

How Deep Is Your Humor?

Bartol Freškura, Filip Gulan, Damir Kopljar

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{bartol.freskura, filip.gulan, damir.kopljar}@fer.hr

Abstract

In this paper, we consider the task of comparative humor ranking in two manner: detecting which tweet of two is more humorous and ranking the given tweets by how humorous they are in three classes. We opted for a different approach based on recent deep neural models in order to eschew manual feature engineering. In evaluation section we experimented with the bi-directional LSTM and CNN, in combination and separately. For constructing feature vectors we used pre-trained Twitter *GloVe* word embeddings along with learned character embedding. The system was trained, tuned, and evaluated on the SemEval-2017 Task 6 dataset for which it yields comparative results.

1. Introduction

Understanding humor expressed in text is a challenging Natural Language Problem (NLP) which has not yet been addressed extensively in the current AI research. Humor is often subjective and relies on the vast knowledge base which is sometimes hard to reason even for humans. It is also important to note that what is humorous today might not be humorous tomorrow due to the fact that humor can be trend dependent.

In this paper, we describe a system for humor comparison and ranking. Our system is designed to solve two tasks. For the first task, further referred as Task A, the system is given two tweets and with a goal of predicting the funnier tweet. For the second task, further referred as Task B, the system is given a set of tweets and it should rank them in three categories (2 for the most humorous tweet, 1 if the tweet is in the top ten humorous tweets, and 0 otherwise).

To learn and test our model we used a novel dataset that was created as a part of the SemEval-2017 Task 6 (Potash et al., 2016). The dataset consists of the tweets viewers sent as a part of the Comedy Central show @midnight¹. For every episode, topic was defined and viewers were asked to send humorous tweets on the given topic.

2. Related Work

In the last few years there were many approaches in humor detection (Mihalcea and Strapparava, 2005; Reyes et al., 2013; Zhang and Liu, 2014; Barbieri and Saggion, 2014; Yang et al., 2015). Some of those works (Reyes et al., 2013; Zhang and Liu, 2014; Barbieri and Saggion, 2014) have also acquired humor dataset from Twitter. Most of the related works separate the apprehension of humor into two groups: humor and non-humor, basically a binary classification. This representation ignores the continuous nature of humor, while also not accounting for the subjectivity in perceiving humor (Potash et al., 2016).

SemEval-2017 Task 6 has finished in February 2017 and most of the top-ranked participants (Donahue et al., 2017; Kukovačec et al., 2017; Baziotis et al., 2017) published their work describing used techniques and models.

Donahue et al. with *HumorHawk* team have won the contest on Task A (no data available for Task B). Their system utilizes recurrent deep learning methods with dense embeddings. They used *GloVe* embeddings in combination with a novel phonetic representation which is used as input to LSTM. After LSTM, they have stacked character-based CNN model, and an XGBoost (Chen and Guestrin, 2016) component in an ensemble model which achieves 0.675 accuracy on the evaluation data for the Task A.

Kukovačec et al. with *TakeLab* team have participated in both A and B tasks and have won the second place in both of them. Unlike the other two teams (Donahue et al., 2017; Baziotis et al., 2017) who used deep learning models with automatic feature extraction, they used a manually handcrafted rich set of features: cultural reference features, binary and count-based features, readability-based features, humor-specific features and sentiment-based features. As a classifier they used gradient boosting model which achieves 0.641 accuracy on the evaluation data for the Task A, and a distance of 0.908 on the Task B evaluation data (lower is better with a worst score equal to 1).

Baziotis et al. with *DataStories* team have won the third place on Task A (no data available for Task B). They used a Siamese architecture (Bromley et al., 1993) with bi-directional LSTMs, augmented with an attention mechanism. Their model leverages word embeddings trained on a huge collection of unlabeled Twitter messages. The system achieves 0.632 accuracy on the Task A evaluation data.

3. Architecture

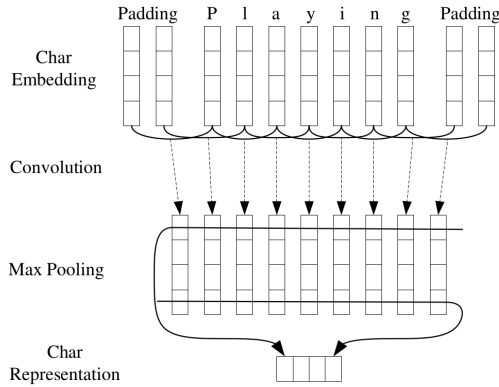
In this section, we describe the architecture of our system. Our most complex architecture consists of the bi-directional *Long Short Term Memory* (Hochreiter and Schmidhuber, 1997) layer, further referred as the Bi-LSTM, a convolutional layer (Lecun et al., 1998), further referred as the CNN, and a fully connected layer with a softmax layer at the end for predicting class probabilities.

