Towards Integrated Information Extraction and Facetted Search Applications in Nephrology

Danilo Schmidt[1], Hans-Jürgen Profitlich[2], and Daniel Sonntag[2]

[1] Nephrology Department
Charité - Universitätsmedizin Berlin
10117 Berlin, Germany
[2] German Research Center for Artificial Intelligence (DFKI)
66123 Saarbrücken, Germany

Abstract. This work focusses on our first integration steps of complex and partly unstructured medical data into a clinical research database. Our main application is an integrated facetted search tool in nephrology based on automatic information extraction results from textual documents. We describe the details of our technical architecture which is based on open-source tools—to be replicated at other universities, research institutes, or hospitals.

1 Introduction

As medical records may cover a very long history of diseases (up to 30 years) and include a vast number of diagnoses, symptoms, results, medications, and laboratory values, we could highly benefit from advanced search capabilities in clinical information systems to allow for the retrieval of relevant data. However, medical information systems often suffer from good search capabilities for data which has many unstructured text parts. Therefore, concepts to implement knowledge based systems, based on textual information extraction in medicine, are in focus of many recent research initiatives [11].

In this paper, we propose a three stage process: (1) offline textual information extraction from medical records in transplant medicine; (2) the generation of interesting facetted search capabilities on the results of the previous stage; (3) the combination of the information extraction results with structured laboratory values (ongoing work). Such a facetted search application uses techniques for accessing information organised according to a facetted medical classification system, allowing users to explore a collection of diagnoses, symptoms, results, medications, and laboratory values by applying multiple filters. Thus, facetted search allows clinicians to analyse complex data sets along a medical and cognitive (reflective) chain of decision-making; in particular, facetted search applications allow physicians to identify groups of patients with similar attributes. This can provide valuable decision support, when physicians are confronted with situations where rare or complex diseases require a high degree of specialist knowledge to filter and interpret (unstructured) medical data.

2 Background and Related Work

The facetted search application is based on the nephrology database TBase®. The web-based electronic patient record TBase® has been implemented in a German kidney transplantation programme as a cooperation between the Nephrology of Charité Universitätsmedizin Berlin and the AI Lab of the Institute of Computer Sciences of the Humboldt University of Berlin [3,10]. Currently, TBase® automatically integrates essential laboratory data (9.9 million values), clinical pharmacology (237.000 prescribed medications), diagnostic findings from radiology, pathology and virology (146.000 findings), and administrative data from the SAP-system of the Charité (70.000 diagnoses, 25.000 hospitalisations). Two groups of use cases for the application of facetted search in the medical field, and nephrology, can be identified: first, the use in clinical research, and second, the implementation in the individual treatment as a decision support system in the clinical routine.

Sacco [7] describes an approach of a guided interactive diagnostic system based an dynamic taxonomies. Biron et al. [2] describe an information retrieval system for computerised patient records We extend these approaches by a special multi-facet functionality. Our approach shows the following main advantages:

- In our facetted search application, the user may remove *any* restriction he or she may have made in previous steps. This allows for a much better navigation through the search space where related systems only allow the subsequent thinning [8, 9].
- The ranking of facet values by cardinality supports the survey of remaining subsets.
- We base automatically generated facets (e.g., disease/symptom relationships and negations) on multi-term extraction and relation extraction, by employing state-of-the-art, high-precision textual information extraction modules.

Only recently, new text mining approaches on Web-based medical literature have been proposed. For extracting adverse drug events from text [6] or automatic symptom extraction from texts on rare diseases [4], for example. However, clinical information extraction from patient records is still underrepresented and underdeveloped in clinical settings. Earlier work includes evaluating context features for medical relation mining on medical abstracts; the identification of semantic relations, such as substance A treats disease B, remains a non-trivial task [13]. Recent work and comparative baseline experiments include temporal information extraction [5]. A special trend becomes apparent, the need for ontology modelling of medical terminology and corresponding information extraction results [12]. Because of enormous annotation costs, mainly unsupervised methods are being used [1]. In industry and in the context of reliable clinical relevance, however, very detailed (and labor-intensive) supervised rule-based approaches represent the state-of-the-art.[3]

[3] Here, we use our research project partner's solution (Averbis), which is based on shallow text parsing, see https://averbis.com/en/research/

3 System Architecture

The annotated texts are transferred in XMI format[4] and stored in a local database at DFKI (see figure 1). Important components are the Solr search platform, the information extraction module, and the facetted search and presentation user interface modules. Solr[5] is an open source enterprise search platform

