Automated Tools for Phenotype Extraction from Medical Records

Meliha Yetisgen-Yildiz, PhD[1,2], Cosmin A. Bejan, PhD[1], Lucy Vanderwende, PhD[5,1], Fei Xia, PhD[2,1], Heather L. Evans, MD, MS[3], Mark M. Wurfel, MD, PhD[4]
[1]Biomedical and Health Informatics, [2]Department of Linguistics, [3]Department of Surgery, [4]Pulmonary and Critical Care Medicine, University of Washington, Seattle, WA; [5]Microsoft Research, Redmond, WA

Abstract

Clinical research studying critical illness phenotypes relies on the identification of clinical syndromes defined by consensus definitions. Historically, identifying phenotypes has required manual chart review, a time and resource intensive process. The overall research goal of Critical Illness PHenotype ExtRaction (deCIPHER) project is to develop automated approaches based on natural language processing and machine learning that accurately identify phenotypes from EMR. We chose pneumonia as our first critical illness phenotype and conducted preliminary experiments to explore the problem space. In this abstract, we outline the tools we built for processing clinical records, present our preliminary findings for pneumonia extraction, and describe future steps.

Introduction

With the introduction of comprehensive electronic medical records (EMRs), all aspects of patient care can now be captured in both structured and free-text format. The existence of such data provides an opportunity (1) to improve patient care by identifying critical illness phenotypes in a timely manner and (2) to facilitate clinical and translational studies of large cohorts of critically ill patients, a task that would not be feasible using traditional screening/manual chart abstraction methods. At the University of Washington, we built a series of general purpose tools to process free-text medical reports including a statistical section segmentation approach[1] to chunk a given medical record into its main sections and an assertion analysis tool[2] to analyze the certainty level of a given concept in the context it appears in text (e.g., present, absent). To capture the syntactic and semantic knowledge, we extensively used an NLP toolkit called SPLAT developed by Microsoft Research for tokenization, POS tagging, and dependency parsing as well as UMLS taggers such as MetaMap for capturing the medical concepts in reports.

Example Phenotype – Pneumonia Extraction from ICU Reports

The main components of the system architecture for pneumonia extraction are depicted in Figure 1. As illustrated, we designed the architecture on top of a supervised machine learning framework, where features associated with each data instance (i.e., patient) are automatically extracted to be used by a binary classifier. We implemented a statistical feature extraction methodology to select only the most informative features for pneumonia identification. To train and test our approach, we used physician notes from a cohort of ICU patients in our institution. The dataset included 5313 ICU physician notes (including admit notes, daily ICU progress notes) created for the 426 patients. Our initial pneumonia screening approach was compared to determinations of the presence or absence of pneumonia during each patient's ICU stay obtained through manual abstraction (positive cases: 66, negative cases: 360). With this dataset, we achieved 0.82 F1-measure[3].

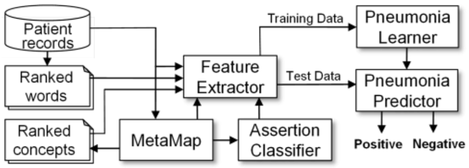

Figure 1. System Architecture

Future Steps

The presented methods are phenotype and institution independent and can be easily applied for other phenotypes with enough training data. We are currently at work on another important clinical application designed to determine the time-of-onset of the phenotypes. We are in the process of building our training sets for different types of pneumonia and will present our preliminary results in the poster.

Acknowledgements

Microsoft Research Connections, P50 HL073996, RC2 HL101779, UL1TR000423

References

1. Tepper M, Capurro D, Xia F, Vanderwende L, Yetisgen-Yildiz M. Statistical Section Segmentation in Free-Text Clinical Records. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), 2012.
2. Bejan CA, Vanderwende L, Xia F, Yetisgen-Yildiz M. Assertion modeling and its role in clinical phenotype identification. To appear in J Biomed Inform. 2012.
3. Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. J Am Med Inform Assoc. 2012;19(5):817-23.