CS221 Project Proposal

melvinj, icaswell, zsilver
Melvin Johnson Premkumar, Isaac Caswell, Zane Silver

We have decided to participate in the SemEval 2015 task titled **Paraphrase and Semantic Similarity in Twitter**.

1 Task Definition

Given two sentences, we have to determine whether they express the same or very similar meaning and optionally a degree score between 0 and 1. The input to the system is two sentences from Tweets in Twitter. The system has to predict whether they express the same meaning and also produce a similarity number between 0 and 1.

Example:
Input
Roberto Mancini gets the boot from Man City
Roberto Mancini has been sacked by Manchester City with the Blues saying
Output
Yes (prediction)
0.84 (similarity score)
Input
WORD OF JENKS IS ON AT 11
Word of Jenks is my favorite show on tv
Output
No (prediction)
0.23 (similarity score)

2 Scope

The task of paraphrasing and textual similarity are critical to many NLP applications, such as summarization, sentiment analysis, textual entailment and information extraction etc. Also, this task tries to promote this line of research in the new challenging setting of social media data, and help advance other NLP techniques for noisy user-generated text in the long run.

3 Data

The training dataset contains about 17,790 annotated sentence pairs, and comes with tokenization, part-of-speech and named entity tags. Hashtags have been removed. It is well balanced, consisting of 35% positive examples (paraphrases) and 65% negative examples (non-paraphrases). The data set was selected semi-randomly from among Twitter's trends and annotated by 5 Mechanical Turkers, with good correlation to expert judgements. The testing data set consists of a further 1K examples from a different time period, annotated by an expert.

4 Baseline and Challenges

The task has already produced two different baselines for us.

4.1 Supervised: Logistic Regression

This is a logistic regression model using simple lexical features. This baseline uses the NLTK toolkit to extract the features. It achieves a maximum F-score of 0.6 when trained on train.data and tested on dev.data.

4.2 Unsupervised: Weighted Matrix Factorization

Weighted Matrix Factorization (WTMF) is a latent variable model to extract nuanced and robust latent vectors for short texts/sentences, such as tweets, SMS data, short forum posts/comments. To overcome the sparsity problem in short texts/sentences (e.g. 10 words on average), it explicitly models the missing words.

Humans are exceptionally good at understand the implicit meaning of sentences and are particularly good at recognizing paraphrases of sentences. Typically, humans are known to get very high F1 scores in paraphrasing. Hence there is a huge gap between the baseline system achieving **0.6** F1 and human performance.

We plan to use the machine learning techniques taught in this class, especially neural networks, a special advancement of which known as deep learning or deep neural networks have been found to very good at paraphrasing [1]. We also plan to experiment with both supervised and unsupervised forms of learning.

5 Related Work

There exists a panoply of twitter classification research in the academia and industry. Extensive research has already been conducted on Stanford's campus. However, the focus of the existing twitter classifiers do not explicitly exhibit the relational aspect between messages.

5.1 Twitter Sentiment Classification using Distant Supervision

This research focuses on classifying twitter messages based on sentiment. This can be helpful classifying our twitter messages to increase the confidence that two messages are related if they share the same sentiment.
http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

5.2 Global Vectors for Word Representation

This here presents an unsupervised learning algorithm for word vectorization. This project (and related projects) will be a key part of text interpretation and classifying twitter messages.
http://nlp.stanford.edu/projects/glove/

5.3 Google word2vec

This a google tool used for computing continuous distributed representation of words.
https://code.google.com/p/word2vec/

5.4 Build Your Own Twitter Sentiment Analysis Tool

Building a Twitter sentiment analyzer is a popular topic and many baseline tools exist today. This is just one example of the many Twitter classifiers generators that exist.
http://blog.datumbox.com/how-to-build-your-own-twitter-sentiment-analysis-tool/

References

[1] Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection, Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. Advances in Neural Information Processing Systems (NIPS 2011)