An approach for a medical ontology based on UMLS to improve information retrieval in German language clinical text documents

Georg Petritsch[ab], Stephan Spat[a], Christian Gütl[b], Peter Beck[a]

[a] *Institute of Medical Technologies and Health Management, JOANNEUM RESEARCH, Graz, Austria*
[b] *Institute for Information Systems and Computer Media, Graz University of Technology, Graz, Austria*

Abstract and Objective

A large number of free text documents are created in clinical practice every day. These documents often contain critical information for clinical decision-making however automatic information extraction is complicated. Here we present an approach whereby an existing prototype for a medical information retrieval system is extended. Using a suitable knowledge representation, the mapping of German language free text terms and phrases to corresponding concepts is facilitated.

Keywords:

UMLS, Semantic, Information retrieval, NLP

Introduction

Clinical notes commonly contain large amounts of unstructured data. Medical text is quite specific, with its own grammar and terms, and a considerable number of acronyms and homonyms can be found. Therefore methods of classic information retrieval modelled on a „bag of words" analogy might not be that effective. We aim to improve the recall and precision of an existing Medical Information Retrieval System (MIRS) [1] by applying linguistic methods and medical ontologies. A related approach for English documents is described elsewhere [2].

Methods

One of the major challenges faced is that the documents are written in German, which limits the options for suitable medical vocabulary. Since Medical Subject Headings (MeSH) are translated into German, MeSH terms will be used to act as a gateway to UMLS. As a thesaurus, MeSH is limited to a very basic set of relations and does not possess the expressiveness of a full-scale ontology like SNOMED. Using UMLS will enable further semantic relations to be discovered.

In addition to the underlying semantic network, a natural language processing (NLP) pipeline is required to perform the pre-processing and semantic annotation of the medical narratives. We intend to make use of Apache UIMA as the NLP framework in combination with cTakes and medKAT. The latter two modules have been released by the Open Health Natural Language Processing (OHNLP) consortium[1] and focus on the processing of medical documents.

[1] https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP

Results and Conclusion

Figure 1 sketches the existing MIRS and the planned extension. Using the NLP pipeline, analysed terms from the texts will be mapped to concepts from the knowledge base. The extracted concepts and existing terms will be indexed. Users will define queries with the GUI Module. The query builder offers an interface to define concept-based search requests and relays the query to the search module. The search module will return relevant documents to the users. Since one of the main features of the existing MIRS is the classification in specialist medical subjects, special emphasis will be placed on these results.

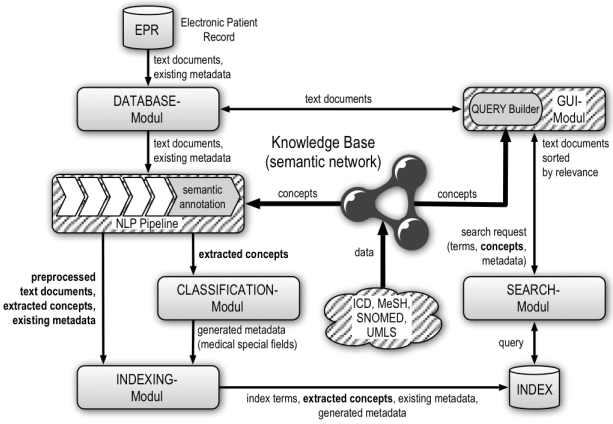

Figure 1 – Knowledge Base extending the existing MIRS (shaded areas show extended modules)

References

[1] Spat S, Cadonna B, Rakovac I, et al. Enhanced information retrieval from narrative German-language clinical text documents using automated document classification. Stud Health Technol Inform. 2008;136:473-478.

[2] Martin-Valdivia M, Montejo-Raez A, Diaz-Galiano M, et al. Expanding Terms with Medical Ontologies to Improve a Multi-Label Text Categorization System. In: Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration. Medical Info Science Reference; 2009:38–57.