Proposal- SemEval-2020 Task 7: Assessing Humor in Edited News Headlines

Maisha Binte Moin
bintemoi@ualberta.ca

Uduak Ituen
ituen@ualberta.ca

Abstract

This document contains the details of the project proposal submitted for CMPUT501, Fall 2021.

1 Task

The task embarked on for this project is obtained from the SemEval-2020 task-7, titled "Assessing Humor in Edited News Headlines". This task provides a dataset rated by crowdsourcing which contains edited versions of news headlines to make them funny. The project is divided into two subtasks. The first is a “Regression” task aiming to estimate the degree of humor on a scale of 0-3 (0 - Not Funny; 1 - Slightly Funny; 2 - Moderately Funny; 3 - Funny). Second is a “Classification” task to predict the funnier version between the original and edited news headline.

Humor is a complex form of human language expression. It is an essential component needed for a computer to communicate effectively with humans (Binsted et al., 2006). By modelling humor generation in computer systems, we are able to better understand how humans process language and cognition (Binsted et al., 2006). Although steady progress has been made by researchers in training intelligent systems to express humor, it has not gained as much traction as it should, partly due to the complexities involved in emulating the intrinsic nature of Humor (Hossain et al., 2019).

Two metrics would be employed for evaluation of the tasks, first is the Root Mean Squared Error (RMSE) for Subtask 1 and Accuracy for Subtask 2. RMSE measures the difference between the rating generated by the system model and the benchmark rating given by crowdsourcing. While Accuracy would be adopted to evaluate the correctness of the model’s prediction in determining the funnier headline.

2 State-of-the-Art

Computational humor can be divided into two groups: recognition and generation (Docekal et al., 2020). Both humor recognition and generation prove to be challenging problems as they involve in-depth world-knowledge, common sense and various levels of understanding (Hossain et al., 2019). While earlier humor recognition systems used statistical machine learning algorithms such as support vector machine (SVMs), decision tree and Naive Bayes (Jin et al., 2020), nowadays given the advancement of deep learning technologies we can see both the utilization and success of such sophisticated models in achieving state of the art results.

For the SemEval 2020 task 7, the leading teams made use of pre-trained language models (PLMs) such as BERT, RoBERTa, ELMo, GPT-2 (Hossain et al., 2020). The winning team for both tasks, Hitachi (Morishita et al., 2020), formulated the problem as a sentence-pair regression and used the results generated by their model from Subtask 1 for the prediction problem defined in Subtask 2. A notable general trend based on the performance of the submitted models is the importance of contextual information in humor recognition tasks for achieving better results (Hossain et al., 2020).

3 Available Data

The dataset for this task contains approximately 5,000 news headlines obtained from Humicroedit (Hossain et al., 2019) and 8,248 training data obtained from the FunLines competition (Hossain et al., 2020).

The zip file downloaded from the proceeding contains two folders for the subtasks, each

containing train, test, train_funlines and dev CSV files, which would serve as the input data. The file also contains a baseline output CSV file for metric measurements.

Below is a sample expected input and output from the model for the two subtasks to be designed.

Original Headline	Edited	Rating (subtask 1)	Estimation (Subtask 2)
4 arrested in Sydney raids to stop terrorist **attack**	4 arrested in Sydney raids to stop terrorist **kangaroo**	2.6	1.06
CNN's Jake Tapper to **interview** Paul Ryan following retirement announcement	CNN's Jake Tapper to **wrestle** Paul Ryan following retirement announcement	2.8	1.17

Table 1: Sample extracted from the original proceeding (Hossain et al., 2019) showing original headline, edited version, rating score (0-3) scale and estimated level of funniness.

4 Available Code

For this project, we plan to employ both statistical machine learning and deep learning approaches to solve the tasks. The statistical machine learning approaches we plan on using are support vector machines, decision tree, k-NN and Naïve Bayes. For the deep learning approach, we plan to implement a convolutional neural network architecture basing on the submission by Docekal et al., 2020 and a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) neural network with attention mechanism (Bahdanau et al., 2014) basing on the submission by Zhang et al., 2020. These approaches have been chosen as they have been previously employed and it will enable us to do a comparative study with previous results.

Our steps will be as follows: preprocessing the data, implementing the models, training and testing the models using the provided train and test datasets, hyperparameter tuning for the deep learning models and finally conducting a comparative analysis of the results based on the models performance for the tasks. The metric that will be used to determine the performance of the models is RMSE for Subtask 1 and Accuracy for Subtask 2. The results will be evaluated using the official scripts that are available in the associated GitHub repository for the task.

Ongoing review is being done on past submissions to evaluate the code used.

Repository:

Task: https://github.com/n-hossain/semeval-2020-task-7-humicroedit

Project: https://github.com/UOFA-INTRO-NLP-F21/f2021-proj-udituen

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Kim Binsted, Anton Nijholt, Oliviero Stock, Carlo Strapparava, G Ritchie, R Manurung, H Pain, Annalu Waller, and D O'Mara. Computational humor. IEEE Intelligent Systems, 21(2):59–69, 2006.

Martin Docekal, Martin Fajcik, Josef Jon, and Pavel Smrz. Jokemeter at semeval-2020 task 7: Convolutional humor. arXiv preprint arXiv:2008.11053, 2020.

Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long short-term memory. Neural Comput., 9(8):1735–1780, November.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In Proceedings of the 2019 Conference of the NorthAmerican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 133–142, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020a. Semeval-2020 Task 7: Assessing humor in edited news headlines. In Proceedings of International Workshop on Semantic Evaluation (SemEval-2020), Barcelona, Spain.

Shuning Jin, Yue Yin, Xiane Tang, and Ted Pedersen. Duluth at semeval-2020 task 7: Using surprise as a key to unlock humorous headlines. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 986–994, 2020.

Terufumi Morishita, Gaku Morio, Hiroaki Ozaki, and Toshinori Miyoshi. Hitachi at semeval2020 task 7: Stacking at scale with heterogeneous language models for humor recognition. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 791–803, 2020.

Tiantian Zhang, Zhixuan Chen, and Man Lan. Ecnu at semeval-2020 task 7: Assessing humor inedited news headlines using bilstm with attention. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 995–1000, 2020