

Natural Language Processing for Clinical and Translational Research

The Mayo Clinic - Rochester

PIs: Hongfang Liu, Serguei Pakhomov, and Hua Xu

Grant Number: 3R01GM102282

Rapid growth in the clinical implementation of large electronic medical records (EMRs) has led to an unprecedented expansion in the availability of dense longitudinal datasets for clinical and translational research. This growth is being fueled by recent federal legislation that provides generous financial incentives to institutions demonstrating aggressive application and "meaningful use" of comprehensive EMRs. Efforts are already underway to link these EMRs across institutions, and standardize the definition of phenotypes for large scale studies of disease onset and treatment outcome, specifically within the context of routine clinical care. However, a well-known challenge for secondary use of EMR data for clinical and translational research is that much of detailed patient information is embedded in narrative text. Natural Language Processing (NLP) technologies, which are able to convert unstructured clinical text into coded data, have been introduced into the biomedical domain and have demonstrated promising results. Researchers have used NLP systems to identify clinical syndromes and common biomedical concepts from radiology reports, discharge summaries, problem lists, nursing documentation, and medical education documents. Different NLP systems have been developed at different institutions and utilized to convert clinical narrative text into structured data that may be used for other clinical applications and studies. Successful stories in applying NLP to clinical and translational research have been reported widely. However, institutions often deploy different NLP systems, which produce various types of output formats and make it difficult to exchange information between sites. Therefore, the lack of interoperability among different clinical NLP systems becomes a bottleneck for efficient multi-site studies. In addition, many successful studies often require a strong interdisciplinary team where informaticians and clinicians have to work very closely to iteratively define optimal algorithms for clinical phenotypes. As intensive informatics support may not be available to every clinical researcher, the usability of NLP systems for end users is another important issue. The proposed project builds upon first-hand knowledge and experience across the research team in the use of NLP for clinical and translational research projects. There are several big informatics initiatives for clinical and translational research but those initiatives generally assume one shoe fits all and follow top-down approaches to develop NLP solutions. Complementary to those initiatives, we will use a bottom-up approach to handle interoperability and usability: i) we will obtain a common NLP data model and exchange format through empirical analysis of existing NLP systems and NLP results; ii) we will develop a user-centric NLP front end interface for NLP systems wrapped to be consistent with the proposed NLP data model and exchange format incorporating usability analysis into the agile development process. All deliverables will be distributed through the open health NLP (OHNLP) consortium which we intend to make it more open and inclusive. PUBLIC HEALTH RELEVANCE: Rapid growth in the clinical implementation of large electronic medical records (EMRs) has led to an unprecedented expansion in the availability of dense longitudinal datasets for clinical and translational research. We propose the development of a novel framework to enable the use of clinical information embedded in clinical narratives for clinical and translational research.