Project Report

Team Members:

Chukwubuikem Ume-Ugwa

Text Corpus: Corpus

In this project, I performed text categorization by attempting to reproduce the experimental results in this paper. The goal is to classify Reddit posts as one of ten discourse acts.

The rationale behind the project is to see how well I can replicate the work of others and to improve my understanding of the paper. I implemented a system that given a Reddit post title, subreddit and post content, it classified the post into one of the ten discourse acts identified in the paper.

I followed the typical methodology for creating a classification model. Since the data was structured, the main challenge was to figure out which features provide better predictive capability. I didn't have much time to explore all the different combinations, I chose to use the title, subreddit and post content as the features. These features were then combined together and used to compute the embedding for a Reddit post. The embedding was computed by combining 1-3 grams TF-IDF vector at the character level for text features and numerical structural information vector that was normalized with a minimax scaler. The numerical structural information include number of characters in the title and content, number of words in the content, number of sentences in the content and the post depth in the sureddit thread.

Two classification models were explored to find the best model for the problem. The classification models explored were gradient boosting classifier and logistic regression classifier. These were chosen because they are good for multilabel classification tasks. The two models were trained on 80% of the data and validated on 20%. The gradient boosting classifier performed better with an accuracy score of 67% on validation set versus 65% by logistic regression. Consequently, the gradient boosting classifier was chosen as the model for this task.

The gradient boosting classifier was then trained on the full corpus and achieved an accuracy score of 70%.

The system comes with a gui, which provides a user with input fields to supply the post title, subreddit, and content. Using this information, the system loads an already trained gradient boosting classifier and uses it to compute the discourse act for the given subreddit post.