Feasibility of pooling annotated corpora for clinical concept extraction

**Kavishwar Wagholikar, MBBS, PhD[1], Manabu Torii, PhD[2],
Siddhartha Jonnalagadda, PhD[1], Hongfang Liu, PhD[1]**

[1]Mayo Clinic, Rochester, MN; [2]Georgetown University, Washington, DC

Abstract

Availability of annotated corpora has facilitated application of machine learning algorithms to concept extraction from clinical notes. However, it is expensive to prepare annotated corpora in individual institutions, and pooling of annotated corpora from other institutions is a potential solution. In this paper we investigate whether pooling of corpora from two different sources, can improve performance and portability of resultant machine learning taggers for medical problem detection. Specifically, we pool corpora from 2010 i2b2/VA NLP challenge and Mayo Clinic Rochester, to evaluate taggers for recognition of medical problems. Contrary to our expectations, pooling of corpora is found to decrease the F1-score. We examine the annotation guidelines to identify factors for incompatibility of the corpora and suggest development of a standard annotation guideline by the clinical NLP community to allow compatibility of annotated corpora.

Introduction and Background

Many medical institutions have an interest in using Natural Language Processing (NLP) to utilize unstructured text in their electronic medical record (EMR) systems. Individual institutions could benefit from using shared and publicly available resources. There have been similar efforts to pool datasets in the biomedical domain. In this paper we investigate whether pooling of similar datasets from two different sources, can improve performance and portability of resultant machine learning taggers for medical problem detection.

Methods

We trained and tested taggers on a dataset from Mayo Clinic, Rochester (MCR) and a dataset from the 2010 i2b2/VA NLP challenge. The taggers were trained to recognize medical problems, including signs/symptoms and disorders. Firstly, we trained the tagger on i2b2 dataset and tested it on MCR dataset and vice versa. We then performed 5 fold cross-validation on each of the datasets. We repeated the cross-validation on MCR dataset after supplementing the training fraction with the i2b2 dataset. This design was repeated for the i2b2 dataset by using MCR data to supplement the training. Precision, recall and F1-score performance measures were computed for the experiments.

Results and Discussion

Taggers trained on annotated corpus from the same institution performed the best. Pooling of corpora decreased the F1-score. We examined the annotation guidelines to identify factors that led to the incompatibility of the datasets. These included differences in concept definition, and whether articles, possessive pronouns, prepositional phrases and conjunctions where included in the concept spans. We suggest the development of a standard annotation guideline by clinical NLP community to allow compatibility of annotated corpora.