Extraction and Processing of Rich Semantics from Medical Texts

Kerstin Denecke[1], Yihan Deng[2], and Thierry Declerck[3]

[1] Department of Medical Informatics,
Bern University of Applied Sciences
Höheweg 80, 2501 Biel, Switzerland
kerstin.denecke@bfh.ch
[2] Innovation Center Computer Assisted Surgery,
Semmelweissstr 14, 04103 Leipzig, Germany
yihan.deng@medizin.uni-leipzig.de
[3] German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3 66123 Saarbrücken
declerck@dfki.de

Abstract. Important information is captured in medical documents. To make use of this information and intepret the semantics, technologies are required for extracting, analysing and interpreting it. As a result, rich semantics including relations among events, subjectivity or polarity of events, become available. The First Workshop on Extraction and Processing of Rich Semantics from Medical Texts, is devoted to the technologies for dealing with clinical documents for medical information gathering and application in knowledge based systems. New approaches for identifying and analysing rich semantics are presented. In this paper, we introduce the topic and summarize the workshop contributions.

1 Introduction

Pharmaceutical companies and individual patients exploiting advances in translational medicine and informational infrastructure are joining clinical interests in recording detailed patient records. The latter comprise a broad range of clinical documents including nurse letters, discharge summaries and radiology reports describing a patients health status, diagnoses, applied procedures and observations of the health care team. The rich semantics such as facts, experiences, opinions or information that are hidden in those medical documents could when extracted automatically - support a broad range of applications including clinical decision support systems. Physicians could learn about the experiences of their colleagues, get hints to critical events in the treatment of a specific patient or receive information for improving treatment. Studies on the effectiveness of clinical treatment could be realised based on the text material. The workshop on Extraction and Processing of Rich Semantics from Medical Texts (RichMedSem) brings together researchers and their work in this upcoming field. Relevant topics include on the one hand extraction methods specifically designed for the

extraction of rich semantics from medical texts. Further, the representation and storage of extracted rich semantics for further analysis is an important aspect. Semantic Web technologies are of particular interest, since they could provide the means to link this type of medical information with other data sets in a Linked Data infrastructure. What makes rich semantics, what are the challenges to be addressed and which approaches are suggested? The work presented in the RichMedSem workshop addresses these questions and answers are summarized in the following.

2 Rich Semantics

With rich semantics, we refer to concepts and their relations and characteristics described in written text. Existing methods for information extraction from clinical texts addressed the extraction of mentions of diagnoses, clinical treatments or medications mainly for the purpose of clinical coding, detection of drug interactions or contraindications. Extraction or addition of rich semantics goes beyond this. Rich semantics can include descriptions of clinical events, relations among clinical events (e.g. causality relations), but also subjectivity, polarity, emotion or even comparison, for example

- A change in the health status (e.g., a patient can suddenly feel better or worse),
- Critical events, unexpected situations or specific medical conditions that impact the patient's life (e.g., tumour is malignant as such is a fact, but this medical condition is negative for the patient since it might lead to health problems or death),
- The outcome or effectiveness of a treatment (e.g., a surgery can be successfully completed),
- Experiences or opinions towards a treatment or a sort of drug (e.g., a patient or a physician can describe serious adverse events after drug consumption),
- The certainty of a diagnosis (e.g., a physician can be certain of some diagnosis).

3 Challenges of Extracting Rich Semantics

One big challenge for extracting rich semantics from clinical texts are language peculiarities, content diversity, streaming nature of clinical documents that pose many challenges to an automatic processing, Finding the trade-off of filtering noise at the cost of losing potentially relevant information is crucial. In contrast to biomedical documents, clinical documents are often not well formulated. They can for example contain verbless clauses, writing errors, many idiosyncratic abbreviations and sentence complexity of such document ranges from few word phrases to complex sentence structures. New technologies are required for dealing with the peculiarities of clinical documents, in particular for extracting,

analysing or adding semantics, which can be included into corresponding knowledge based systems. The challenges of extracting rich semantics stem from the ultimate ambiguities caused by the objective nature of the medical text. The medical knowledge bases have also a deep influence on the outcome of the semantic meanings. For instance, events such as a bleeding can be positive or negative, critical or less critical. The phrase blood pressure decreased could express a positive or negative change depending on the previous state of blood pressure. A decrease of blood pressure can be good when it was too high before. This also shows that sentiment in clinical narratives cannot always be manifested in single terms of phrases, but the context is important.Moreover, the medical semantics include a large amount of aspects, a coherent definition and standard data schema should be established to facilitate the linking and the usage of current available ontologies in the biomedical domain.

