University of Massachusetts Lowell, Computer Science

Week 3 Report

Michael Meding

March 19, 2015

Over spring break I got our baseline up and running. The baseline code is supplied by SemEval but spending the time to understand how it works is important. The baseline logistic regression model using the supplied data of 11530 samples gives us an accuracy using the dev test set of about 60%.

This result is actually not bad and being able to improve on this may prove to be quite challenging. According to the results from the competition, more than a quarter of the teams did worse than the initial baseline.

In addition to running and reading through the baseline code I read the article Named Entity Recognition in Tweets:An Experimental Study by Alan Ritter, Sam Clark, Mausam and Oren Etzioni. This article was one of the recommended readings on the SemEval task page and seemed to be a useful feature for this project. The article details how their new T-ner system outperforms the Stanford NER system and how the model is trained for this task. This is useful for our project as tweet segmentation is largely dependent on identifying the subject of that tweet.