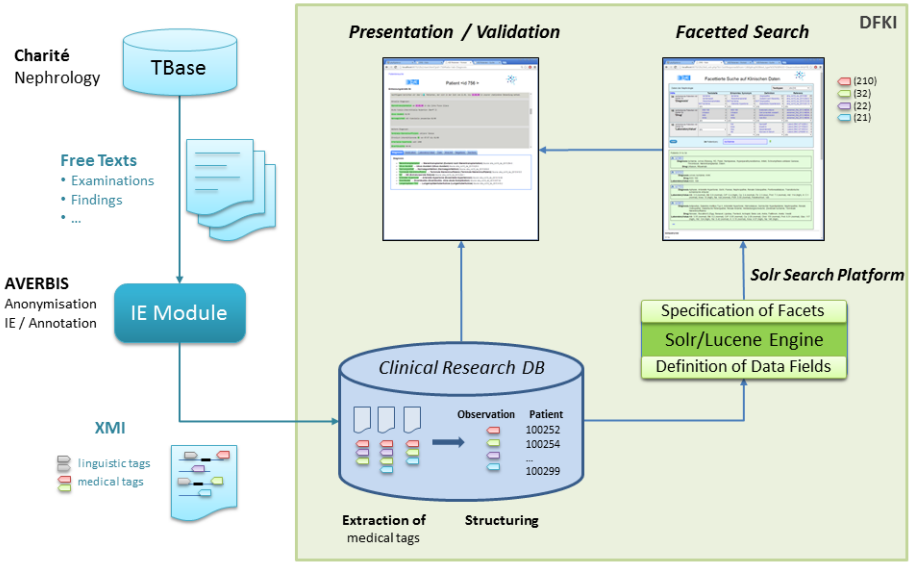

Fig. 1. System Architecture of the Integrated Information Extraction and Facetted Search Application

used in many large websites and applications and is one of the most popular enterprise search engines.[6] Solr runs as a standalone full-text search server and uses the Lucene Java search library at its core for full-text indexing and (facetted) search. We chose the Solr system mainly because of some interesting features like facetted navigation, a query language that supports structured and textual search, the possibility for automatic result clustering based on Carrot2[7], its scalability and extensibility through plug-ins, and its various APIs for input (text, xml, JSON, etc.) and output (JSON, XML, PHP, python, etc.).

The first step in our process is offline informative extraction. The text data for our system originate from the TBase® database of Charité Berlin containing medical information about nephrology patients. In the first phase we only

[4] http://www.omg.org/spec/XMI/
[5] http://lucene.apache.org/solr/
[6] http://db-engines.com/en/ranking/search+engine
[7] http://project.carrot2.org/

used about 5000 unstructured, free texts (no meta data or structured data of patients) of four types: 'Befunde" (findings), "Untersuchungen" (visits), "Entlassungsbriefe" (clinical reports), and "Verläufe" (progress reports). These free texts are processed by the project partner Averbis, which anonymises the texts and adds annotations based on several medical reference systems and dictionaries (LOINC[8], ICD10[9], ABDAMED[10]).

A software module extracts the relevant medical tags and features and stores these in a database structure similar to the i2b2 star structure (in order to simplify the updates of the target system i2b2[11]). The user interface to search and explore the annotated text database by using facets is built as a web service based on the Solr extension "solarium" for PHP systems.[12] This extension provides for an API to specify all parameters necessary to create complex Solr requests. The presentation/validation user interface of the system (see figure 2) consists of two parts: the upper part shows the original text with highlighted annotations, the lower part contains tabs listing the different relevant annotations. Clicking on an item in the lower part scrolls the text above to the corresponding