The whole pipeline was built using the open source frameworks *TensorFlow*² (Abadi et al., 2015) and *scikit-*

¹<http://www.cc.com/shows/-midnight>

²<https://www.tensorflow.org/>

Figure 1: Word embeddings and the CNN architecture, from (Ma and Hovy, 2016)



*learn*³ (Pedregosa et al., 2011).

3.1. Recurrent Neural Networks

The main idea behind RNNs lies in retaining information from "history". In the context of NLP, history refers to observing the context of the sentence up to the currently processed word. Despite the promising results in short sentences, RNN loses its performance dramatically with the increasing sentence length due to the gradient vanishing (Bengio et al., 1994) and exploding problems (Pascanu et al., 2013). LSTMs were designed with the purpose of correcting the RNNs shortcomings. Although LSTM can successfully capture the past context, it is sometimes good to have an insight at the future sentence context. Bi-LSTMs model implements this by adding an extra LSTM layer which has a reversed information flow meaning that the information is propagated from the end of the sentence towards the beginning. Output of a Bi-LSTM is a concatenated vector of the two opposite LSTM layers.

3.2. Convolutional Neural Networks

CNN networks are famous for their appliance in the *Computer Vision* domain but have also demonstrated an ability to extract morphological information from word characters, encoding them into neural representations. We first create a hash map of all characters that appear in the dataset where values are arbitrarily assigned integer values. All sentence characters are then represented using their mapped integer values but the padding is also applied on the word level as shown in Figure 1. Encoded sentence represents an input which is fed into a trainable character embedding layer of $C_e \times V$ dimensions, where C_e is the character embedding size, and V is the number of unique characters in the corpus.

3.3. Used models

To test how every architecture handles humor detection we have created three models based on the following architectures: CNN, Bi-LSTM, and a combination of the former two.

3.3.1. CNN network

Previous work (Potash et al., 2016) has demonstrated that CNN are top performers when it comes to capturing humor. Our first model consists of the character embedding layer followed by a 1-D CNN layer for extracting features. CNN tweet features are joined and fed into a fully connected layer with a softmax layer for predicting classes.

3.3.2. Bi-LSTM network

Our second model is a pure bi-directional recurrent model with LSTM cells. As input features we use 100 dimensional *GloVe*⁴ (Pennington et al., 2014) vectors trained on the 27B Twitter dataset. Fully connected and softmax layers are also placed at the networks end.

3.3.3. CNN + Bi-LSTM network

To use the best of each of the convolutional and recurrent networks, we construct a merged network consisting of word embeddings, CNN, Bi-LSTM, and a fully connected layer. We hope this network can capture inter word relations using the Bi-LSTM module as well as the inter character features with the CNN layer.

4. Dataset

We trained and evaluated models on the dataset that was created from tweets of viewers who watched TV show @midnight. As part of this *game-show* viewers were asked to write a humorous message about a topic that was announced in the show. The day of the ensuing episode, @midnight would create a post that announced the top-10 tweets from the previous episode. The whole dataset consists of 11,685 tweets about 106 different topics. Every tweet in the dataset is labeled with one of the three labels. Label 2 indicates the winning tweet, label 1 a top-10 tweet, and label 0 is intended for all other tweets. A number of tweets per topic varies among different topics, but 71% of topics contain at least 90 tweets. Topics with the lowest number of tweets have around 20 tweets, and topics with the highest number of tweets have around 180 tweets. It is also important to note that for some topics like *Fast-FoodBooks* external knowledge is required to understand the joke, while for others like *IfIWerePresident* it isn't.

Since given dataset was part of a competition with two tasks, we decided to tackle both of them with single base model. For the task A, we have implemented system which on its input accepts two tweets and predicts which tweet of the given two is more humorous. For the task B we used the same model in voting manner. More specifically, we created all possible pairs of all tweets for given hashtag and fed them to the described system. Then we counted how many times a tweet was more humorous than other tweets, and ranked the tweets by that number.

5. Experiments

We divide this section into two parts: the former one reporting results on the non-evaluation data, and the latter results on the official evaluation data.

³<http://scikit-learn.org/stable/index.html>

⁴<https://nlp.stanford.edu/projects/glove/>

Table 1: Final network hyper-parameters

Hyperparameter	Value
Dropout rate	0.5
BLSTM hidden state	128
Sentence timestep	30
Learning rate	0.0002
CNN filter size	60
FC layer size	256
Character embedding layer size	50

Table 2: 95% confidence scores on the non-evaluation data for all models (in %)

	Baseline	Bi-LSTM	CNN	CNN + Bi-LSTM
Accuracy	50.1 \pm 0.2	67.8 \pm 1.7	52.9 \pm 1.4	67.0 \pm 1.7
Precision	49.8 \pm 0.3	68.2 \pm 1.7	54.1 \pm 1.4	67.4 \pm 1.7
Recall	50.3 \pm 0.2	67.8 \pm 1.7	52.9 \pm 1.4	67.0 \pm 1.7
F1	50.0 \pm 0.2	67.8 \pm 1.7	53.2 \pm 1.4	67.0 \pm 1.7

5.1. Non-evaluation data results

During the model training and evaluation we used a k-fold cross-validation technique ($k = 35$) to properly optimize our models hyper-parameters. Grid search method was not feasible in our case due to large parameters search space and slow training time. Our method was to change each parameter value from the low to high extremes and see how it affects the model performance, in the end interpolating to find near optimal parameters. In addition to using dropout for regularization, we also employed the early-stopping (Caruana et al., 2000) method to achieve the best possible validation performance. Table 1 shows the final hyper-parameters used in training and validation. For the optimization algorithm we used Adam (Kingma and Ba, 2014).

