Assessing Humor in Edited News Headlines

Vinh Ngu
2ngu@inf...

Finn Rietz
5rietz@inf...

Abstract

In this paper we will present our results and methodology in the context of a Competition "**SemEval-2020 Task 7: Assessing Humor in Edited News Headlines**". In total, we have developed three different approaches based on different assumptions and interpretations of the provided data set. We achieved an accuracy of **24%**. Given the fact that there are 31 possibilities and thus an expected value of 3%, we have achieved a **seven times greater** accuracy than blind guessing.

1 Introduction

For the applied project part of the Deep Learning Seminar, we took part in the SemEval2020 competition titled "Assessing the Funniness of Edited News Headlines". The concrete task, as formulated in the competition, is to evaluate the *funniness* of an atomic edit made to a newspaper headline. Thus, the high-level goal of this challenge is to gain a deeper understanding of general indicators that make for a funny, simple change to a given (short) text like a newspaper headline.

Generally, humor-related datasets are usually annotated in a categorical fashion (e.g. funny vs not funny), thus, the intensity of the humor can not be accessed and ranking in terms of funniness is not possible. Specifically, in terms of atomic edits, the ability to rank them according to funniness is desirable for various applications, including humor-generation scenarios (Hossain et al., 2019).

Motivated by this, the competition is formulated in terms of a regression problem, where participants need to develop a system that can predict a regression score for a given edit. The organizers offer an additional, second task, that is only considered with the direct comparison of two edits, where the funnier of the two must be determined. However, as this problem can be solved implicitly by solving the regression problem of predicting the funniness score of an edit, we set our focus on the regression task.

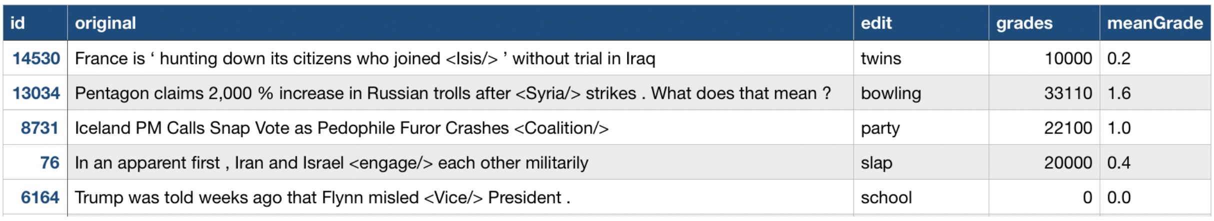
2 Dataset

The underlying dataset for the competition is the *Humicroedit* dataset as presented by Hossain, Krumm and Gamon (Hossain et al., 2019). The dataset contains 15,095 edited headlines. One item in the dataset consists of the original headline, an indicator for which word will be edited or changed, and the proposed edit. Each item has been rated in terms of its funniness by five jurors, that could assign the edit one of the following ratings: **0**: Not funny, **1**: Slightly funny, **2**: Moderately funny and **3**: Funny.

Based on the five scores, a mean grade for the edit is determined. This mean grade operates as the primary measure of interest, as the individual submissions to the competition are ranked according to the *root mean squared error* (RMSE) between the predicted funniness grade and the ground truth mean grade.

2.1 Artificial data addition

When investigating the tasks in greater detail and brainstorming for possible approaches, it became quickly apparent that there is a significant and informative relationship between the unedited and edited versions of the headline. While there are numerous possibilities to model this relationship, we tried to incorporate this by adding both versions of the headline to our dataset. We accomplished this by adding the unedited version of each headline with a ground truth score of zero for the funniness grade. Thus, our *adjusted* version of the Humicroedit dataset contains two items for every original one, where every item is a tuple of the headline and the assigned funniness grade.



id	original	edit	grades	meanGrade
14530	France is ‘ hunting down its citizens who joined <Isis/> ’ without trial in Iraq	twins	10000	0.2
13034	Pentagon claims 2,000 % increase in Russian trolls after <Syria/> strikes . What does that mean ?	bowling	33110	1.6
8731	Iceland PM Calls Snap Vote as Pedophile Furor Crashes <Coalition/>	party	22100	1.0
76	In an apparent first , Iran and Israel <engage/> each other militarily	slap	20000	0.4
6164	Trump was told weeks ago that Flynn misled <Vice/> President .	school	0	0.0

Figure 1: A screenshot of the first 5 datapoints of the published training-dataset. Each line contains a sentence where a word is marked. If you replace this marked word with the word from the column 'edit' the sentence gets a score of funniness (meanGrade).

