Clinical Phenotyping Using Semantic Knowledge Graphs

Ramiz Iqbal¹, BS, MS, Ken Chen¹, PhD, Elmer Bernstam², MD, MSE
¹MD Anderson Cancer Center, Houston, TX, USA; ²UT Health Science Center at Houston
School of Biomedical Informatics, Houston, TX, USA

Brief Abstract:

Clinical phenotyping is a difficult task to conduct at scale. Automating this task is confounded by the need for deep expertise in a particular domain and the terminology. We propose to construct a knowledge graph where the relatedness of UMLS terms is determined by their semantic relationship in biomedical literature. This approach would both help overcome the variability inherent to the UMLS terminology and the lack of deep domain knowledge needed to phenotype a patient.

Introduction

Clinically phenotyping a patient as having a particular disease is often a difficult task to conduct at scale and a persistent problem for healthcare systems. Automating the task of phenotyping is confounded by the need for deep expertise in particular domains and the terminology itself. The Unified Medical Language System (UMLS), a controlled vocabulary developed for the task of extracting biomedical concepts, has been utilized to overcome some of the aforementioned issues¹. However, the UMLS is known to be fraught with inconsistencies and errors which has affected the ability of researchers to use this vocabulary to automate tasks².

Methods

Previous work related to identifying phenotypic information in full-text literature, such as the PhenoCHF annotated dataset, have paid special attention to specific disease-entity relationships such as risk factors and treatments³. We are interested in extracting UMLS entities that have at least one of four basic relationships to a disease: risk factor, diagnostic criteria, treatment, and signs/symptoms. These relationships will be used as the basis for constructing a knowledge graph where the relatedness of UMLS terms is determined by their semantic relationship in biomedical literature rather than where they are in the UMLS hierarchy. Grounding the relationship between the terms on the literature filters the UMLS for the most utilized terms for a specific disease. The approach we suggest would both help overcome the variability inherent to the UMLS terminology and the lack of deep domain knowledge needed to phenotype a patient.

We have identified about 19K articles in the Pubmed Open Access Archive that are related to Congestive Heart Failure (CHF). From these articles, sentences containing the aforementioned relationships will be identified with the aid of Snorkel, a system used to create weakly labeled training data⁴. With the weakly labeled data, we will be able to train classifiers to identify the disease-entity relationships we are interested in and evaluate our extractions against the PhenoCHF corpus.

Conclusion

A patient would be phenotyped by comparing the UMLS terms in a patient’s EHR to the knowledge graph to yield a probability of the patient having a phenotype, e.g., breast cancer. This process can be conducted across thousands of records in far less time than the gold standard of having a physician review the patient’s record.

Lastly, our method is based on literature and can be expanded to phenotype other disease, therefore it is widely applicable to any healthcare system and disease.

References