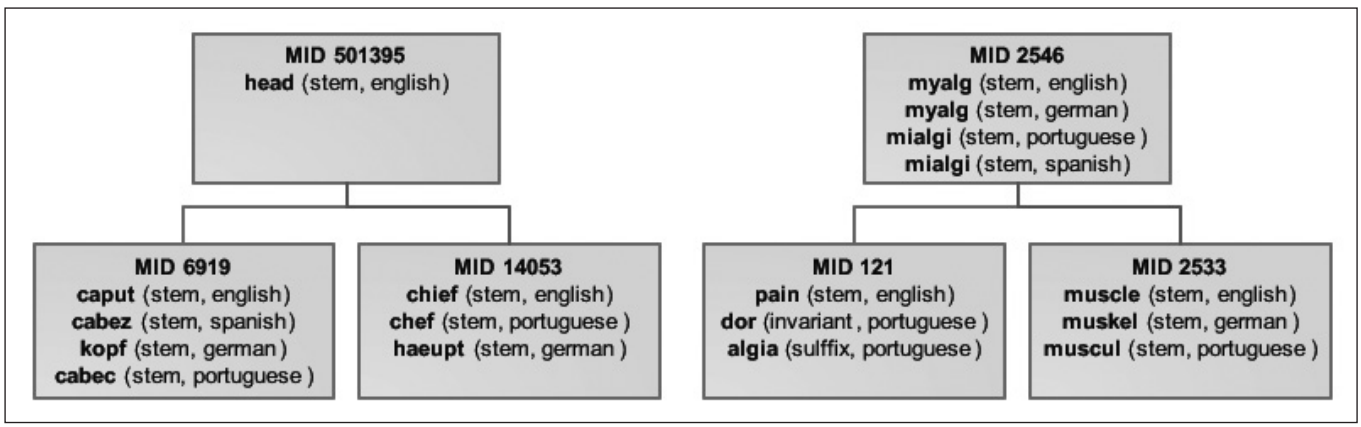

Fig. 1 Semantic relations supported by the MorphoSaurus thesaurus

medical terms in different languages, so-called cognates. We developed a semi-automated approach to acquire lexical entries of a new language thus optimizing the lexicon acquisition process [4]. This procedure is based on manually acquired string substitution rules (▶Table 1). Applying these rules, many translation candidates were generated that were then checked against lists of words in the target language for validation. The subsequent exploration of parallel corpora further enriched the raw vocabulary in the target language.

Table 1 Some string substitution rules and examples

44 Rules:	Portuguese	Spanish	7 Rules:	English	Swedish
ss → s	fracass	fracas	c → k	cramp	kramp
lh → j	mulher	mujer	ph → f	phosphor	Fosfor
+ça → za	cabeça	cabeza	ce → s	iceland	island

2.4 The MorphoSaurus Architecture

The architecture of the MorphoSaurus system is depicted in ▶Figure 2. First, the orthographic normalization step removes insignificant words and character substitutions are applied (e.g., elimination of capitalization, accents). A morphosyntatic parser then splits each remaining word into subwords. A high-performance extraction of subwords from large amounts of text is best achieved by the use of finite-state techniques for lexicon-based decomposition, derivation and deflection such as described in [4]. In the semantic normalization step these units are finally linked to a language-independent MID representation.

Queries are processed in an analogous way, thus allowing multilingual search. ▶Table 2 shows the main stages of this process, with English, German and Portuguese examples.

Lexical ambiguity is treated after the normalization process. For example, in one context the Portuguese word *lobo*[PT] can denote an animal (*wolf*), in another case a part of human brain (*lobe*). MorphoSaurus uses a lexical sense disambiguation routine, trained by multilingual corpora. It has been proven that disambiguation substantially increases the overall performance of the system [12].

3. Applications

We present two application scenarios in which the MorphoSaurus technology has been employed. Whereas the first one focuses on the multi-language retrieval in published literature, the second one demonstrates how clinical users can benefit from the semantic retrieval approach for performing cross-patient searches in the electronic health record.

