SO at SemEval-2020 Task 7: DeepPavlov Logistic Regression with BERT Embeddings vs SVR at Funniness Evaluation

Anita Soloveva
Lomonosov MSU
nit-sol@mail.ru

Abstract

This paper describes my efforts in evaluating how editing news headlines can make them funnier within the frames of SemEval 2020 Task 7. I participated in both of the sub-tasks: Sub-Task 1 "Regression" and Sub-task 2 "Predict the funnier of the two edited versions of an original headline". I experimented with a number of different models, but ended up using DeepPavlov logistic regression (LR) with BERT English cased embeddings for the first sub-task and support vector regression model (SVR) for the second. RMSE score obtained for the first task was 0.65099 and accuracy for the second – 0.32915.

1 Introduction

Humor is inherent in human beings, but difficult to be detected by the machines. Therefore, a challenge of automatic humor recognition and analysis, based on data of different genre and language, has lately received a great amount of attention. Specifically, several shared tasks were organized within the frames of evaluation workshops, i.e. SemEval-2017 Task 6 "Learning a Sense of Humor", aimed to analyze humorous responses submitted to a Comedy Central TV show @midnight in English (Potash et al., 2017), HAHA task at IberEval 2018 with the sub-tasks of automatic detection and rating of humor in Spanish tweets (Castro et al., 2018) and etc.

SemEval-2020 Task 7, however, presented a slightly different type of challenge, namely, an attempt to investigate how small edits can turn a text from non-funny to funny. Sub-task 1 was to predict the mean funniness of the edited headline and sub-task 2 was intended to predict which of the two given edited versions of the original headline was the funnier or were they equal. All the data were in English, for more details about the dataset, see (Hossain et al., 2019), about the task itself, see the Task paper (Hossain et al., 2020).

The aim of this paper is two-fold: describing my approaches for both sub-tasks and analyzing the obtained results. In this study, I experimented with two models: for the first sub-task I used a relatively new one – DeepPavlov logistic regression with BERT English cased embeddings (Burtsev et al., 2018) and for the second task I chose a well-known one – SVM version for regression, namely SVR (Drucker et al., 2003), with word n-gram features. In sub-task 1 I obtained RMSE equal to 0.65099 and in sub-task 2 accuracy was 0.32915. I also tried to use DeepPavlov BERT-based model during the post-evaluation period, which perfomed better than two previously-mentioned ones. My repository can be found on github https://github.com/aniton/SO_SemEval-2020_News_Headlines.

2 Background

In this section, I would briefly introduce the input and the output sets in order to form a better idea of the sub-tasks.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1055–1059
Barcelona, Spain (Online), December 12, 2020.

2.1 Sub-task 1

The training set consisted of 9652 news headlines modified using short edits. The aim of the challenge was to assign a funniness grade to 3024 headlines from the test set in the [0,3] interval. Systems were ranked by RMSE. An example of a headline from the training set is the following:

Original	Substitute	Grade
Trump wants to make Wall Street great again	fail	2.0

2.2 Sub-task 2

Second sub-task was to predict which of the two given edited versions of one headline was the funnier. Training and test sets consisted of 9381 and 2960 non-edited headlines respectively. Systems were ranked by prediction accuracy. One typical example of the input would be the following (the second version is the funnier):

Orig.	Sub. 1	Gr. 1	Orig.	Sub. 2	Gr. 2	Label
Trump wants to make Wall Street great again	asphalt	1.8	same	fail	2.0	2

3 System description

DeepPavlov (Burtsev et al., 2018) as an open source framework based on TensorFlow (Abadi et al., 2015) and Keras (Chollet, 2015) has lately gained much attention among the developers. It contains three available models, namely, BERT, Keras and Sklearn. For sub-task 1 I used Sklearn Logistic Regression. The parameters for this system were chosen as following: C = 1, solver = 'lbfgs'.

DeepPavlov also trained various word and sentence multilingual BERT-based embeddings. As the embeddings features have a great impact on system performance, for the first sub-task I decided to employ TF-IDF weighted 100-dimensional BERT English cased embeddings.

For sub-task 2 I chose RBF kernel support vector regression (SVR). I used word n-grams as features, maximum word n-gram size was set to 5 and the parameter C to .1. This time I did not use any pre-trained word embeddings.

As data preprocessing has a great impact on system performance, I applied it to the news headlines in both sub-tasks (see Section 3.1). I also made use of 14 billion word iWeb corpus (in particular, I used a list of the most frequent English bigrams) (see Section 3.2). I did not use any external sources except for the mentioned one.