4 Summary of the Workshop

The papers accepted in the RichMedSem workshop this year have focused on the extraction, retrieval of biomedical semantics and corpus generation in the biomedical domain, which covered the foundation of text analysis in the biomedical domain. Shafahi et.al [1] present a controlled experiment with the task of clinical guideline updating. Text-based and semantics-based methods are evaluated with the corpus obtained from the PubMed query service. The proposed methods simulated the updating process of the Dutch national guidelines of breast cancer from 2004 to 2012. New existing research literatures was used to generate updates. The approach has focused on the retrieval of new evidences at linguistic and semantic level in a specific domain. Natural language processing and concept based methods were used. Deng and Denecke [2] describe the creation of a corpus for the task of biomedical sentiment analysis. The sentiment analysis in the biomedical domain is different to the sentiment analysis in the general domain. In the biomedical domain, the polarity of the patient status is related to the clinical events. The paper describes the generation of a corpus with 300 intensive care unit nurse letters and the corresponding annotation guidelines for biomedical sentiment. Zhukova et.al [3] introduce a system architecture to process medical documents in Russian, stressing the necessity to process unstructured documents in the medical domain and to transform them in structured information that can be managed by computers. Schmidt et.al [4] present a system that allows users to perform faceted navigation over a large corpus of medical documents. These documents concern either clinical research or patient records. The case presented here focusses on the domain of nephrology and makes use of large database of patients' records. Anna Kolliakou describe in her invited talk Social media platforms and clinical records: Trend detection and intervention a very interesting aspect of interrelating medical information detected in social media and the information included in patient records. The aim is to enable healthcare professionals and policy makers to evaluate online information for emerging rumours and other health-related issues. Findings can

then be used to (i) to develop education materials for service users and the general public and (ii) link to analysis of electronic health records (EHR). Analyses of clinical data utilised in clinical record resources.

5 Concluding remarks

There is an increased awareness that rich semantics such as sentiments, opinions and other qualitative factors are relevant in ensuring individualized care. New research topics are coming up (e.g. sentiment analysis from clinical texts). Rich semantics can be in the future be used in clinical decision support systems, with particular support of semantic web technologies. There is still a huge potential to connect semantic web technologies with extraction and storage of rich semantics from biomedical texts. Clinical decision support can only be realised when the amounts of unstructured texts can be processed, on the one hand to establish databases with rich semantics; on the other hand to extract decision relevant information for specific patients targeting at improving treatment.

References

1. Mohammad Shafahi, Qing Hu, Hamideh Afsarmanesh, Zhisheng Huang, Annette ten Teije, and Frank van Harmelen. A task-based comparison of linguistic and semantic document retrieval methods in the medical domain. In *Proceedings of 1st Workshop International Workshop on Extraction and Processing of Rich Semantics from Medical Texts at ECSW, May 29, 2016 to June 2, 2016 in Anissaras, Crete, Greece*, 2016.
2. Yihan Deng and Kerstin Denecke. The generation of a corpus for biomedical sentiment analysis. In *Proceedings of 1st Workshop International Workshop on Extraction and Processing of Rich Semantics from Medical Texts at ECSW, May 29, 2016 to June 2, 2016 in Anissaras, Crete, Greece*, 2016.
3. Nataly Zhukova, Mikhail Lushnov, Timur Safin, and Maksim Lapaev. Medical text processing for smda project. In *Proceedings of 1st Workshop International Workshop on Extraction and Processing of Rich Semantics from Medical Texts at ECSW, May 29, 2016 to June 2, 2016 in Anissaras, Crete, Greece*, 2016.
4. Danilo Schmidt, Hans-Juergen Profitlich, and Daniel Sonntag. Integrated information extraction and facetted search applications in nephrology. In *Proceedings of 1st Workshop International Workshop on Extraction and Processing of Rich Semantics from Medical Texts at ECSW, May 29, 2016 to June 2, 2016 in Anissaras, Crete, Greece*, 2016.