eHealth2015 – Health Informatics Meets eHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-524-1-9

Semantic Retrieval and Navigation in Clinical Document Collections

Markus KREUZTHALER[a,1], Philipp DAUMKE[b] and Stefan SCHULZ[a]
[a]*Institute for Medical Informatics, Statistics and Documentation,*
Medical University of Graz, Austria
[b]*Averbis GmbH, Freiburg, Germany*

Abstract. Patients with chronic diseases undergo numerous in- and outpatient treatment periods, and therefore many documents accumulate in their electronic records. We report on an on-going project focussing on the semantic enrichment of medical texts, in order to support recall-oriented navigation across a patient's complete documentation. A document pool of 1,696 de-identified discharge summaries was used for prototyping. A natural language processing toolset for document annotation (based on the text-mining framework UIMA) and indexing (Solr) was used to support a browser-based platform for document import, search and navigation. The integrated search engine combines free text and concept-based querying, supported by dynamically generated facets (diagnoses, procedures, medications, lab values, and body parts). The prototype demonstrates the feasibility of semantic document enrichment within document collections of a single patient. Originally conceived as an add-on for the clinical workplace, this technology could also be adapted to support personalised health record platforms, as well as cross-patient search for cohort building and other secondary use scenarios.

Keywords. Information Storage and Retrieval; Health Records, Personal; Data Mining

1. Introduction

For patients with chronic diseases, large quantities of documents accumulate in their electronic health record (EHR). An important document type produced in hospitals is the discharge summary [1]. For patients with numerous treatment episodes, the sum of their discharge summaries is an essential source of information about their current and past health problems and the progression of chronic diseases, encompassing signs, symptoms, allergies, diagnoses, medications, symptoms, and procedures, embedded in contexts such as time or diagnostic certainty. Whereas discharge summaries, such as findings reports or progress notes are primarily produced by clinicians to be read by clinicians, the potential of *secondary use* scenarios is increasingly valued. Here the use of natural language processing (NLP) technology plays a crucial role [2-6].

In this paper, we demonstrate how advanced NLP and innovative interfaces can also improve *primary use* scenarios, for example the presentation of EHR content to clinicians. The main rationale for this is the fact that the amount of clinical texts about one patient may conflict with the limited time clinicians have to carefully read these documents, which bears the risk that important passages are read over and wrong

[1] Corresponding Author: Markus Kreuzthaler, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria, E-Mail: markus.kreuzthaler@medunigraz.at

decisions are made. With an ageing population, the number of documents relating to a particular patient tends to increase, and future personal record systems like ELGA [7] will bring together clinical documents from different platforms. Clinical users familiar with search and navigation features offered by current search engines for the Web or for specialized databases (e.g. PubMed), will increasingly request these functionalities in their clinical workplace.

Textual content in medical records exhibits many peculiarities that distinguish it from the language used in medical publications, as the following text snippet demonstrates:

```
"Ca. 2 x 1 cm, große ovaläre Verschattung im UF li. Rez. re. lat. frei, li.teilhärent"
```

("An approx. 2 by 1 cm sized, oval-shaped shadow (in the chest X-ray) over(-laying) the left lower lobe; pleural recess (on the right side of the lobe) free of fluids, on the left side partly adhesive (after non-recent inflammation)").

As much as this kind of highly condensed text is understandable for physicians, it is challenging for computer-based morphological and syntactic processing, as well as for semantic interpretation and annotation. Typical language idiosyncrasies have to be considered for setting up an advanced natural language processing (NLP) pipeline for clinical narratives: ambiguous terms, acronyms, abbreviations, single-word compounds, derivations, spelling variants, uncorrected spelling, typing and punctuation errors, jargon expressions, a telegram style, non-standardized numeric expressions, and non-standardized variations of negations all exist.

Our system aims at handling the special requirements needed for clinical text processing to support browser-based search and navigation within multiple documents. This use case differs significantly from many other search scenarios. As the size of the document collection under scrutiny is relatively small – pertaining to only one, pre-selected patient – compared to, for example, the complete clinical document pool in a hospital or the collection of abstracts in a literature database, a retrieval approach geared towards high recall seems acceptable. That (mostly) all the facts that address given information needs are actually retrieved is valued by the user, who then would also accept a certain amount of false-positive results.

