

Deep Learning for Speculative Language Recognition in Biomedical Articles

Jia-shen Boon Rasiga Gowrishankar Akshay Sood

Fall 2014

Abstract

1 Introduction

//include motivation here

This project applies deep learning to the task of speculative language recognition in biomedical articles, which has potential applications in Information Retrieval, Question Answering systems, Multi-Document Summarization and Knowledge Discovery.

Hypothesis Applying deep learning to the task of speculative language recognition in biomedical articles improves the performance in the task, compared to the baseline algorithms.

2 Approach

// describe your work, the modifications made to existing algos.
// provide complete reference to descriptions.
// dont describe code / implementation details. (describe algo in sufficient detail)

Paragraph vector representation

Our neural network approach comprises two networks. One uses the Paragraph Vector algorithm[1], an unsupervised approach to learn a fixed length feature representation of a sentence, both at training and test time. The second network takes the sentence vector learnt from the first network to predict whether this sentence contains speculative language. We also cross validate the window

size.

Recursive Neural Networks

- Parse Tree Generation
//Stanford parser
//running example
- Labelling parse tree
//algo used for labelling
// labelled example tree
- Training the model
//parameters that can be varied
- Testing new sentence
//show example - parse tree - final label given by model
- Graph (include here or include all graphs in evaluation?)
// number of epochs Vs. Accuracy/ F-Score?
// ROC curve
// Precision Recall curve

3 Baselines

- Support Vector Machines
- Naive Bayes
- Substring matching

4 Evaluation

Empirical Evaluation

// data set description
// selection of data (training, testing , validation)
// setting parameters for algos
// what are you trying to test/demonstrate in your experiments

Analytical Evaluation

// proofs or other formal results

5 Discussion

A better word relationship evaluation set should be developed specific to biomedical language. The list of word relationships provided by the word2vec project includes semantic relationships (e.g., Italy is to Rome as France is to Paris) and syntactic relationships (e.g. run is to running as grow is to growing). However, as much of the vocabulary in the semantic relationships are absent in biomedical language, and hence absent in our trained models, we had to exclude these relationships during evaluation. We suggest creating a set of semantic relationships for biomedical language (e.g., headache is to aspirin as cough is to dextromethorphan).

//lessons of your exp?

//limitations of the approach

// future work direction

6 Bibliography