EE/CprE/SE 492 GROUP PROGRESS REPORT

Group number: 17
Project title: Mining and Evaluating Verb tags and Other Important POS tags inside Software Documentation
Client: Hung Phan, Hiep Vo, Arushi Sharma
Advisor: Ali Jannesari
Team Members: William Sengstock, Austin Buller, Kelly Jacobson, Jacob Kinser, Sam Moore, Zach Witte, Dan Vasudevan

---

Project Summary:

The goal of our project is to research and experiment with different natural language processing (NLP) techniques to try and find the best way to tag (part of speech) software documentation. We want to use existing NLP models and techniques to create our own model that is best suited for processing software documentation and source code. The direction of the project has not greatly changed, but we think taking a neural network/deep learning approach might be the best way to address the issue of processing software documentation and code.

Accomplishments

As a group, we spent this last week researching neural networks and how they can be applied to the processing of source code.

Kelly - One resource we looked at was CuBERT, which is a research project by Google that is in the same area as our project -- applying natural language processing to source code. I spent some time reading about this project and looking at the code to get a better understanding of what they were doing. The biggest takeaway was that tokenization of code lines is an extremely important step, and we will most likely have to develop our own extensive library of token labels to train a neural network with.

Zach - I looked into what role deep learning and neural networks have when processing source code. I found a paper titled “Deep Learning for Source Code Modeling and Generation: Models, Applications and Challenges.” This gave me some insight into why deep learning is probably the best way to approach the problem of processing source code. For example, deep learning is able to perform four main things: “Automatic feature generation”, “capture long-term dependencies as well as sequential property”, “end-to-end learning”, and “generalizability”.

Jacob - I researched the basics of neural networks and what they are typically used for. In addition, I looked into existing projects that used neural networks when analyzing source code. I took note of the techniques shown in these projects as we look to perform similar steps when analyzing software documentation.

Austin - Recently, I've been researching neural networks and how they learn. I looked into the math of how neurons pass along values to the next layer. While most of the math is complex there are some simpler factors like the bias that can have an effect on the network output. We

will also need to create or find a tagged data set to train our neural network on because properly labeled data sets decrease the amount of time it takes for a network to learn.

Dan - I reviewed all of the findings we had last semester and did some research into Neural Networks and how to connect them to POS tagging. Based on my research I concluded that RNNs/LSTM models are the best way to go about this. I also made a presentation with my findings and showcased them to the group.

William - Throughout the course of this week I brushed up on my understanding on neural networks. We briefly started to learn about them at the end of last semester, so I did some research on what they are and how they operate with natural language processing. Once I had some background information about them, I looked into some techniques that are used for data, such as BERT and CuBERT, analyzing how they will relate to the experiments our group will conduct later down the road.

Pending issues None

One issue was scheduling a time for our group to meet because one of our clients is currently in a different time zone. That being said, we are proceeding with the meeting around our normal meeting time, conversing with another client while the situation gets figured out.

Advisor Input/Signature:
Please select one of the options below and sign.
___X____ I am pleased with the progress the team is making.
_____ The teams progress could use some minor improvements which I will discuss with them.
_____ The team's progress has some major concerns that I will discuss directly with Dr. bigelowbigelow@iastate.edu, 515-294-4177

Signature: Ali Jannesari

Client Input/Signature:
Please select one of the options below and sign.
___x____ I am pleased with the progress the team is making.
_____ The teams progress could use some minor improvements which I will discuss with them.
_____ The team's progress has some major concerns that I will discuss directly with Dr. bigelowbigelow@iastate.edu, 515-294-4177

Signature: Arushi Sharma
