University of Massachusetts Lowell, Computer Science

Week 2 Report

Michael Meding

March 2, 2015

This week Hoanh and I met up and decided that we would be doing the SemEval 2015 Task 1. This task involves paraphrasing Tweets to detect if they are referencing the same topic. SemEval was the logical choice for our task this semester as it did not require us to spend time generating a dataset to serve as our gold standard. Additionally, included with this task is baseline python code for us to get started thus streamlining our design process and allowing us to focus on our algorithm for the task.

Andrew, Goldberg (2007). Automatic Summarization was one of the articles that I read while beginning my research into text paraphrasing and topic detection. The article details several different approaches using both unsupervised and supervised machine learning and the pros and cons of each. TextRank is another technology that the article talks about for use with finding lexical similarity between text units similar to Word2Vec as discussed in class.

Hercules, Dalianis (2003). Porting and evaluation of automatic summarization. This article talks about attempting to implement a text summarization algorithm to several Scandinavian languages and contains some good information about implementing state-of-the-art text extraction algorithms which may prove useful for this task.

To conclude, I found that text summerization an paraphrasing is quite ambiguous and relies heavily on supervised models that are trained specifically to the articles to which they will be analysing. Furthermore, Twitter data is hard to train for as it is very short with only 140 characters or less and the topics are extremely broad. This leads to a very challenging supervised model or a rather complicated unsupervised model to fit the data correctly which is what will be required for this task.