Text Processing Tools from the University of Washington Biomedical Language Processing Group

Meliha Yetisgen-Yildiz, PhD[1,2], Cosmin A. Bejan, PhD[3], Prescott Klassen, MS[2], Michael Tepper, MS[2], Lucy Vanderwende, PhD[4,1], Fei Xia, PhD[2,1]
[1]Biomedical and Health Informatics, [2]Department of Linguistics, University of Washington, Seattle, WA; [3]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN; [4]Microsoft Research, Redmond, WA

Abstract

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). In addition we released a statistical feature selection tool[3] for our classification tasks. 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.

Statistical Section Segmenter

Although the clinical reports are in free-text, they are structured in terms of sections to describe the clinical information. We developed a statistical approach to identify the boundaries of the sections and their types[1]. Our approach requires (1) construction of an ontology of section headers for a selected clinical report type (e.g., discharge summary) and (2) annotation of a corpus of notes with the constructed ontology to be used for training. Our basic methodology for section segmentation is to classify each line in a document to indicate its membership to a section. We built two separate models for section segmentation and classification. First, the section boundaries are identified by labeling each line with a B (beginning of section), I (inside of section), O (outside of section) tag. Then the unlabeled sections from the first step are passed to the second step, where a separate classifier is called upon to label each section with the appropriate section category. We achieved 0.932 precision, 0.911 recall, and 0.921 f-score based on a gold standard composed of 100 manually annotated radiology reports with respect to both section boundaries and section label.

Assertion Analyzer

We built a statistical assertion analyzer based on the 2010 i2b2/VA NLP challenge assertion task data[2]. The purpose of the assertion task was to classify the assertion value of a medical concept expressed in a free-text report as present, absent, possible, conditional, hypothetical, or associated with someone else. While building our assertion classifier, we defined a comprehensive list of semantic and syntactic features to model the assertion values and achieved the state-of-the-art 0.942 micro-averaged f1-score with support vector machines.

Statistical Feature Selector

Feature selection algorithms have been successfully applied in text categorization in order to improve the classification accuracy. We built a statistical feature selection tool for our classification tasks[3]. Our tool uses statistical significance tests to measure the association between each feature from the training set and two categories of the classification task. It ranks the features based on the χ^2 and t statistics values and selects only the most relevant features that are within a specific threshold. The selected features are used for training and testing.

All the tools described in this abstract are available for download at the UW-BioNLP group website (http://depts.washington.edu/bionlp/software.htm).

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. J Biomed Inform. 2013; 46(1):68-74.
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.