Ontology Based Clinical Query Extraction

Pinar Wennerberg[1], Manuel Möller[2], Paul Buitelaar[3], Sonja Zillner[1]
[1]Siemens AG, Corporate Technology, Knowledge Management CT IC 1, Otto-Hahn-Ring 6, 81739, Munich Germany
[2]DFKI GmbH, Knowledge Management Dept., Kaiserslautern, Germany
[3]DERI - NLP Unit, National University Ireland Galway, IDA Business Park, Lower Dangan, Galway, Ireland
{pinar.wennerberg.ext | sonja.zillner}@siemens.com, manuel.moeller@dfki.de, paul.buitelaar@deri.org

Abstract

Knowledge about human anatomy, radiology and diseases that is essential for medical images can be acquired from medical ontology terms and relations. These can then be analyzed using domain corpora to observe statistically most relevant term-relation-term patterns. We argue that such patterns are the basis for more complex clinical search queries and describe our approach for deriving them. These patterns can then be used to support the knowledge elicitation process between the domain expert and the knowledge engineer by providing a common vocabulary for the communication.

Experiences along the Theseus-MEDICO[1] use case, which aims at a semantic medical search engine, have shown that collecting clinical queries through expert interviews is not effective. It requires solid medical background knowledge from the knowledge engineer, which is in most cases not available. Hence, we apply a ‘*query pattern derivation*’ technique to elicit knowledge from the medical experts about their search queries to retrieve medical images and the related patient text data. An example query pattern is: [[RADIOL. IMAGE]Modality] *is_about* [ANATOMICALSTRUCTURE] **&&** [[RADIOL. IMAGE]Modality]*shows_symptom*[DISEASESYMPTOM]

To obtain an overview of mostly likely expressed, therefore most likely queried terms we performed statistical analysis by using medical ontologies and domain corpora. The three domain ontologies and terminologies were: *Foundational Model of Anatomy*[1] for human anatomy, *The Radiology Lexicon*[2] *(RadLex)* and *The International Classification of Diseases ICD-9 CM*[3] for radiology and diseases, respectively. Our corpora about human anatomy, radiology and diseases were compiled based on the Wikipedia categories Anatomy, Radiology and Disease. Details of the ontology driven statistical data processing and the corpora are provided elsewehere[2,3]. To extract the relations between the statistically most relevant terms we implemented an algorithm that traverses each sentence, looking for the pattern [Term:Verb+Preposition:Term], where Verb+Preposition is the relation we look for. Subsequently, we extracted relations, e.g. ‘*secreted by*’ and obtained a set of term-relation-term triplets i.e. the ‘*query patterns*’, e.g. “*disorder associated with leukemia*”.

FMA	**Score**	**RadLex Term**	**Score**
lateral	338724,0	x-ray	81901,64
anterior	314721,0	imaging modality	58682,00
artery	281961,0	volume imaging	57855,09

Table 1: top 5 relevant FMA and RadLex terms in anatomy and radiology corpora respectively

We presented the most relevant query patterns to the radiology experts and received encouraging feedback. Upon seeing the patterns, he made references to other anatomy, radiology and disease related terms and added that most commonly searched terms will most likely be those that are rarely mentioned in text. That is, one searches for the less known and the less mentioned terms for purposes of self education. In future we want to assess the quality of our patterns in a semi-automatic way by comparing them to actual clinical questions to identify overlaps. Therefore, we are currently preparing a corpus of clinical questions.

Acknowledgements: This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. Paul Buitelaar contributed most of the work reported here while at DFKI.

References

1. Möller M., Sintek M., Buitelaar P., Mukherjee S.,.Zhou X.S., Freund J. “Medical Image Understanding through the Integration of Cross-Modal Object Recognition with Formal Domain Knowledge”, Healthinf 2008.
2. Buitelaar P.,Wennerberg P.,Zillner S. Statistical Term Profiling for Query Pattern Mining BioNLP-ACL, 2008.
3. Wennerberg, P. Buitelaar P., Zillner S, *Towards a Human Anatomy Data Set for Query Pattern Mining based on Wikipedia and Domain Semantic Resources*, LREC, 2008

[1] http://sig.biostr.washington.edu/projects/fm/
[2] http://www.radlex.org/viewer
[3] ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD9-CM/2007/