3.1 Optimizing and Evaluating a Medical Search Engine

The German National Library of Medicine (ZB MED), Europe's largest medical library, maintains numerous bibliographic databases, which are accessible online via the MEDPILOT-Portal[b], a joint development between ZB MED[c] and German Institute of Medical Documentation and Information (DIMDI)[d]. With a single request, users of MEDPILOT can simultaneously research a wide range of medical sources and databases by a meta-search system.

Using the MorphoSaurus indexing strategy we are providing a semantic intra- and cross-language retrieval system to MEDPILOT. Three different bibliographic resources have been made searchable with the new technology: the 2007 version of Medline, Current Contents Medicine (CCMED), and the online public access catalog (OPAC) of the German National Library of Medicine, totaling 15.7 million bibliographic items and covering all major European languages.

A major objective is to analyze the searchers' needs in order to customize the new search engine, and to provide a gold standard for enabling a benchmark for measuring MEDPILOT's performance before and after the integration of the new search technology.

[b] http://www.medpilot.de
[c] German National Library of Medicine (ZB MED): http://www.zbmed.de
[d] German Institute of Medical Documentation and Information (DIMDI): http://www.dimdi.de

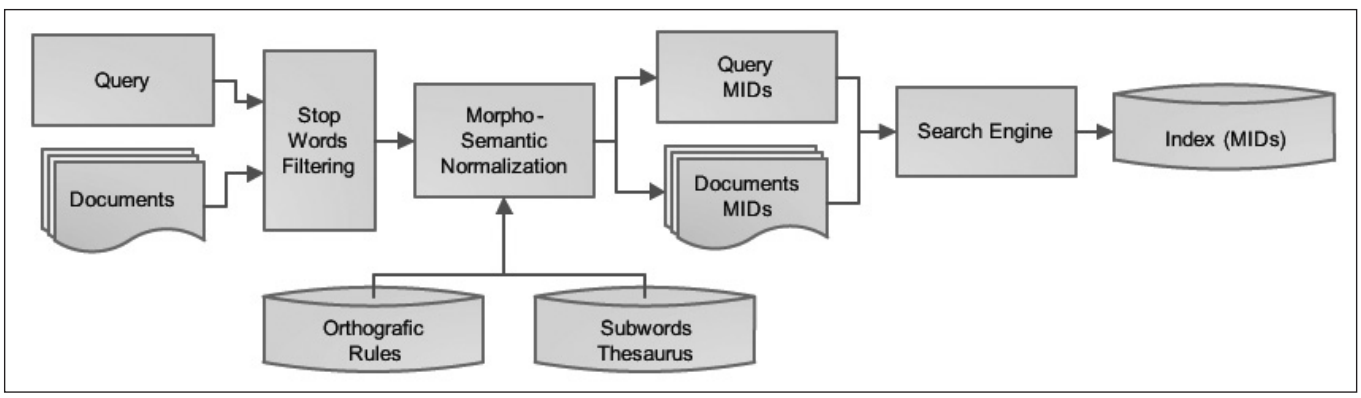

Fig. 2 The MorphoSaurus architecture

Table 2 Morpho-semantic indexing pipeline

	Orthographic Normalization *Normalization Rules*		Morphosyntactic Parser *Subword Lexicon*		Semantic Normalization *Subword Thesaurus*
High TSH values suggest the diagnosis of primary hyperthyroidism	→	high tsh values suggest the diagnosis of primary hyperthyroidism	→	high tsh values suggest the diagnos is of primar y hyper thyroid ism	→	#up tsh #value #suggest #diagnost #first #small #thyre
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose	→	erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose	→	er hoeh te tsh wert e erlaub en die diagnose e einer primae ren hypo thyre ose	→	#up tsh #value #permit #diagnost #first #small #thyre
A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário	→	a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario	→	a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primary o	→	#current tsh #value #up tsh #suggest #diagnost #small #thyre #first

The evaluation process comprised three stages:

- a content analysis of user queries to explore the user's information needs and to discover the scope of special contents 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 [13];
- optimizing the usability of the search engine user interface. Usability plays an important role for satisfaction and users trust on a website [14, 15].

