CS 410 Project Progress Report

Yiao Ding (Captain, Team Leader)

yiaod2@illinois.edu

1. **Progress made thus far**
   As mentioned in the project proposal, the three main steps for this project are first to work on an algorithm to extract the syntactic structure of a sentence, then work on transforming the extracted structure into vector representation, and finally working on creating a machine learning model to train the data and make predictions. Currently I have successfully completed the first step and the second step. For the first step, I used the code and algorithm which is inspired by Stanford Parser to extract the subject, predicate and object from a sentence, where those sentences are the data from Microsoft research paraphrase identification (MSRP) dataset. For the second step, I utilized Gensim word2vec glove-twitter-50 model and pretrained data, with this I can turn each word into a 50-dimensional vector, so the extracted subject, predicate and object will from a 50 x 3 matrix which can later be the input into the machine learning model.

2. **Remaining tasks**
   The remaining tasks is the third step which including creating the machine learning model to train the data and make predictions. The overall architecture of the machine learning model to be created will include two convolution layers and one pooling layer.

3. **Any challenges/issues being faced**
   There are no obvious issues for now. In the future when creating the machine learning model mentioned above, parameters in the model like kernel size, padding and step size are not determined, efforts are required to optimize these parameters so that the model can predict the similarity between sentences with a high accuracy.