3.1 Preprocessing steps

The preprocessing pipeline included the following basic steps:

- Removing ids of the headlines
- Removing all the following charachters " :. ,— ˜ ", digits and single quotation marks
- Making substitutes

3.2 14 billion word iWeb corpus

Since the news headlines are short and consists of 10 to 30 tokens, the humour is usually produced not due to the given context, but due to some background knowledge of assessors. Thus, the more famous the situation, the statement or the person is, about which/whom the joke is, the higher score of funniness it gets. This is also confirmed by some researches on the evaluation of humor in short texts (Boylan, 2018), (Braslavski et al., 2018). According to these articles, the joke receives a high score, when it is equally understood by native/non-native speakers, people of different age and gender. One possible strategy to achieve this, is to use some phrases, collocations, associated with the event or celebrity, which are easily

recognized by people of different groups. Therefore, I tried to make use of the list of the most frequent English bigrams from 14 billion word iWeb corpus.

In both sub-tasks, we used a binary feature as input for the systems, in particular '1' in the situation, when a substitute word with its left/right context was in an above-mentioned list, and '0', when it was not.

3.3 System pipeline for sub-task 1

Here, I present a description of the system architecture for the first sub-task covering preprocessing steps, features used and the model employed (see Figure 1).

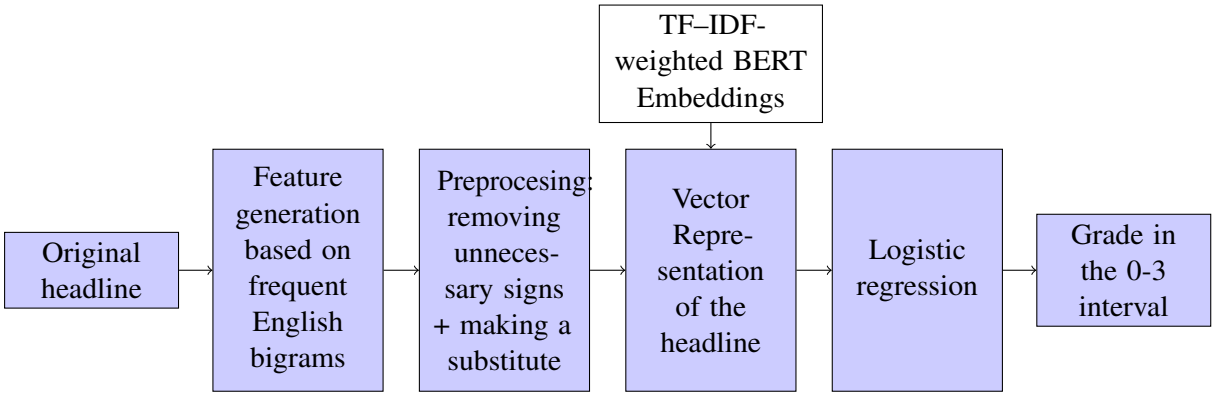

Figure 1: Pipeline of sub-task 1

3.4 System pipeline for sub-task 2

The system architecture for the second sub-task is illustrated in Figure 2.

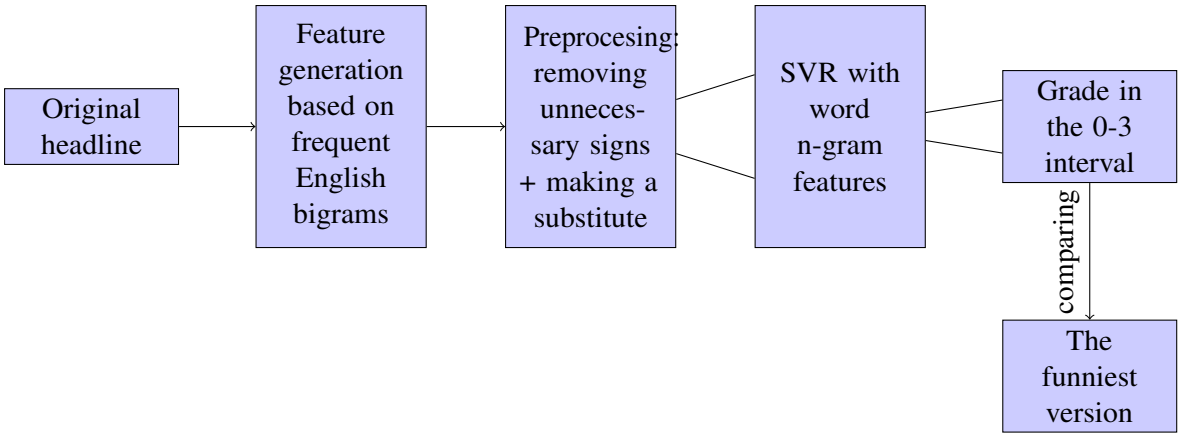

Figure 2: Pipeline of sub-task 2

In the frames of this task I firstly predicted the grade $\in$ [0,3] for each of the two versions of the original headline and then compared them. The output was '0', when both headlines had the same funniness, '1', when the first edited headline was the funnier, '2', when the second one was the funnier.