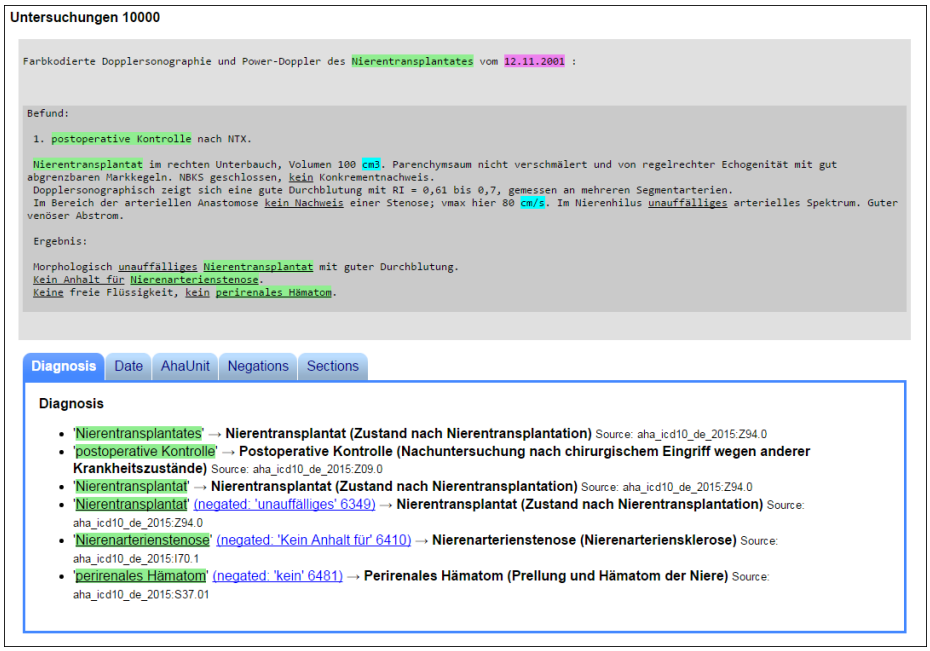

Fig. 2. Presentation/Validation User Interface

[8] https://loinc.org/
[9] http://www.icd-code.de/
[10] http://www.wuv-gmbh.de/abdata-pharma-daten-service/datenangebot/abdamed/
[11] https://www.i2b2.org/about/intro.html
[12] http://www.solarium-project.org/

position. The original XMI contents representing the complete original annotation information is shown in a pop-up window when a highlighted annotation in the text is clicked. Accordingly, this page serves two different purposes: (1) the presentation of the original text snipped found by the facetted search, and (2) the validation of the annotations.

4 Conclusion and Outlook

We demonstrated that new facetted search applications in the use case of transplant medicine in nephrology, based on open-source software tools and exchangeable information extraction modules, are feasible and a very suitable decision-support tool for the doctor: this type of a knowledge based system provides physicians with a practicable tool for the analysis of medical data and decision support for cohort selection. We developed a user interface for facetted search which is based on the Solr Engine. In the next project phase, we will extend the capabilities of the facetted search application, mainly including the following aspects: (1) integration of existing structural information about patients and treatments which includes numerical values, in relation to laboratory values or medications in particular; (2) extending the user interface by adding visual search and presentation techniques like "foamtree"[13] to further facilitate the users exploration of the search space; (3) the integration of facetted search into special use cases moving towards individualised medicine [11].

Acknowledgements

This research is part of the project "clinical data intelligence" (KDI) which is founded by the Federal Ministry for Economic Affairs and Energy (BMWi).

References

1. Alicante, A.: Unsupervised entity and relation extraction from clinical records in italian. Computers in Biology and Medicine 72(1), 263–275 (2016)
2. Biron, P., Metzger, M., Pezet, C., Sebban, C., Barthuet, E., Durand, T.: An Information Retrieval System for Computerized Patient Records in the Context of a Daily Hospital Practice: the Example of the Leon Berard Cancer Center (France). Applied Clinical Informatics 5(1), 191–205 (2014)
3. Lindemann, G.: A web-based patient record for hospitals - the design of tbase2. In: Bruch, H.P. (ed.) New Aspects of Hight Technology in Medicine: Hannover (Germany), pp. 409–414. Monduzzi Editore, International Proceedings Division (2000)
4. Métivier, J., Serrano, L., Charnois, T., Cuissart, B., Widlöcher, A.: Automatic symptom extraction from texts to enhance knowledge discovery on rare diseases. In: Artificial Intelligence in Medicine - 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015. Proceedings. pp. 249–254 (2015)

[13] https://carrotsearch.com/foamtree-overview

5. Mkrtchyan, T., Sonntag, D.: Deep parsing at the CLEF2014 IE task. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 138–146 (2014)
6. Odom, P., Bangera, V., Khot, T., Page, D., Natarajan, S.: Extracting adverse drug events from text using human advice. In: Artificial Intelligence in Medicine - 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015. Proceedings. pp. 195–204 (2015)
7. Sacco, G.: Guided interactive diagnostic systems. In: Computer-Based Medical Systems. pp. 117–122 (2005)
8. Sacco, G.: Dynamic taxonomies and guided searches. Journal of the American Society for Information Science and Technology 57(6), 792–796 (2006)
9. Sacco, G.: Dynamic taxonomies for intelligent information access. In: Khosrow-Pour, M. (ed.) Encyclopedia of Information Science and Technology, pp. 3883–3892. 3 edn. (2014)
10. Schröter, K.: Tbase2, a web-based electronic patient record. Fundamenta Informaticae 43(1-4), 343–353 (2000)
11. Sonntag, D., Tresp, V., Zillner, S., Cavallaro, A., Hammon, M., Reis, A., Fasching, A.P., Sedlmayr, M., Ganslandt, T., Prokosch, H.U., Budde, K., Schmidt, D., Hinrichs, C., Wittenberg, T., Daumke, P., Oppelt, G.P.: The clinical data intelligence project. Informatik-Spektrum Journal pp. 1–11 (2015)
12. Sonntag, D., Wennerberg, P., Buitelaar, P., Zillner, S.: Pillars of ontology treatment in the medical domain. J. Cases on Inf. Techn. 11(4), 47–73 (2009)
13. Vintar, S., Todorovski, L., Sonntag, D., Buitelaar, P.: Evaluating context features for medical relation mining. In: Proceedings of the ECML/PKDD Workshop on Data Mining and Text Mining for Bioinformatics (2003)