Extraction of Pneumonia Cases from Free-Text Intensive Care Unit Reports

Meliha Yetisgen-Yildiz, PhD[1,3], Bradford J. Glavan, MD[2], Fei Xia, PhD[3,1], Lucy Vanderwende, PhD[4,1], Mark M. Wurfel, MD, PhD[2]
[1]Biomedical & Health Informatics, [2]Pulmonary and Critical Care Medicine, School of Medicine, [3]Department of Linguistics, University of Washington, Seattle, WA
[4]Microsoft Research, Redmond, WA

Abstract

Clinical research studying critical illness phenotypes relies on the identification of clinical syndromes defined by consensus definitions. A prime example is the pneumonia phenotype. Historically, identifying pneumonia has required manual chart review, a time and resource intensive process. The overall research goal is to develop automated approaches that accurately identify critical illness phenotypes. In this poster, we outline our approach to the identification of pneumonia from EMR, present our preliminary results, and describe future steps.

Introduction

While large administrative datasets of Intensive Care Unit (ICU) patients exist, they lack the granular data necessary to accurately identify complex phenotypes and determine the relative timing of events during the course of critical illness. With the introduction of comprehensive electronic medical records (EMRs), all aspects of ICU care can now be captured in both structured and free-text format. The existence of such data provides an opportunity to identify critical illness phenotypes and 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. Our goal is to build automated tools to identify critical illness phenotypes in ICU data. In this study, we chose *pneumonia* as our first critical illness phenotype and conducted preliminary experiments to explore the problem space.

Preliminary Experiments

To determine the feasibility of developing an automated classification approach to identify pneumonia cases, we used physician notes from a cohort of ICU patients at Harborview Medical Center. The dataset included 5313 free-text ICU physician notes (including Admit notes, daily ICU progress notes, and transfer/discharge notes) created for the 426 patients during their ICU stay. 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 (# of cases positive for pneumonia: 66, # of cases negative for pneumonia: 360). In our initial experiments, we represented the content of free-text reports with word unigrams. We also used MetaMap (Available at: http://metamap.nlm.nih.gov/) to identify UMLS concepts in free-text reports and used the concepts and their semantic types in our representation as domain specific features. For each patient, we created a feature vector from the features extracted from the patient's ICU in-patient reports and trained a Maximum Entropy (MaxEnt) classifier.[1] To measure the performance, we used 5-fold cross validation. For the positive class, the classifier produced 0.59 precision, 0.41 recall, and 0.48 F1-score. For the negative class, the classifier produced 0.90 precision, 0.95 recall, and 0.92 F1-score. The overall accuracy of the classifier was 0.86. More detailed information about the experiments can be found in a recently published paper.[2]

Conclusion & Future Work

The preliminary results demonstrate that (1) our initial set of features can achieve promising classification performance as measured with respect to manual chart annotation, and (2) more work needs to be done to improve precision and recall for the positive class prediction for pneumonia. As next steps, we will introduce more sophisticated features that capture the syntax and semantics of the free-text reports. We will also incorporate the structured numeric values (e.g., white blood cell counts, body temperature) as features for the classification process.

Acknowledgements

P50 HL073996, RC2 HL101779, Microsoft Research Connections

References

1. Nigam K, Lafferty J, and Mccallum A. Using maximum entropy for text classification. In Proc. of IJCAI-99 Workshop on Machine Learning for Information Filtering. 1999; 61-67.
2. Yetisgen-Yildiz M, Glavan BJ, Xia F, Vanderwende L, and Wurfel MM. Identifying Patients with Pneumonia from Free-Text Intensive Care Unit Reports. In Proc of Learning from Unstructured Clinical Text Workshop of the 28th International Conference on Machine Learning (ICML), July 2, 2011.