For the content analysis we analyzed the logs of MEDPILOT queries in detail such as done in [16]. To this end we extracted queries covering seven months (142,922 queries) and draw a random sample of 10,000 queries. Afterwards, we developed a category system with 24 classes which was constructed by interplay of deductive and inductive procedures [17]. On the one hand, it is based upon what is known about medical information seeking [18], and on the other the content of the material was analyzed. A validation of the category-system was done by two raters who evaluated 150 of the queries. For this task, an intercoder reliability of 88% was achieved which can be seen as a sufficient correlation [19]. Then, each of the 10,000 queries was assigned to one or more categories by a domain expert.

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

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

The evaluation was performed by domain experts familiar with both medical and biological terminology. Due to the fact that an exact recall calculation is impossible for huge document sets as Medline, we had to choose a pragmatic way to evaluate our data. For this purpose, two indicators of the performance of information retrieval systems have been 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 special linguistic aspects we examined, such as, amongst others, misspellings, acronyms/abbreviations, synonyms, single word compounds, translations, layman/expert queries.

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

The MEDPILOT evaluation produced the following preliminary results. About a third (35.9%) of the search queries in MEDPILOT consists of one single word. Other 30% of the 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. Summarized, one- and two-word queries are the most frequent strategies for searching medical content (nearly 2/3 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 5.4% of the queries. Misspellings have been discovered in 4.6% of the queries.

Analyzing the type of medical content users of MEDPILOT are searching for, the top categories are "diseases, syndromes, symptoms" (30.9%), "methods of treatment, therapy and diagnostics" (28.5%), and content about "social medicine, statistics, studies, and epidemiology" (15.6%). One reason for carrying out a content analysis was to investigate to what extent the queries contained classical chemistry concepts. When looking at the results, we noticed that only 2.1% of the queries were classical chemistry terms. The other information needs in category 4 affects biochemistry, drugs, pharmaceutics and brands.

Preliminary results indicate that the MorphoSaurus technology 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 system. In 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.

3.2 Biomedical Data and Text Mining in an EHR System

The MorphoSaurus system was used in a second scenario to search within a database which stores the electronic health records (EHR) of the dermatology department at the University Medical Center in Freiburg, Germany. Information and document retrieval occupies a significant part of physicians' work [20], which are often dissatisfied with software tools that should – in theory – give them fast access to both patient-centered and scientific information. Meta-search solutions where users can easily browse through different data sources using a single intuitive interface ("Google for doctors") are a high demand. Traditional hospital information systems (HIS) usually focus on patient-centered, "intra-patient" issues, like storage of laboratory results, medical reports, discharge letters, coded diagnoses and procedures, and other clinical documents. The EHR, "a computer-stored collection of health information about one person linked by a person identifier" [21–22] should also fulfill the requirement of being "a secure, real-time, point-of-care, patient-centric information resource for clinicians", according to HIMSS [23]. The EHR also supports the collection of data for uses other than direct clinical care, such as billing, quality management, outcomes reporting, resource planning, and public health disease surveillance and reporting. According to those definitions, access to the medical information contained in current hospital information systems is mostly horizontal, i.e. patient-centered: by opening a patient's electronic record the doctor gets all information about this patient. The HIMSS definition suggests additional scenarios that aggregate information in a vertical view across large numbers of EHRs. As long as the scope is limited to structured entries like billing information, coded diagnoses and procedures, structured laboratory or microbiology results, it easily can be selected using appropriate and well-known database and data warehouse technologies. However unstructured narrative sources like discharge letters and finding reports, are of the same relevance for patient care: the more information the HIS stores, the more interesting are its vertical, i.e. interpatient interdependencies.

Some questions the doctor could pose:

- "What patients did I treat that had the same disease?"
- "What were the outcomes/adverse effects of that treatment?"
- "Did I have patients with disease X and symptoms Y?"
- "What was the name of the patient with symptoms Y I treated three weeks ago?"

Although several promising technologies like the Clinical Document Architecture [24, 25] and medical terminologies have been developed in order to standardize and structure clinical information, there is still a gap between this clinical need and today's practice.

