SemEval 2015 Task 1

Mike Meding & Hoanh Nguyen

March 4, 2015

1 Introduction

Our choice for the project this semester is SemEval 2015 Task 1. This task has to do with paraphrasing tweets to divulge similarity between them. In bold below is the problem as stated directly from the SemEval website. **Given two sentences, the participants are asked to determine whether they express the same or very similar meaning and optionally a degree score between 0 and 1.**

2 Relevant Text

- Andrew, Goldberg (2007). Automatic Summarization
- Hercules, Dalianis (2003). Porting and Evaluation of Automatic Summarization
- Wei Xu and Alan Ritter and Chris Callison-Burch and William B. Dolan and Yangfeng Ji (2014). Extracting Lexically Divergent Paraphrases from Twitter
- **These are in addition to the related readings on the SemEval page**

3 Approach

The great part about doing a project from SemEval is that all baseline models and data are given. For our baseline we have the option of choosing between two models. One is a logistic regression model with an f-score of 0.6 using the included test and training data, the other model is a weighted factorization matrix with similar results. Both of these baselines are written in Python and for the purposes of our implementation may be rewritten to Java. This remains undecided as of this writing.

The tools we plan to use are the standard NLP state-of-the-art toolkits such as POS taggers and sentiment analysers. All of which are well documented for use in both Java and Python.

4 Experiments & Evaluation

Given these baseline models we will likely improve them using semantic analysis tools such as word2vec. We may also experiment with sentiment scores to see how it effects our results. The evaluation of our results has also been standardized for this task. Included with this task is an output evaluator which will grade our results giving us an f-score thus eliminating the need for complicated evaluation methods. This method was used for the competition and is the accepted standard to which we can reflect upon any improvements that we make.

5 Data

The data for this project has been provided for us by the SemEval team. This data includes 4727 development sentence pairs, 972 test pairs, and 13063 pairs for training our model. The pairs are completely random and are organized into sets of semantically similar tweets with their respective paraphrases extracted into categories for use as a gold standard. Additional information regarding the structure of the data can be seen in the Extracting Lexically Divergent Paraphrases from Twitter paper above.