Biological Entity and Relationship Extraction

Robert Tung and Adrian Lin
Yale University, New Haven, CT
robert.tung@yale.edu
adrian.lin@yale.edu

Dragomir Radev (Advisor)
Yale University, New Haven, CT
dragomir.radev@yale.edu

Abstract

It is a well-known problem that in many fields of study today, there is an excess of data and a lot of room for drawing aggregate, concrete findings from them. In fact, NBCNews reported in 2005 that $95 billion dollars a year are spent on medical research [?].

For this project, we have particularly been interested in biomedical papers concerning the relationships between genes and diseases. This project seeks to use natural language processing methodologies to create a pipeline that ultimately determines the probability that each pair of genes and diseases in a paper is related.

This project begins constructing this pipeline by scraping a corpus of papers and converting them to a usable format. Next, this project uses the becas library to recognize genes and diseases in the text, as well as an original syntactic named-entity recognizer for genes. Then, features are extracted including distances between key words and others in the sentence, word2vec representations of each word, and features concerning the dependency-tree parse of each sentence. Finally, this is fed into a convolutional neural network architecture that predicts the probability of each gene-disease pair being related.

This project also includes a Web Application that can take in a new biomedical paper, use the model trained as described above, and in turn determine the probability that each pair of genes and diseases in that paper are related.

This abstract and the respective paper focus primarily on the named-entity recognition, the word2vec and dependency-tree feature extraction, and the Web Application; it additionally concerns small portions of the neural network. This comprises the contributions of Robert Tung to this project; an analogous abstract and report exist for those of Adrian Lin.