In our study we extracted almost 30,000 clinical documents from the HIS database of the dermatology department system at the University Medical Center in Freiburg, Germany. These were mainly discharge letters, but also surgical reports, immunodermatological findings and different other kinds of finding reports. Using our semantic retrieval approach, all data were made available to the clinicians in a Google-like search interface. In order to evaluate its clinical impact, an evaluation scenario was set up. Users were intermittently encouraged to assess the benefit they perceived of the system and rate its potential for their work. The following topics have been examined in the evaluation:

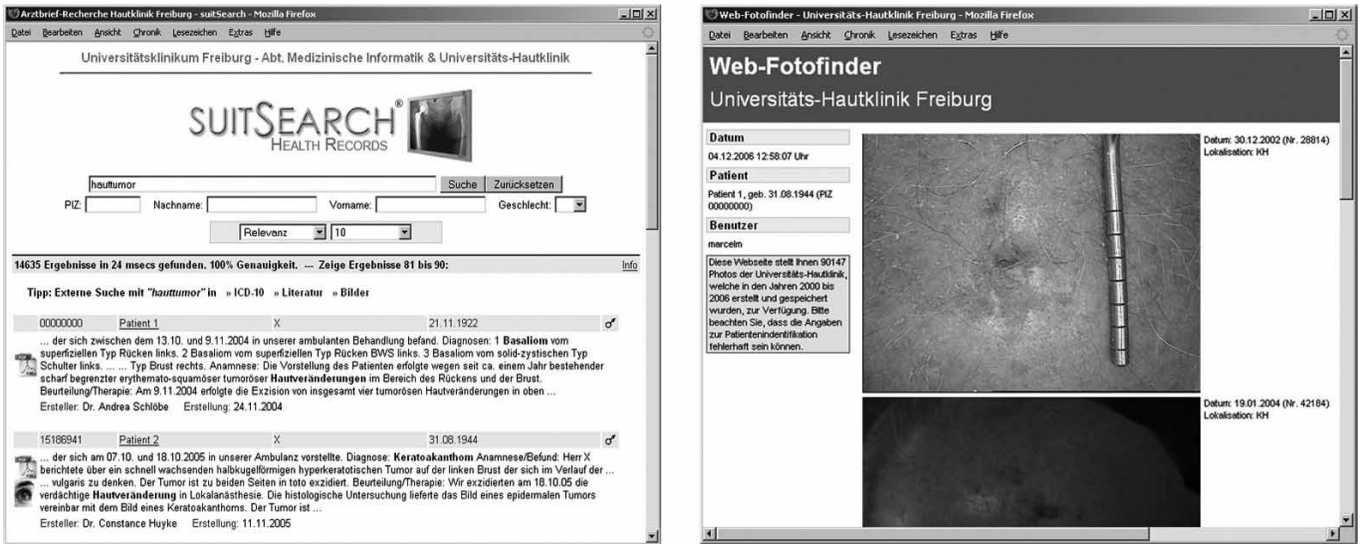

Fig. 3 Web-based user interface and image retrieval

- How shall text mining be presented to the physicians?
- What features are important for clinical routine?
- What additional data can enhance doctors' compliance to this system?
- How useful is text mining for clinical routine as well as for scientific or educational purposes?

Since patient photographic images play an important role in dermatological record keeping and training [26], a large image repository, consisting of 90,000 photographs, had been patient-specifically linked to the web interface, in order to enhance the overall benefit of the system.

The main design principles for the web-based application interface were combining ease of use with specific search options according to clinical needs (cf. ▶Fig. 3).

An edit box allows free-text search in all documents. Results can be ranked by relevance, date, patient name and date. In the results section, the relevant part of the underlying document is shown depending on the search terms. Search words or their derivations are highlighted. Information about the patient, the author and creation date of the document is shown in the respective context. The original document can be viewed as a PDF file that is linked to the record.

Another link opens the image library of the respective patient (▶Fig. 3), where all images are shown together with localization and creation date. This is particularly helpful for dermatologists, so they can analyze the images in the context of the underlying documents. Further linking, e.g. to laboratory results, are currently due.

Our evaluation showed that users' acceptance was remarkably high and enthusiastic from the very beginning: For the first time, they had access to thousands of documents and images independently of single patient-centered view. Time savings were attested for the whole clinical documentation process. Furthermore the system facilitated the targeted retrieval of case reports for clinical/epidemiological studies.

