#HashtagWars: Learning a Sense of Humor

Language Models in Humor Detection

Xinru Yan & Ted Pedersen
Department of Computer Science University of Minnesota Duluth
yanxx418 & tpederse @d.umn.edu

UMD
UNIVERSITY OF MINNESOTA DULUTH
Driven to Discover℠

Introduction

This research is associated with the SemEval-2017 Task6 #HashtagWars: Learning a Sense of Humor. The task aims to characterize the sense of humour of a particular source, which consists of humorous tweets submitted to a comedy show @*midnight*. Two subtasks are involved:

- Subtask A: Pairwise Comparison – a system should predicts which tweet is funnier for every possible combination of tweet pairs from a given hashtag file.
- Subtask B: Semi-Ranking – a system should produce a ranking of tweets from funniest to least funny given an input file of tweets for a specific hashtag file.

We developed a system called Duluth that participated in the task. The system completed Task A and Task B using ngram language models (LMs), ranking well during evaluation.

Background

1. N-gram models: predict the upcoming word from the previous N-1 words. An N-gram is a contiguous sequence of N words. For example, in tweet "tears in Ramen #SingleLifeIn3Words", "tears" is an unigram; "tears in" is a bigram and "tears in Ramen" is a trigram.
2. Markov assumption: the probability of a word depends only on a small number of previous words. For tri-gram language model, we have:
$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-2}, w_{n-1}) \quad (1)$$
3. Tri-gram LMs: using trigrams to computer the probability of a complete sequence of words. General equation:
$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-2}, w_{k-1}) \quad (2)$$
4. How funny the tweet is compare to "common language": in the study on how phrasing affects memorability, in favor of evaluating how memorable a movie quote is, researchers evaluate its likelihood with the respect of the language models trained on news data [1]. movie quotes that are less like the "common language" are more memorable. The idea of using LMs to assess the memorability of a quote is suitable for our purpose of detecting how humorous a tweet is.

Dataset

Our system estimates tweet probability using ngram models. Except for using tweets provided by the task to train ngram LMs, our system also trained ngram LMs on English news data in order to evaluate how funny a tweet is. Tweets that were more like the tweets model, or less like the news model, were ranked as being more funny.

- The tweets data – provided by the SemEval task ; consists of 106 hashtag files, about 21,580 tokens.

The hashtag: #SingleLifeIn3Words	
tweet	tweet_label
tears in Ramen #SingleLifeIn3Words @midnight	2
Ben and Jerry's #SingleLifeIn3Words @midnight	1
Pet is kid. @midnight #SingleLifeIn3Words	0

[1] http://www.statmt.org/wmt11/translation-task.html#download

- The news data – We collected in total of 6.2 GB of English news data, about 2,002,655 tokens, from the News Commentary Corpus and the News Craw Corpus (2008, 2010 and 2011) [1].

Method

Our system solves the given two tasks in four steps:

1. Corpus preparing and pre-processing: Collect training data files to form one corpus. Pre-processing includes filtering and tokenization.
2. Language model training: Build ngram LMs by feeding the corpus to KenLM Language Model Toolkit [2].
3. Tweet scoring: Get log probability for each tweet based on the trained ngram language model.
4. Tweet prediction: According to the log probability
   - Subtask A – Given two tweets, comparing them and predicting which one is funnier.
   - Subtask B – Given a set of tweets associated with one hashtag, ranking tweets from the funnest to the least funny.

Results

Run # 1 respresents trigram LM trained on the tweets data. Run #2 represents trigram LM trained on the news data.

Task A				Task B			
Rank	**Name**	**Run**	**Accuracy**	**Rank**	**Name**	**Run**	**Distance**
1	djd1283	2	0.675	1	Xinru	2	0.872
2	djd1283	1	0.637	2	sylar	1	0.908
3	cbaziotis	1	0.632	3	xiwu	1	0.924
4	Xinru	2	0.627	4	xiwu	2	0.924
5	acattle	1	0.523	5	Rutal	2	0.938
6	Rutal	1	0.506	6	sylar	2	0.944
7	sylar	1	0.403	7	Rutal	1	0.949
8	Xinru	1	0.397	8	Xinru	1	0.967
9	sylar	2	0.359	9	AlexandraFlescan	1	1.0
10	xiwu	1	0.187	NA	NA	NA	NA

Table 1: SemEval 2017 Task 6 Results

Discussion

- Lack of tweets data could cause the failure on the tweets LMs. We would like to train news models using about as much text as we have for the tweets and see how the results compare.
- To improve the system : collect more tweets that participate in the hashtag wars for tweets models and more news data for news models.
- Try machine learning techniques.

References

[1] Cristian Danescu-Niculescu-Mizil, Justin Cheng, and Lillian Kleinberg, Jonand Lee. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 892–901. Association for Computational Linguistics, 2012.

[2] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August 2013.