Does This Sound like a Joke to You?!

A Computational Approach to Humour Evaluation: The Pursuit of a Scoring Model

Éanna Morley, Ryan Jenkinson, Angus Brayne - NLP Group 2, University College London

Abstract

- The field of computational humour is cross disciplinary, bringing together sociological and psychological theories.
- Determining whether a piece of text is funny is a subjective and nontrivial problem.
- Automatic joke evaluation could be very useful in the training of generative humour models, as well as providing a principled, scalable methodology for the evaluation of models.
- We set out to automate this evaluation process by training a scoring model that predicts Reddit joke "scores".
- Our models outperformed random prediction on a data set often used in the generative joke context.
- We sought to better understand how our models were learning and why they were not achieving better results.

Introduction

- What makes something funny? This question is incredibly complex. The most common theory in the field of computational humour posits "incongruity" as a key mechanism to the success of puns and jokes.
- (Kao et al., 2013) defines incongruity as "perceiving a situation from different viewpoints and finding the resulting interpretations to be incompatible".
- Theories of humour suggest that there are underlying features of comedy that could be learnt by a model in order to score or generate jokes successfully.
- Much of the prior work on generative joke models (Ren and Yang, 2017; Chippada and Saha, 2018) relies on human evaluation. This process greatly impedes progress in the field, as this evaluation methodology is subjective, time consuming and sometimes costly.
- We investigate various models for predicting Reddit "scores" on the "r/Jokes" subreddit, treating the score as a proxy for joke quality.
- An accurate model could be used for a multitude of tasks, such as automatically evaluating the quality of generative joke models or used as a discriminator for these generative models during training.
- Humour has utility in human-computer interaction interfaces e.g chatbots and AI care/personal assistants.

Methods

- Jokes were represented using 300 dimensional pre-trained GloVe embeddings (Pennington et al., 2014). We also experimented with BERT.
- Thresholds were set on the joke scores in order to create 3 classes roughly corresponding to 'Not funny', 'Somewhat funny' and 'Very funny'.
- A number of model architectures were trained and compared; Simple Multi-Layer Perceptrons, LSTM's and CNN's as well as "advanced models" such as (Kim, 2014) and a BERT based classifier.
- tSNE was used for dimensionality reduction to visualize mean GloVe embeddings for each joke.
- We independently rated a sample of jokes to assess the learnability of the Reddit Score (see Figure 4).

Champion Model Architecture

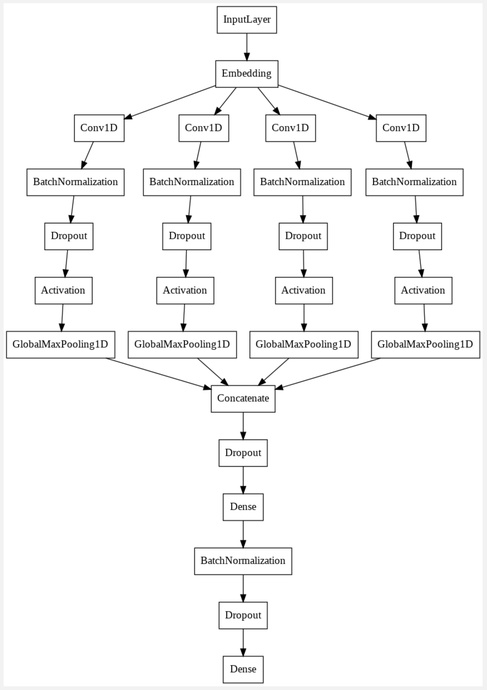

Figure 1: Proposed model architecture for text classification (Kim, 2014) which performed well on the classification task. This model will be tuned to improve performance on our data set. Early results show that BERT may achieve better performance than the above.

t-SNE Visualization

Figure 2: These plots show a 2-D tSNE embedding of the mean GloVe vectors for each joke in the test set. Our baseline model appeared to learn a correlation between joke length and score. Some clusters of conceptually similar jokes have been annotated in the top figure.

Results

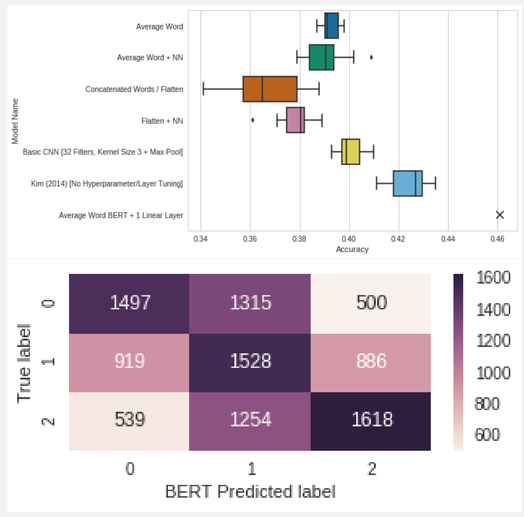

Figure 3: Test accuracy for various models over ten runs. Note: Preliminary results show that BERT Contextual Embeddings will perform best, but this was only one run.

Self Scoring

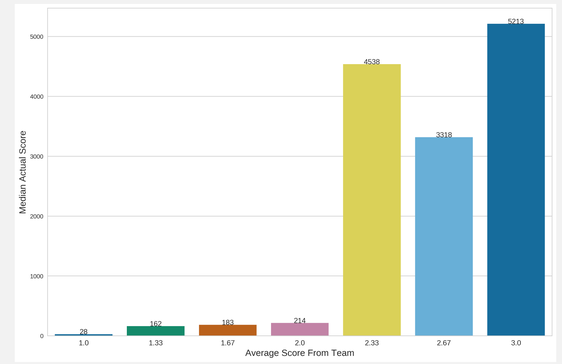

Figure 4: The team's scores of jokes were weakly correlated with actual scores, attesting to the difficulty of our task. These results motivated our score threshold choices.

Conclusion

- The exact method Reddit uses to calculate the score is unknown and thus may not be easily learnable.
- Popular jokes are more likely to be seen by users and therefore attract more upvotes ("rich get richer" effect), thereby creating a heavily skewed distribution of scores.
- The CNN architecture proposed by Kim (2014) performed the best on this data set after hyperparameter tuning.
- The use of contextual word embeddings such as BERT are likely to improve performance further.
- The subjective nature of comedy and its variation across populations make learning humour an especially difficult task. There is more work to be done in this area.

References

Bhargav Chippada and Shubajit Saha. 2018. Knowledge amalgam: Generating jokes and quotes together. *CoRR*, abs/1806.04387.

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2013. The funny thing about incongruity: A computational model of humor in puns. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society, CogSci 2013, Berlin, Germany, July 31 - August 3, 2013*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

He Ren and Quan Sheng Yang. 2017. Neural joke generation.