In our questionnaires, 82% of the users stated that the system could enhance their clinical performance. Almost 89% thought that this kind of biomedical data mining and the integration of narrative text with dermatological images have a very positive impact on their scientific work. The impact on dermatologic education has been less estimated: 52% of the users saw a potential benefit. Users generally wished the integration of more clinical data, like radiology images and reports, laboratory results and other clinical findings. The integration of standardized vocabularies (e.g. SNOMED, MeSH) would dramatically enhance the clinical value. Patient data could be cross-linked to external information sources as Medline or the Cochrane Database [27]. This was not the focus of the presented work but will be a major cornerstone in the future.

3.3 Facilitating Access to Consumer Health Information

The "Weisse Liste" (white list) platform [28] is a joint effort of the German Bertelsmann Foundation and the umbrella association of major patient and consumer organizations. It is centered around a free, non-commercial Internet portal that supports patients, their families and patient counselors in their search for a suitable health care provider. In its first phase the portal offers information about the services and quality of around 2000 hospitals in Germany. In the future other health care providers such as inpatient care facilities will be integrated.

The portal allows the user to formulate queries according to their prior knowledge and their own wishes. The centerpiece – an interactive search assistant for specific illnesses – guides users step-by-step to a customized result. It provides them with easy-to-understand explanations so they can interpret the sometimes complex information on the structural and performance-related qualities of the health care institutions.

The information about the hospitals is obtained by "Structured Quality Reports" (SQB), a compulsory report following nationwide uniform specification for German hospitals. These specifications are to ensure that all hospitals will regularly publish comparable details on quality-relevant aspects of their services.

The SQB has two parts – a basic section containing service and organization data of the hospital concerned, and a system section which shall give a systematic overview of the active involvement in quality management processes. Hospital quality indicators include diagnoses and procedure statistics using the German Modification of the International Classification of Diseases (ICD-10-GM) and the German procedure classification OPS [29].

One of the main challenges of the consumer portal described is to give medical laypersons user-friendly access to the SQB data. To this end, the "Weisse Liste" uses an adaptation of the MorphoSaurus system, combined with a large synonym and hyponym list including about 200,000 expert and laypersons' terms. It enables patients and the interested public to use their own words for querying 4000 ICD-10-GM and OPS codes.

The portal is independent of the interests of health insurance companies and service providers. Its initiators want to use the portal as a way to provide patients and their families with assistance in making decisions and to support the work of patient counselors and physicians. The easy-to-understand and well-organized presentation of the information about quality, which previously was often only available to industry professionals, is intended to enable patients to conduct a dialog with patient counselors and physicians and actively participate in selecting health care providers.

4. Conclusions

In this paper, we presented an interlingua-based retrieval system called MorphoSaurus, that takes into account the peculiarities of various Germanic languages and allows both intra- and cross-language search in biomedical data. Three real-world applications based on this system were described.

The MorphoSaurus system is based on a thesaurus using subwords as lexical units and a semantic layer in which subwords are mapped to language-independent semantic identifiers. The usefulness of this approach that had previously been shown in different scenarios such as terminology

mapping, intralingua and cross-language document retrieval could now be demonstrated in large-scale settings and assessed by real users. The library use case showed that a language-independent, semantic document retrieval is useful for exploring a heterogeneous collection of bibliographic items in the biomedical field. The EHR experience has shown that horizontal retrieval across many cases addresses the user's needs and opens the way to new data analysis scenarios. Finally the health portal use case demonstrated that semantic indexing approach facilitates the access to health service quality information using medical terminology by laypersons using their own expressions.

Acknowledgments

This work has been supported by the CNPq, Brazil, (550830/05-7) and the International Bureau of the BMBF, Germany (BRA 05/022).

