SemEval-2020 Task 7: Assessing Humor in Edited News Headlines - Subtask 1 (Regression)

Edward Lang
Deep Learning for Natural Language Processing – SS20 – Philipp Cimiano and Philipp Heinisch

August 31, 2021

1 Introduction

We understand humour as something that amuses us, but it is very hard to lay out the precise properties of something that is considered funny. The complexity of this problem comes with not being universally understood e.g., something that may be humorous to one person could not be to another. It can have different dependencies such as culture, language or even taste in humour e.g., "dark humour". The given task (Hossain, Krumm, Gamon, and Kautz, 2020) is to automatically determine the grade of humour given an original and edited headline. The neural network should be able to predict a mean grade on a scale from 0 to 3, whereby 0 is considered as "Not Funny" and 3 is considered as "Funny". Exploring and trying to solve this task may have implications on what makes a headline funny by making a small edit and maybe on humour in general. This mini-paper showcases an existing approach, my approach, an evaluation of my approach and lastly a conclusion.

2 Related work

There are many teams who worked on this specific problem as participants of the competition (Hossain, Krumm, Gamon, and Kautz, 2020). Luo and Tang (Luo and Tang, 2020) build a neural network system primarily consisting of two components. The first part of their model includes a bidirectional LSTM (Hochreiter and Schmidhuber, 1997), a TF-IDF representation (Salton and McGill, 1986) and a vector substraction between the original and the edited word, which represents the input. The second part is a two-layer-feed-forward neural network. The idea is to take advantage of the LSTM architecture which is suitable for processing sequences or in our case sentences. They also reduced the importance of frequently occuring words with the help of the TF-IDF representation and used this as a feature for the neural network. After also calculating the difference between the original and edited headline, all these features are concatenated as one input and fed into the feed-forward neural network, which finally makes the prediction. A visual representation can be seen in Figure 1.
They used the Humiocroedit (Hossain, Krumm, and Gamon, 2019) and FunLines dataset (Hossain, Krumm, Sajed, et al., 2020) for training. They split the whole dataset randomly into 64% training, 16% validation and 20% as test data. At the end of the competition their neural network achieved a score of 0.57237 RMSE and they conducted an 3-pair-subsystem study in order to determine which sub-parts did contribute best (Figure 2).

3 Method

3.1 Data

I exclusively used the Humiocroedit dataset (Hossain, Krumm, and Gamon, 2019), which was already split into 64% training, 16% validation and 20% test data. The input had the following representation: "[Original sentence][SEP][Edited sentence]" and was tokenized for training. I wanted to make sure

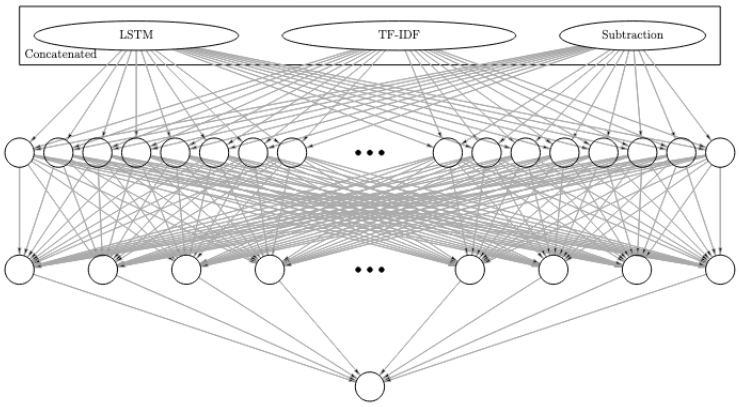

Figure 1: Luo and Tang, 2020

Ablation Subsystem	RMSE
Baseline	0.57469
LSTM with final cell state	0.59105
LSTM with all timeslice hidden state	**0.58611**
LSTM without subtraction	0.58611
LSTM with subtraction	**0.58361**
TF-IDF without subtraction	0.58990
TF-IDF with subtraction	0.58619
Competition System	0.57237
Modified all-together System	**0.56786**

Figure 2: Luo and Tang, 2020

that the neural network would learn the mean grade of the edited headline in dependence of the original sentence, because I think the relation of both sentences is important to consider.