4 Results and Analysis

In both sub-tasks I tried to experience with traditional machine learning approaches, namely LR and SVR. Below, I present and discuss the results for each sub-task.

4.1 Sub-task 1

The official competition metric to evaluate the systems in this sub-task was RMSE (Root-Mean-Square Deviation). My system achieved 0.65099 score. In the post-evaluation period, I also tested the second model in this sub-task. It performed slightly better: 0.63252. First, this can be explained by the fact that logistic regression performs better in binary/multiclass classification tasks, not in regression ones. Second, in spite of the applied regularization, the model might have suffered from overfitting: the maximum and mean true grades are 2.8 and .93 respectively, in comparison with the maximum and mean grades, predicted by LR: 1.8 and .77.

4.2 Sub-task 2

In the second sub-task participating teams were ranked by the accuracy. This time my system obtained 0.32915 score. The class '0' (with equal grades for both versions) had the least quantity of true positives (TP) (PPV = .11, see Figure 3). Other positive and negative predictive values were more than .5. This could happen, since I reduced the parameter 'C' to .1 and predicted grades were too diverse, contrary to the situation in sub-task 1 with parameter 'C' of 1.

Figure 3: Positive and negative predictive values

	Class: 0	Class: 1	Class: 2
Pos Pred Value	0.11444	0.5156	0.5278
Neg Pred Value	0.88963	0.5959	0.5769

5 Post-evaluation experiments

In this section, I will discuss the findings of the post-evaluation experiments. This time, to deal with the first sub-task, I tried to test DeepPavlov multilingual-cased BERT-based model, since BERT models (Devlin et al., 2018) has recently demonstrated excellent performance in various NLP tasks. I included all previously mentioned preprocessing steps and added a binary feature, based on iWeb corpus. The specific setting was the following: the batch size of 256, the maximum sequence length of 64, the learning rate of .5. I trained the model for 3 epochs. This system performed better than two others: it achieved RMSE of .5529.

6 Conclusion and Future directions

In this paper I presented the contribution of SO team to SemEval-2020 Task 7. During the evaluation period I experienced with traditional machine learning algorithms, namely LR and SVR, in order to evaluate the funniness of the edited news headlines. Despite the fact that I used pre-trained BERT embeddings as input features to LR model, during the post-evaluation period I discovered that BERT-based model achieves much better results in the given regression task.

While analyzing the test dataset, I also noticed that edited news headlines had top grades if they contained names of some famous people/events regardless of the substitute. For instance, there were a great amount of headlines about Trump, which were highly evaluated. Thus, in the future I could try to make use of the different lists of popular political and media figures, since they are a target of admiration, jokes and hate. A list of representatives of top twitter profiles from different countries (`https://www.socialbakers.com/statistics/twitter/profiles`) could serve as an example of such lists. In (Bansal et al., 2019) this source was also used within the frames of SemEval-2019 Task 6 (OffensEval), see (Zampieri et al., 2019), in order to predict whether the tweet was aggressive or not and if yes, who or what was the target of aggression.

7 Acknowledgments

This work was supported by MSU Development Program, School of Artificial Intelligence Technologies.

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 Mane, 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 Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Himanshu Bansal, Daniel Nagel, and Anita Soloveva. 2019. Had-tübingen at semeval-2019 task 6: Deep learning analysis of offensive language on twitter: Identification and categorization. In *SemEval@NAACL-HLT*.

James R. Boylan. 2018. The cognitive psychology of humour in written puns.

Pavel Braslavski, Valeria Bolotova, Vladislav Blinov, and Katya Pertsova. 2018. How to evaluate humorous response generation, seriously? In *CHIIR 2018 - Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*, volume 2018-March, pages 225–228, United States, 2. Association for Computing Machinery (ACM).

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia, July. Association for Computational Linguistics.

Santiago Castro, Luis Chiruzzo, and Aiala Rosa. 2018. Overview of the haha task: Humor analysis based on human annotation at ibereval 2018. 09.

François Chollet. 2015. Keras. `https://keras.io`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Harris Drucker, Chris C, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 2003. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9, 11.

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 North American 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. 2020. Semeval-2020 Task 7: Assessing humor in edited news headlines. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada, August. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.