References

1. Scott RE. e-Records in health – Preserving our future. Int J Med Inform 2007; 76 (5–6): 427–431.
2. Lobach DF, Detmer DE. Research Challenges for Electronic Health Records. American Journal of Preventive Medicine 2007; 32 (5, Suppl 1): S104-S11.
3. Wen H-C, Ho Y-S, Jian W-S, Li H-C, Hsu Y-HE. Scientific production of electronic health record research, 1991–2005. Computer Methods and Programs in Biomedicine 2007; 86 (2): 191–196.
4. Markó K, Schulz S, Medelyan O, Hahn U. Bootstrapping Dictionaries for Cross-Language Information Retrieval. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brasil, 2005. pp 528–535.
5. Honeck M, Hahn U, Klar R, Schulz S. Text Retrieval Based on Medical Subwords. In: Surjan G, Engelbrecht R, McNair P (eds). Health Data in the Information Society. Proceedings of the MIE2002. Amsterdam: IOS Press; 2002. pp 241–245.
6. Oard DW, He D, Wang J. User-assisted query translation for interactive cross-language information retrieval. Information Processing & Management 2008; 44 (1): 181–211.
7. Gey FC, Kando N, Peters C. Cross-Language Information Retrieval: the way ahead. Information Processing & Management 2005; 41 (3): 415–431.
8. Kishida K. Technical issues of cross-language information retrieval: a review. Pergamon Press, Inc. 2005. pp 433–455.
9. Franz M, McCarley JS, Ward RT. Ad hoc, cross-language and spoken document information retrieval at IBM. TREC-8. Gaithersburg, MD: National Institute of Standards and Technology. 2000.
10. Markó K, Hahn U, Schulz S, Daumke P, Nohama P. Interlingual Indexing across Different Languages. 7th International Conference "Recherche d'Information Assistée par Ordinateur" (RIAO'04); 2004. Avignon, France: Proceedings of RIAO'04; 2004. pp 82–99.
11. K. Markó, P. Daumke, S. Schulz, R. Klar, U. Hahn. Large-Scale Evaluation of a Medical Cross-Language Information Retrieval System. Proceedings of the 12th World Congress on Medical Informatics, Brisbane, Australia (Studies in Health Technology and Informatics, Vol. 129). 2007. pp 392–396.
12. Markó K, Schulz S, Hahn U. Unsupervised Multilingual Word Sense Disambiguation via an Interlingua. Proceedings of the 20th National Conference on Artificial Intelligence (AAAI '05), Pittsburgh, Pennsylvania. 2005. pp 1075–1080.
13. 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. Cologne: Herbert von Halem Verlag; 2007. pp 243–258.
14. 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.
15. Dzeyk W. Vertrauen in Internetangebote. Saarbrücken: VDM Verlag; 2007.
16. 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.
17. Rustemeyer R. Praktisch-methodische Schritte der Inhaltsanalyse. Eine Einführung am Beispiel der Analyse von Interviewtexten. Münster: Aschendorff; 1992.
18. Davies K. The information-seeking behaviour of doctors: a review of evidence. Health Information and Libraries Journal 2007; 24: 78–94.
19. Wirtz M, Caspar F. Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Göttingen: Hogrefe; 2002.
20. Koch O, Kaltenborn R. Mehr Zeit für Patienten durch bessere Information. Deutsches Ärzteblatt 2005: 28–29: A2008–2011.
21. Waegemann CP. The vision of electronic health records. J Med Pract Manage 2002: 18 (2): 63–65.
22. Waegemann CP. The five levels of electronic health records. MD Comput 1996: 13 (3): 199–203.
23. HIMSS Electronic Health Record, Definitional Model. http://www.himss.org/content/files/EHRAttributes.pdf. Last accessed: January 2009.
24. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo SA. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc 2006: 13(1): 30–39.
25. Müller ML, Uckert F, Bürkle T, Prokosch HU. Cross-institutional data exchange using the clinical document architecture (CDA). Int J Med Inform 2005: 74 (2–4): 245–256.
26. Scheinfeld N. Photographic images, digital imaging, dermatology, and the law. Arch Dermatol 2004: 140 (4): 473–476.
27. The Cochrane library. http://www.cochrane.org.
28. Weisse Liste. http://weisse-liste.de.
29. Klassifikationen im Gesundheitswesen. Deutsches Institut für Medizinische Dokumentation und Information (DIMDI). http://www.dimdi.de/static/de/klassi/index.htm.