3.2 Model

I used the pretrained BERT (Devlin et al., 2018) model from Hugging Face (https://huggingface.co/) especially build for sequence classification as this task is a regression problem. The model consists of the pretrained main BERT layer, a dropout layer to avoid overfitting and a dense layer. BERT is a transformer pretrained by Google for NLP tasks, which fits nicely to our task. By design it is able to consider the context from both the left and the right sides of each word. There is also no requirement to considerably modify the model in order to fit it to specific NLP tasks. Unfortunately BERT is pretty big and needs a substantial time to train without access to appropriate hardware. The model used adam as the optimizer and MSE for the loss function to learn.

```
Model: "tf_bert_for_sequence_classification"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bert (TFBertMainLayer)       multiple                  108310272
_________________________________________________________________
dropout_37 (Dropout)         multiple                  0
_________________________________________________________________
classifier (Dense)           multiple                  769
=================================================================
Total params: 108,311,041
Trainable params: 108,311,041
Non-trainable params: 0
_________________________________________________________________
```

Figure 3: Model architecture

4 Evaluation

The metric used for evaluation is RootMeanSquaredError. The BERT model accuracy was tested against the provided test data set. In table 1 are results from my model, the model from the related work and others which participated in the competition. BERT did actually well in comparison to other models, which were specifically build and modelled for this task. Evaluating the table, it is clear that Roberta produces the best results by far, almost closely followed by BERT and then the Bi-LSTM models.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

Model	Reference	Accuracy (RMSE)
BERT for Sequence Classification (Hugging Face)	From this paper	**0.5362**
Bi-LSTM (TF-IDF, Substraction)	Luo and Tang, 2020	0.56786
Roberta	Ballapuram, 2020	0.516
Embedding Layer with Bi-LSTM	Miraj and Aono, 2021	0.6164

Table 1: Results (Lower is better)

5 Conclusion

Even though BERT is a general model to tackle NLP tasks it was fairly easy to set up for this specific task and produce good results without substantially modifying the model. One could change the input with focus on more interesting features such as the distance between the original and edited word in hope for better results. Another approach could be changing the surrounding architecture of the model. All in all the results were satisfying enough for me if you take the simplicity of the approach into account and compare them to more thoughful methods.

References

Ballapuram, Pramodith (Dec. 2020). "LMML at SemEval-2020 Task 7: Siamese Transformers for Rating Humor in Edited News Headlines". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, pp. 1026–1032. URL: https://aclanthology.org/2020.semeval-1.134.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Hossain, Nabil, John Krumm, and Michael Gamon (2019). ""President Vows to Cut Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines". In: *CoRR* abs/1906.00274. arXiv: 1906.00274. URL: http://arxiv.org/abs/1906.00274.

Hossain, Nabil, John Krumm, Michael Gamon, and Henry A. Kautz (2020). "SemEval-2020 Task 7: Assessing Humor in Edited News Headlines". In: *CoRR* abs/2008.00304. arXiv: 2008.00304. URL: https://arxiv.org/abs/2008.00304.

Hossain, Nabil, John Krumm, Tanvir Sajed, et al. (Jan. 2020). "Stimulating Creativity with FunLines: A Case Study of Humor Generation in Headlines". In: pp. 256–262. DOI: 10.18653/v1/2020.acl-demos.28.

Luo, Xuefeng and Kuan Tang (Dec. 2020). "Funny3 at SemEval-2020 Task 7: Humor Detection of Edited Headlines with LSTM and TFIDF Neural Network System". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, pp. 1013–1018. URL: https://aclanthology.org/2020.semeval-1.132.

Miraj, Rida and Masaki Aono (2021). "kdehumor at semeval-2020 task 7: a neural network model for detecting funniness in dataset humicroedit". In: *CoRR* abs/2105.05135. arXiv: 2105.05135. URL: https://arxiv.org/abs/2105.05135.

Salton, Gerard and Michael J. McGill (1986). *Introduction to Modern Information Retrieval*. USA: McGraw-Hill, Inc. ISBN: 0070544840.