The remainder of this document is organized as follows: Section 2 describes the components used for enhanced semantic enrichment of clinical texts. Section 3 presents the obtained results for meta-data annotations and describes the web-based search interface. In the last two sections we discuss pitfalls and challenges that must be addressed by any general EHR navigation system, independent of the clinical domain.

2. Methods

2.1. Data

In order to harvest a sample of discharge summaries for training and testing, we performed a search across all documents from the Department of Dermatology of the Graz University Hospital, limited to a six-year period, filtered by ICD-10 codes describing malignant melanoma. The extraction was done using an **ETL** (Extract Transform Load) workflow with Talend Open Studio [8], which is scalable and has a huge resource of supported database connectors.

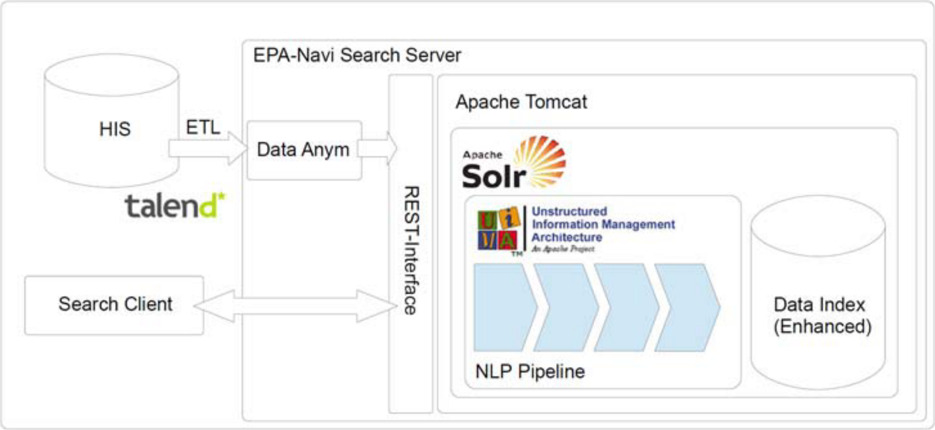


Figure 1. Main components and data flow.

Both data extraction and de-identification were mandated by the data owner and conducted by our Scientific Service Area – Medical Data Management group, with the sole purpose of producing a non-identifiable medical document collection. The anonymised data (**Data Anym**) set comprises 1,696 summaries from 175 patients (max 82; min 1; avg. ~10 documents per patient). These data were converted into XML following the technical requirements of the Apache Solr search server [9].

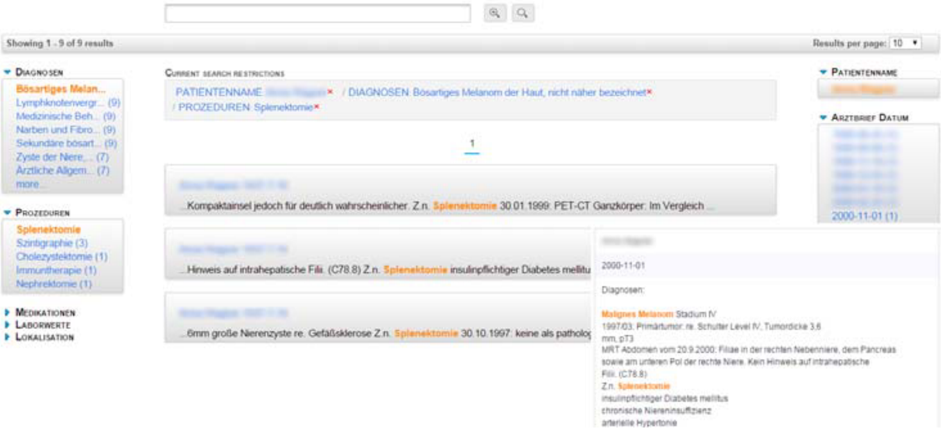
2.2. Architecture

Figure 1 describes the main components of the framework and its technical setting.

Several **NLP** components are aggregated into a pipeline and customised using different **UIMA** [10-14] (Apache Unstructured Information Management Architecture) based analysis engines (AEs) where the required NLP functionality is encapsulated. The following components were used by the pipeline: Sentence detection and tokenization (OpenNLP [15]); Stemming (Lucene Snowball [16]); Decompounding based on morphosemantic analysis [17] (Myokarditis = #muscle #heart #inflamm; Herzmuskelentzündung = #heart #muscle #inflamm; Inflammation of the heart muscle = #inflamm #heart #muscle); Drug annotation (dosage, regimen); Lab values (numbers, units); Negation detection (based on Negex [18]); Date and time recognition; and Abbreviation resolver.

