ANLY 580 Natural Language Processing

Professor Linda Moreau & Dan Loehr

Cheng Zhong

Oct 28th, 2018

# Project Proposal

**Summary:**

This project will be mainly focus on identifying questions pairs intent from Quora. By applying deep learning and nature language processing techniques, the project will give a way to classify whether question pairs are duplicates or not. Hence it will make the overall improved experience for Quora users.

**Project's Goal and Objectives:**

This project will be based on the Kaggle Competition sponsored by Quora. Quora is a platform to ask questions and connect with people who contribute unique insights and quality answers. According to Quora's description, over 100 million people visit Quora every month, and there exists lots of people ask similarity worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. The Goal of this project will be predicting which of the provided pairs of questions contain two questions with the same meaning using deep learning and nature language processing techniques in order to help the Quora platform to have a greater user experience.

**Data:**

The Data of this project will be offered by Quora through Kaggle. The dataset will be containing two csv files. One is called ***train.csv*** and the other is called ***test.csv***. The train.csv and test.csv dataset will have 6 columns of data including following information:

- Id – the id of a training set question pair
- qid1, qid2 – unique ids of each question (only available in train.csv)
- question1, question2 – the full text of each question
- is_duplicate – the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

The train dataset has more than 400k records to train the model and we have over 100k record as the test dataset to test the result. The ground truth is the set of labels that have been supplied by human experts. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty.

**Assessment Metrics:**

Since this is a classification problem, the result will be evaluated on the log loss between the predicted values and the ground truth. The reason chooses log-loss function here is because log-loss function heavily penalizes classifiers that are confident about and incorrect classification. For example, if the classifier assigns a very small probability to the correct class then the corresponding contribution to the log-loss will be very large indeed. Then it will be going to have a significant impact on the overall log-loss for the classifier. Hence the loss function and assessment metrics use here will be log-loss value. The smaller log-loss value will indicate a better prediction. A submission file will be submitted to Kaggle and it will have a score to evaluate the result.

**Approach:**

Quora currently using the random forest model to identify duplicate questions. The current idea of finishing this project is to use the basic CNN (Convolutional Neural Network) to train the model. Although LSTM model might be more useful in nature language processing, the CNN model could probably have better performance in avoid overfitting. Also, CNN could be somehow faster than LSTM based on a large sample. However, there do have limitations in this CNN approach since it does not have a good sense in dealing with sentences have same word but different word orders. The project will also implement a random forest model to compare if time permits.

**Timeframe:**

The timeframe for this project will be in Nov 5th 2018 – Dec 3rd 2018.