A Semantic Feature Space for Disease Prediction

Mariam Daoud*, Jimmy Xiangji Huang*, William Melek†, C. Joseph Kurian†
* Information Retrieval and Knowledge Management Research Lab
School of Information Technology, York University,Toronto, Canada
Email: {daoud,jhuang}@yorku.ca
†Alpha Global IT, Toronto, Canada
Email: {william,cjk}@alpha-it.com

Extended Abstract

The huge amount of data generated by modern medicine has motivated us to develop decision support systems for improving health care applications. In this paper, we address the problem of clinical disease prediction given patient-reported symptoms and medical signs where patient records lack of semantic code annotation. We propose a novel context-enhanced disease prediction approach based on leveraging semantic and contextual medical entity relations. We have already exploited semantic relations of medical terminology for patient records search [2] but they were never considered for disease prediction in the literature. Patient signs and symptoms are first mapped to SNOMED-CT concepts, which compose a feature space for disease prediction. Our major contributions in this paper consist of expanding the feature space using semantic and contextual concept relations of SNOMED-CT. Based on patient's reported signs and symptoms, we use biomedical text mining tool, namely Metamap [1] to extract concepts of the SNOMED-CT metathesaurus. A "concept" in SNOMED-CT is a clinical meaning identified by a unique numeric identifier (ConceptId) and described via a set of words. For each concept, we define a medical entity context by integrating "defining" and "qualitative" medical aspects through the use of different types of semantic and contextual relationships of SNOMED-CT. Figure 1 illustrates the concept "Pneumonia" and its relations to other concepts.

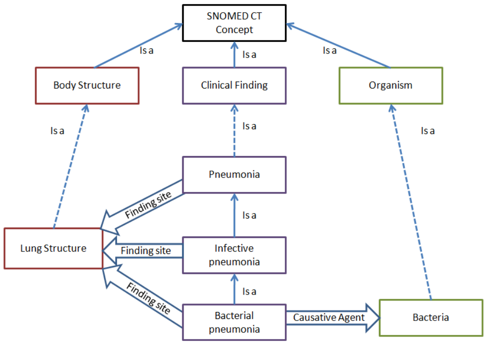

Fig. 1. Illustration of relations in SNOMED-CT

A case study is conducted on a real medical dataset provided by Alpha Global IT healthcare company located in Canada. Patient records are pre-annotated with diseases. We evaluate the impact of our proposed feature space on the disease prediction performance. Figure 2 presents the classification accuracy of the support vector machines classifier (SMO) using different types of medical relations on cardiology patient records dataset. We choose SVM for studying the impact of relations types on disease prediction since it performed best compared to other classifiers. We notice that most of the concept relations types have shown positive impact on the disease prediction accuracy where expanding the concepts with all types of medical relations (labelled "all") has performed best. The positive relations are "interprets" and "has-episodicity". Using all relations types "all" to expand the feature space provides the highest accuracy. The negative impact of some relations types could be due to a high relatedness in symptom descriptions between different diseases. When using the relations "synonyms", "same as" or "replace", the overlapping features between diseases that present few common symptoms increase, which makes the disease type hard to identify. For example, the symptoms "Breathless" and "Palpitation" are common for 16 and 12 cardiology diseases respectively where the total number of diseases is 21.

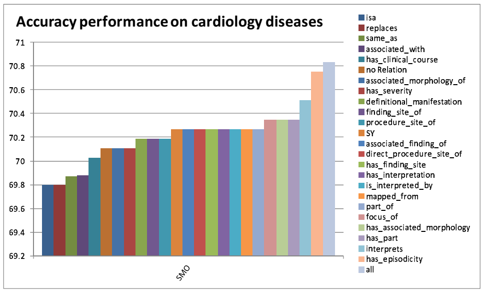

Fig. 2. Impact of relation types on cardiology disease prediction

Acknowledgements

This research is supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and Mathematics of Information Technology and Complex Systems (MITACS). We also thank reviewers for their valuable comments on this paper.

References

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In *AMIA, Annual Symposium*, pages 17–21, 2001.

[2] A. Babashzadeh, J. Huang, and M. Daoud. Exploiting semantics for improving clinical information retrieval. In *SIGIR*, pages 801–804, 2013.

978-1-4799-1310-7/13/$31.00 ©2013 IEEE