In addition, we used the following terminologies to semantically annotate free text passages via a concept mapper. Diseases: ICD-10 DIMDI; Procedures: OPS DIMDI; Medication: ABDAMED; Lab values: Averbis Lab Terminology; and Body parts: RadLex 2.0. We chose to use these categories as they correspond, according to [19, 20], to the top five semantic searches in a typical clinical environment.

After the annotation process using UIMA, the enriched documents are indexed by Apache **Solr.** The annotation and indexing service is accessed via a **REST-Interface** in a one-time batch job. The enhanced document index is locally saved on a server using **Apache Tomcat** as a deployment container. Finally, the documents are searched via a Web-based **Search Client**. The result of the search interface is presented in more detail in the next section.



PATIENTENNAME [redacted] / DIAGNOSEN: Bösartiges Melanom der Haut, nicht näher bezeichnet / PROZEDUREN: Splenektomie

[redacted]

...Kompaktainsel jedoch für deutlich wahrscheinlicher. Z.n. Splenektomie 30.01.1999: PET-CT Ganzkörper: Im Vergleich ...

[redacted]

...Hinweis auf intrahepatische Fili. (C78.8) Z.n. Splenektomie insulinpflichtiger Diabetes mellitu

[redacted]

...6mm große Nierenzyste re. Gefäßsklerose Z.n. Splenektomie 30.10.1997: keine als patholo

[redacted]

2000-11-01

Diagnosen:

Malignes Melanom Stadium IV
1997/03: Primärtumor: re. Schulter Level IV, Tumordicke 3.6
mm, pT3
MRT Abdomen vom 20.9.2000: Filiae in der rechten Nebenniere, dem Pancreas
sowie am unteren Pol der rechte Niere. Kein Hinweis auf intrahepatische
Fili. (C78.8)
Z.n. Splenektomie
insulinpflichtiger Diabetes mellitus
chronische Niereninsuffizienz
arterielle Hypertonie

Figure 2. Main search interface.

3. Results

All discharge summaries related to a patient are presented within a browser (Figure 2). The hybrid semantic and full text search engine follows the look and feel of Web search interfaces. To query clinical data, free text- and concept-based queries implemented as facets (diagnoses, procedures, medications, lab values, body parts) can be compound, therefore coupling both search strategies in a user friendly way. In contrast to these so called *content-based facets* located at the left side of the screen, *metadata facets* exploiting information about the document are placed on the opposite side. They contain, for example, the document date and the patient identification. Concept-based facetted search (AND, OR and NOT operators can be combined) supports a recall oriented navigation, taking into account synonyms and hyponyms, for example. Free text search, using the search bar at the top, exploits the Lucene query syntax and has the advantage that it is independent of predefined terminologies. It is well known that no terminology has 100% coverage, given the wealth and dynamic growth of medical language. For each hit the relevant document snippets are displayed, from which the full document can be inspected via in-depth navigation. Here, as well as in the results section, words or phrases that semantically correspond to the search terms are highlighted and colour-coded (Figure 2).

4. Discussion

In the presented framework, semantic retrieval within the EHR is mainly provided as a support for non-linear navigation, ideally drawing the user's attention to those passages in the document space that best addresses their information needs. The combination of classical search with facets shrinks down the information space within a patient's list of documents. Visual feedback of conceptually related terms in the document is of the utmost importance in order to support quick relevance statements for the user. Nevertheless, some pitfalls and challenges need to be considered, such as:

Redundant information. The functionality is still hampered by a large amount of redundant information in the document corpus. We have not addressed this so far, as our system leaves the documents intact. Redundancy mostly addresses the past history and family history section, and is justified by the fact that each document, in isolation, needs to convey a complete picture of the patient. For instance, an information such as "mother died from breast cancer at the age of 33" would be an important item in the family history section of each single document, which, of course, would appear as redundant in a synoptic view over the complete collection of discharge summaries.

Non-standardised headers. Section headers are important for contextualizing the information in each section, e.g. to distinguish past history from family history. In our document collection we observed 59 different header types. Although the authoring system provides pre-defined headers, these can be overwritten by the user according to their needs and preferences. For example the heading "DIAGNOSE" occurred 1,169 times in contrast to 525 "DIAGNOSEN". In 45 cases, headers were completely missing.

