A Suite of Natural Language Processing Tools Developed for the I2B2 Project

Sergey Goryachev, MS, Margarita Sordo, Ph.D., Qing T. Zeng, Ph.D.
Decision Systems Group, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA

ABSTRACT

Textual medical records contain a wealth of information that needs to be extracted and interpreted by the automated tools. We have developed a collection of natural language processing (NLP) tools to extract various types of information from unstructured medical records. The generic NLP components, when assembled in pipelines and initialized with custom configuration parameters, become a powerful medical data mining instrument. We have successfully extracted such medical concepts as diagnoses, comorbidities, discharge medications, smoking status, and discharges that need to be extracted and/or indexed in order to be manipulated by the automated tools. We took a modularized and parameterized approach in the software development and employed syntactic, statistical, and template-based methods for different parsing tasks. This approach allows users to tailor the NLP tools to extract and index specific information from different domains and institutions.

METHODS AND RESULTS

We have developed 11 modules for text report processing (Figure 1):

1. Section Splitter
2. Section Filter
3. Text Tokenizer
4. Part-of-Speech (POS) Tagger
5. Noun Phrase Finder
6. UMLS Concept Finder
7. Negation Finder
8. Regular Expression-based Concept Finder
9. Sentence Splitter
10. N-Gram Tool
11. Classifier (e.g. Smoking Status Classifier)

These modules were applied to discharge summaries and outpatient notes from 2 institutions, Brigham and Women’s Hospital and Massachusetts General Hospital with minimum changes in the configuration files. They were also used to extract key data items from a set of medical error reports, which involved adding several new modules, but didn’t require any alteration of the original 11 modules.

Figure 1. NLP components for medical report processing assembled into pipelines for various information extraction tasks.