While the adjusted version of the dataset contains a baseline to compare each headline against, it generally has the disadvantage of being largely biased towards a mean funniness grade of zero, as 50% of the items in the dataset have a mean funniness grade of zero, with the other 50% consisting all the other funniness grade present in the original dataset. To account for this, we apply each developed method on both versions of the dataset and report results accordingly.

3 Data Preprocessing

The dataset provided by the competition is not readable on default by the model. Hence, we needed to prepare it first. Given the sentence with the id “8713” 1 “In an apparent first, Iran and Israel **xxxx** each other military” we have unfold the sentence into two phrases.

1. The first sentence is composed of the original word of interest “engage” and the second one is made of the word “slap”. In approach 1 und 2 we have added the score **0.0** and **0.4** for sentence 1 and sentence 2 described above. We have assumed, original sentences are facts and not funny at all, so we scored these with 0.0 level of funniness.

2. Afterwards, we have applied stemming and lemmatizing algorithms on the dataset to remove redundant words such as “eat”, “ate” and “eaten” would be reduced to “eat”.

3. We embedded the words by the frequency of their occurrence within the text corpus in order to make them machine-readable.

4. We have left-padded the sequences by zeros to ensure a consistently defined length of 25 integers.

5. Finally, we normalized the values to a range between 0 and 1.

4 Approach

There are numerous possibilities for modeling the relationship between the edited and unedited versions of each item. Thus, we tried to model this relationship with a total of three distinct approached, where approaches one and two have been applied to both the original and adjusted version of the Humicroedit dataset.

4.1 Model architecture

Independently of the approach, we employed a simplistic fully connected model with dropout regularization. Based on a simplistic hyperparameter optimization we identified adequate neuron numbers and for the hidden layers, which vary slightly across the different approaches. The main difference in the model for the approaches manifests in the output layer, which is adjusted according to each approach, as described in greater detail in the following sections.

4.2 Approach 1

This approach aims at learning the funniness of headlines in general instead of the direct edit. Thus, we try to predict the mean funniness grade of each headline and treat the task as a regression problem. Accordingly, the general model is adjusted to output a floating-point number to represent the mean funniness grade of the headline. Each data-point results in two tuples, according to the steps described in the data preprocessing section 3. An exemplary representation is given below:

$$[0, 0, 0.83, 0.4, 0.6, 0.13, 0.24] \rightarrow [0.24] \qquad (1)$$

Based on the previously reported upon generation of the adjusted dataset, we applied this approach to both versions of the dataset. The results of this are provided in table 1.

4.3 Approach 2

For our second approach, we reformulated the problem from a regression problem towards a classification problem. Here, we discretized the output space into 31 discrete bins, where the bins correspond to the discretized output space of the mean funniness grade, which ranges from 0.0 to 3.0. Thus each bin has a width 0.1 of the previously continuous output space. A training sample could have the following shape:

$$[0, 0, 0.83, 0.4, 0.6, 0.13, 0.24] \rightarrow [0_0, 1_1, ..., 0_{30}] \tag{2}$$

Accordingly, we adapted our model to this kind of problem. Concretely, our output is now a 31-dimensional vector, with a softmax activation function, and the bin with the highest probability is accepted as final prediction.

There are additional possibilities for a more sophisticated activation function. For example, it could be appropriate to report the mean of the activations in the final layer instead of the softmax, depending on the variance of the activations in the final layer. However, further testing would be required to confirm that this yields a reduction of loss.

Again, as for the first approach, we applied this approach to both versions of the dataset and report results in table 1

4.4 Approach 3

Finally, for our third and last approach, we chose a joint vector representation to model the relationship between the unedited and the edited version of the headline. Concretely, this means that we try to predict the binned probability as in approach 2, but the input to the model is the embedding of the concatenation of the edited and unedited version of the headline.