Our model is trained to maximize the class probability $p(y|x)$ using cross-entropy as the loss function. Output 1 means that the first tweet was funnier and 0 otherwise. In our results, we report results in form of four metrics: accuracy, precision, recall and F1 score. All entries represent 95% confidence intervals calculated from the 35 k-fold validation runs. Each of the runs was trained for only one epoch due to severe overfitting problems after the first epoch.

In Table 2 are the results from the non-evaluation data. Baseline model randomly guesses the funnier tweet and is expected to have metrics around 50%. We can see that the Bi-LSTM model performs the best so we further reference it as our best model.

5.2. Evaluation data results

In this section we demonstrate how our best model compares with other solutions on the official evaluation data. Note that the accuracy and distance measurements listed in Table 3 and Table 4 are defined by the task organizers (Potash et al., 2016).

Our top model outperforms the best result in the Task

Table 3: Official Task A results in comparison with our Bi-LSTM model

Team	Accuracy
EuroNeuro	69.1
HumorHawk	67.5
TakeLab	64.1
HumorHawk	63.7
DataStories	63.2
Duluth	62.7

Table 4: Official Task B results in comparison with our Bi-LSTM model (lower is better)

Team	Distance
Duluth	0.872
EuroNeuro	0.881
TakLab	0.908
QUB	0.924
QUB	0.924
SVNIT@SemEval	0.938

A by a margin of 1.6%, placing our model at the top of the list. In Task B, our model ranks second which is still competitive because the best team in Task A doesn't even have top results in Task B, meaning our model can perform well on both tasks.

6. Conclusion

We proposed three different models for solving comparative humor ranking tasks of pairwise comparison and direct ranking classification. All three models use deep learning architecture by combining approaches of recurrent and convolutional neural networks.

For pairwise comparison task best results were achieved using the Bi-LSTM model result in 69.1% accuracy score on unseen evaluation data, and for direct ranking classification task best results were achieved using same Bi-LSTM model and were 0.881 on unseen evaluation data. Model evaluation on final unseen data is done using official evaluation scripts given in SemEval-2017 Task 6.

We have compared our results with the results of other task participants resulting in our model taking the first place on the Task A, and ranking second on the Task B. The main distinction between our model and competitive models is in the lack of hand engineered features which indicates that automatic feature extraction using deep learning framework has a great prospect in this task and requires further work.

For the next step we would experiment with specially adapted word embeddings trained only on the humor containing corpus. We believe it is crucial for word vectors to learn semantic meaning from the domain specific data because of the complex humor structure.

Acknowledgements

We thank TakeLab⁵ team for providing us with the adequate computing resources.

References

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in twitter. In *International Conference on Computational Creativity*, Ljubljana, Slovenia.
- Christos Baziotis, Nikos Pelekis, and Christos Doukeridis. 2017. Datastories at semeval-2017 task 6: Siamese lstm with attention for humorous text comparison. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 381–386, Vancouver, Canada, August. Association for Computational Linguistics.
- Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.
- Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a “siamese” time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems, NIPS’93*, pages 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Rich Caruana, Steve Lawrence, and Lee Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *NIPS*, pages 402–408.
- Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.
- David Donahue, Alexey Romanov, and Anna Rumshisky. 2017. Humorhawk at semeval-2017 task 6: Mixing meaning and sound for humor recognition. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 89–93, Vancouver, Canada, August. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Marin Kukovačec, Juraj Malenica, Ivan Mršić, Antonio Šajatović, Domagoj Alagić, and Jan Šnajder. 2017. Takelab at semeval-2017 task 6: #rankinghumorin4pages. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 387–391, Vancouver, Canada, August. Association for Computational Linguistics.
- Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov.
- Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.
- Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT ’05*, pages 531–538, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Peter Potash, Alexey Romanov, and Anna Rumshisky. 2016. # hashtagwars: Learning a sense of humor. *arXiv preprint arXiv:1612.03216*.
- Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47:239–268.
- Di Yi Yang, Alon Lavie, Chris Dyer, and Eduard H. Hovy. 2015. Humor recognition and humor anchor extraction. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *EMNLP*, pages 2367–2376. The Association for Computational Linguistics.
- Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM ’14*, pages 889–898, New York, NY, USA. ACM.

⁵<http://takelab.fer.hr/>