Method Mention Extraction from Biomedical Text

Progress Report 2

Waqar Kalim

COMPSCI 4490Z

Department of Computer Science

Western University

26th February 2021

Project Supervisor: Dr Robert Mercer, Computer Science Department

Course Instructor name: Prof. Nazim Madhavji

1. **Emerging Result(s)**
   1. **Rule-based extraction of method mentions from research dataset**

We created a rule-based model for extracting method mentions from our research dataset, and the results are moderate. Although it is successful, the simplistic nature of the rules is an issue with the machine learning algorithms.

* 1. **Implemented a Conditional Random Field (CRF) model**  
     Our CRF model was trained on the silver-standard corpus generated by the rule-based model. Performance metrics were excellent, but a closer look shows that it figured out the rules we were using, thus preventing it from generalizing.
  2. **Implement a Bi-directional Long Short Term Memory (Bi-LSTM) model**

We are attempting to implement a Bi-LSTM model that uses the biomedical word embeddings, BioWordVec. We are having issues loading the word embeddings into memory as the binary file containing the word embeddings is too big (26Gb).

1. **Work still to be done to project completion**
   1. **Improve the rules used to construct the silver standard corpus**Unfortunately, the rules we had used to form our corpus were too simple for the machine learning models. And this prevented them from generalizing. We hope to resolve this issue by making more complex rules for our rule-based model.
   2. **Loading the BioWordVec word embeddings file into memory**  
      In order to implement a Bi-LSTM model, we still need to access the BioWordVec binary file. This task is proving to be a challenge; however, we are attempting different workarounds to access the data.
   3. **Reporting performance metrics**  
      When all of the models have been trained and are working, we will need to analyze them and report their performance metrics. In addition to that, we will also need to perform hyperparameter tuning on them where possible.
2. **Any challenges or problem areas, if any.**

* So far, the only serious challenge we have faced is the issue with loading the word embeddings file for the purpose of implementing the Bi-LSTM model.