Who's to say what's funny? A computer using Language Models and Deep Learning, That's Who!

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

UMD
UNIVERSITY OF MINNESOTA DULUTH
Driven to Discover℠

The Problem

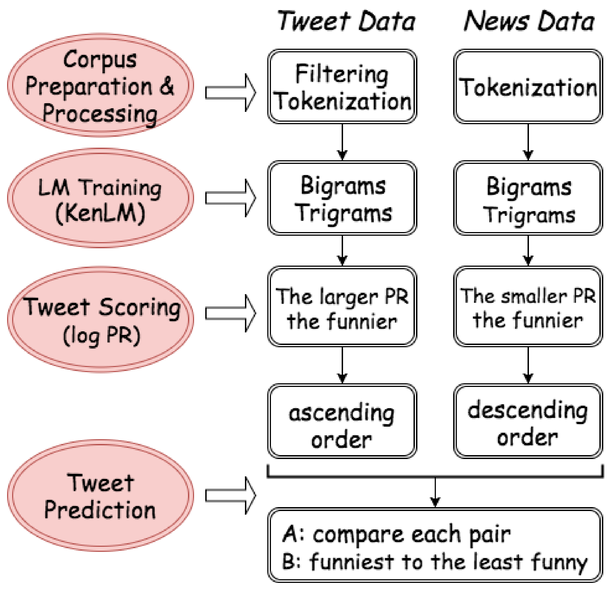
- Traditional humor detection: *binary classification*
  Our focus: learn a *continuous* and *subjective* sense of humor from tweets submitted to the *@midnight* show in response to hashtags by using **Ngram Language Models** (LMs) and **Deep Learning** (DL) methods
- We participated in **SemEval-2017 Task 6** #HashtagWars: Learning a Sense of Humor [1]:
  - Tweets are in three baskets: top most funny tweet, next nine most funny tweets and all remaining
- **Two Subtasks**
  - A: Pairwise Comparison – a system should choose a funnier of two tweets given a hashtag file
  - B: Semi-Ranking – a system should categorize tweets into the right baskets given a hashtag file
- **Dataset**
  - *Tweet Data*: provided by the task, 106 hashtag files, about 21,580 tokens
  - *News Data*: We used 6.2 GB English news, about 2 million tokens [2]

Examples from #BreakUpIn5Words (Trigram LM, News Data)

Tweet	*@midnight*	LM	DL
It's not you, it's meth.	funniest	funny	
Hey, can we NOT talk?	funny	funny	
You need your own Netflix	funny	not funny	?
Figured I'd try being happy.	not funny	funny	
You're a Mac, I'm PC	not funny	not funny	

Language Models

Ngram LMs learn humor from training data and allow ranking by assigning probability for each statement [3][4]



Language Model Results

We seek high accuracy for A and low distance for B.

Dataset	Ngram	Accuracy (A)	Distance (B)
news	**3**	**0.627** (4th)	**0.872** (1st)
news	2	0.624	0.853
tweet	**3**	**0.397** (8th)	**0.967** (8th)
tweet	2	0.406	0.944

- The **type** and the **quantity** of the corpora is what really matters –> *more tweet data, less news data*
- Bigram LMs performed slightly better than trigram LMs –> *Unigram* and *character* level LMs

Deep Learning

- Humor relies on creative use of language which causes too many OOV
  - Jokes often include puns based on invented words
    Barktender #DogJobs
    Tinderella #UpdateAFairyTale
  - Token-level LMs can not understand such puns
  - Character-based CNNs (CharCNN) are not dependent on observing tokens in training data
- Bigram and trigram LMs only use two or three preceding words to predict the next word –> LSTMs are good at making use of sequantial data such as text and are designed for long-term dependencies
  - Some hashtags require tweets to have more than three words and some funny tweets are mostly made up of common bigrams or trigrams
    Complaining makes it better #AmericaIn4Words
    Romantic dinners with the cats #BestWeekendIn5Words
- Ngram LMs do not include external knowledge such as movie titles and song lyrics –> Create word embeddings from domain specific materials
- Our plan: use Keras library to train CharCNN + LSTM LMs on both datasets and investigate ways to include domain knowledge word embeddings in the CharCNN + LSTM LM

References

[1] Peter Potash, Alexey Romanov, and Anna Rumshisky.
SemEval-2017 Task 6: #HashtagWars: learning a sense of humor.
In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, BC, August 2017.

[2] EMNLP 2011 SIXTH WORKSHOP ON STATISTICAL MACHINE TRANSLATION.
`http://www.statmt.org/wmt11/translation-task.html`.

[3] 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*, Sofia, Bulgaria, August 2013.

[4] Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee.
You had me at hello: How phrasing affects memorability.
In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2012.