In this setting, where we represent the relationship between original and edited headline directly in each item, we don't need to store the baseline headlines with a mean funniness grade of zero. Thus, as indicated in table 1, we only apply the algorithm on the joint version of the Humicroedit dataset. Accordingly, a sample can have the following shape

$$\begin{aligned} &[0, 0, 0.83, 0.4, 0.6, 0.13, 0.24, \\ &0, 0, 0.83, 0.21, 0.6, 0.13, 0.24]] \rightarrow [0_0, 1_1, ..., 0_{30}] \end{aligned} \tag{3}$$

We assume that this data representation yields a stronger encoding of the relationship between the edited and unedited version compared to approach 1 and 2.

4.5 Training

Our final approach 3 promises the best results so far among the three introduced approaches. Hence, this section focus on approach 3. As stated, we have not changed the hyperparameters within the 3 approaches. We used the Adam-Optimizer with the learning-rate of 0.001. After 100 epochs we have achieved a rmse loss of **1.4754** and an accuracy of **0.238** on the training-set.

4.5.1 Model

The models input-layer is made of 256 units followed by a hidden-layer composed of 512 units. Both layers are using the activation-function "relu". The input dimension of the input-layer is twice as long as the max embedding length (25*2). Accordingly to the number of classes the output-layer is equipped with 31 units combined with the "softmax" activation-function.

Dataset:	Original	Adjusted	Joint
	Acc | Loss	Acc | Loss	Acc | Loss
Approach 1	0.04 | **0.20**	0.52 | **0.20**	X
Approach 2	0.10 | 1.61	0.51 | 1.38	X
Approach 3	X	X	**0.11** | 1.62

Table 1: Results obtained from the three different approaches on the different versions of the dataset. Metrics were calculated on the test-set.

4.6 Results

Generally, approach three indicates the most promising results, in terms of accuracy. Here, accuracy translates to the prediction of the right bin of discretized mean funniness grade. This indicates that the direct modeling of the relation between the two versions of the headline with the joint vector representation is most potent for this kind of problem. However, this approach also produces a relatively high root mean squared error. This is to be considered for the specifically for the SemEval 2020 challenge where this task originates from because the key criterion there is the RMSE loss over the entire test data set.

Further, from table 1 we can observe a surprisingly low RMSE loss for approach 1, paired with a relatively high accuracy of 0.5 on the adjusted dataset and very low accuracy of 0.04 on the original dataset. We currently have no explanation for this low error in combination with poor accuracy on the original dataset and assume some unintended mechanism in the way Keras calculates the accuracy metric for the regression problem.

The high accuracy scores for approach 1 and 2 on the adjusted dataset can be explained by the significant imbalance of classes in the dataset, as describe in section 2.1.

Even though our obtained accuracy and RMSE loss scores (as reported in table 1) don't indicate strong results, we uploaded our results in the requested form to the SemEval 2020 and are at the time of writing ranked 82 of 85 with a RMSE of 0.72 (for comparison, rank 1 currently achieves a RMSE of 0.51) (ran). This indicates that our results can not compete with the best models that currently take part in the competition, but also that the tasks not fully solved, with even the best performing models achieving a high RMSE of 0.5 (on a scale of 0-3).

5 Conclusion

For the task of predicting the funniness associated with atomic edits, we implemented three distinct approaches. We employed a simplistic fully connected model with dropout and only slightly varied the architecture for each of the approaches. We created multiple data representations in an attempt to model the relationship between the unedited and edited versions of the headlines. The scores of approach three, which we are most confident in, still perform four times better than random guessing (3% accuracy random vs 11% accuracy obtained) on the test-set and seven times better on the training set (3% accuracy random vs 24% accuracy obtained). Our ranking on the development set of the SemEval 2020 competition indicates that a more sophisticated data representation and or model are needed, while even the best performing competitors don't achieve scores largely different to ours, which indicates how challenging it still is to assess humor through machine learning methods.

References

SemEval 2020 competition rankings. `https://competitions.codalab.org/competitions/20970#results`. Accessed: 2020-02-27.

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. Association for Computational Linguistics.