Joke detection with neural networks

Project Exposé

Miriam Amin

WS 2019/20

1 Introduction

Humor is a fundamental property of humans. Although scholars are analyzing and studying humor since the Ancient Times, until today it is not understood completely. In contrast to other NLP-related problems, the computational treatment of humor is far behind.

Former research in computational humor was mainly carried out on humor generation and humor detection. As I showed in earlier work (Amin, 2019), none of the humor generators presented so far were able to produce human-like humor. From my investigations I concluded two approaches which seemed promising for the advancement of joke generators – a generative and a restrictive approach. A generative approach to humor generation would aim at exclusively producing humorous output by preselecting suitable topics to joke about. A restrictive approach on the other hand would consist of two systems: A system that produces texts with structural features of jokes and a second humor detection system that works as a filter letting only the humorous texts pass. One approach for such a filter would be a neural network for text classification with the target classes `joke` and `no joke`.

The aim of this project is to assess the feasability of current neural network architectures for text classification for the application as such a joke detector. In the following I will briefly present related work and earlier systems for joke detection. I will proceed by outlining the intended method and the data set that will be used for training the neural network.

2 Related Work

Several attempts have been made to classify different forms of humor. Chen and Soo (2018) discriminate English and Chinese one-liners and puns from news text with similar vocabulary. de Oliveira and Rodrigo (2015) detect humor in Yelp reviews. Both groups of authors found that Convolutional Neural Networks (CNN) yielded the best results for accuracy.

3 Methods

The training dataset consists of labeled jokes and non-jokes. The positive jokes examples originate from different websites, crawled by Pungas (2017) and Moudgil (2017).

Compiling a negative data set is more challenging. The documents have to be linguistically similar to jokes, i.e. having a similar word order and vocabulary, but should not be funny. To gather these data, the positive samples will be represented as a suitable feature matrix in order to find similar sentences in the English Wikipedia corpus.

After preprocessing and vectorization of the documents using GloVe word embeddings, the training subset will be fed into a CNN using the python scykit-learn library.

4 Expected Outcome

Given the convincing results of Chen and Soo (2018), who applied a similar strategy of assembling a negative dataset for one-liners and then training a CNN-classifier, I assume that my approach will yield similar good results for multiple-sentence narrative jokes. However, the use of such an assembled negative data set can bias the classification. It is possible, that, instead of learning to discriminate jokes from non-jokes, the CNN learns to distinguish sentences from Wikipedia vs. sentences from other sources. However, such an evaluation of the method can only be made after the experiment has been carried out and the results have been assessed.

References

Amin, M. (2019). *Computational Humor - Automatic Generation of Jokes*. Bachelor Thesis, Leipzig University.

Chen, P.-Y. and Soo, V.-W. (2018). Humor Recognition Using Deep Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 of *Short Paper*, pages 113–117.

de Oliveira, L. and Rodrigo, A. L. (2015). Humor Detection in Yelp reviews.

Moudgil, A. (2017). Python scripts for building 'Short Jokes' dataset, featured on Kaggle. *GitHub repository.*

Pungas, T. (2017). A dataset of English plaintext jokes. *GitHub repository.* https://github.com/taivop/joke-dataset.