NLP and concept mapping. Our experiences demonstrate typical challenges and pitfalls in natural language processing: Ambiguity: "Ca" (Calcium) vs. "ca." (approx.) and "Ca."(Carcinoma); Missing synonyms, such as in abbreviations that characterize the jargon of a specific sub-discipline, e.g. "SSM" (superficially spreading melanoma"); Coordination problems ("left and right leg", which should be mapped to the concepts *left leg* and *right leg*); Local context (liver cancer and renal disease, should map to *liver cancer* and *renal disease* and not to *liver disease* and *renal cancer*).

Terminologies. The terminologies we chose for diseases and procedures are classifications (ICD and OPS), and therefore not ideally suited for semantic annotations (residual classes like “not elsewhere classified” or “not otherwise specified”, lack of multiple hierarchical links). SNOMED CT would certainly be a better candidate for semantic annotation, but it is currently not available in German.

Time. Medical documents usually contain numerous date references. Relating them correctly to patient-based events could feed an additional time-related index. This could address information needs that exploit temporal sequences, e.g. between medications and disorders or symptoms.

5. Conclusion and Outlook

We reported on the semantic enrichment of medical texts, in order to support recall-oriented navigation across a single patient's complete documentation. To this end, we built a browser-based platform including a natural language processing toolset for semantic document annotation and indexing. It supports free text and concept-based querying, enhanced by dynamically generated facets. We could demonstrate the feasibility of this approach using a corpus of 1,696 documents belonging to 175 patients. The platform is functional and will be tested with pilot users soon. Future work will consider how to semi-automatically expand a given terminology with the document type’s medical sublanguage and to highlight qualitative aspects of the information retrieval approach in accordance with well-known scientific challenges in this area. Another desideratum is the support of temporal aspects. In addition, human search and navigation behaviour and their information needs regarding the optimised use of medical documents in the clinical workflow have to be studied in a controlled setting, in order to improve the user interface and its instruments [21].

References

[1] Christensen, T., & Grimsmo, A. (2008). Instant availability of patient records, but diminished availability of patient information: a multi-method study of GP's use of electronic patient records. BMC Medical Informatics and Decision Making, 8(1), 12.

[2] Friedman, C., & Elhadad, N. (2014). Natural language processing in health care and biomedicine. In Biomedical Informatics (pp. 255-284). Springer London.

[3] Meystre, S. M., et al. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 35, 128-44.

[4] Hripcsak, G., & Albers, D. J. (2012). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, amiajnl-2012.

[5] Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405.

[6] Chapman, W. W. et al. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. JAMIA 18(5), 540-543.

[7] ELGA, http://www.elga.gv.at/ , last access: 30.01.2015.

[8] talend, http://www.talend.com/products/talend-open-studio, last access: 30.01.2015

[9] Apache Solr, http://lucene.apache.org/solr/, lasst access: 30.01.2015

[10] Apache UIMA, https://uima.apache.org/, last access: 30.01.2015

[11] Savova, G. K. et al. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*,*17*(5), 507-513.

[12] Coden, A. et al. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of Biomedical Informatics*, *42*(5), 937-949.

[13] MedKAT, http://ohnlp.sourceforge.net/MedKATp/#d4e5, last access: 30.01.2015

[14] Savova, G., Kipper-Schuler, K., Buntrock, J., & Chute, C. (2008). UIMA-based clinical information extraction system. Towards enhanced interoperability for large HLT systems: UIMA for NLP, 39.

[15] Apache openNLP, https://opennlp.apache.org/, last access: 30.01.2015

[16] Apache Lucene, http://lucene.apache.org/core/, last access: 30.01.2015

[17] Markó, K. et al. (2005). MorphoSaurus - Design and Evaluation of an Interlingua-based, Cross-language Document Retrieval Engine for the Medical Domain. *Methods Inf Med*, *44*(4), 537-545.

[18] Chapman, W. Wet al. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301-310.

[19] Natarajan, K., Stein, D., Jain, S., & Elhadad, N. (2010). An analysis of clinical queries in an electronic health record search utility. *International Journal of Medical Informatics*, *79*, 515-522.

[20] Zheng, K., Mei, Q., & Hanauer, D. A. (2011). Collaborative search in electronic health records. *Journal of the American Medical Informatics Association*, *18*(3), 282-291.

[21] Hearst, M. (2009). Search user interfaces. Cambridge University Press.