Amobee at SemEval-2020 Task 7: Regularization of Language Model Based Classifiers

Alon Rozental **Dadi Biton** **Ido Blank**
Amobee Inc., Tel Aviv, Israel
alon.rozental,dadi.biton,ido.blank@amobee.com

Abstract

This paper describes Amobee's participation in SemEval-2020 task 7: "Assessing Humor in Edited News Headlines", sub-tasks 1 and 2. The goal of this task was to estimate the funniness of human modified news headlines. In this paper we present methods to fine-tune and ensemble various language models (LM) based classifiers for this task. This technique used for both sub-tasks and reached the second place (out of 49) in sub-tasks 1 with RMSE score of 0.5, and the second (out of 32) place in sub-task 2 with accuracy of 66% without using any additional data except the official training set.

1 Introduction

While most of the studies are interested in whether a chunk of text is funny, this study examined whether an atomic change can transform a non-funny headline into a funny one. The study focused on two sub-tasks. The first task is to the determine how funny the micro change in the headline is. The second task is to choose between two micro changes of the original headline, which one is funnier. This task is part of SemEval 2020 workshop (Hossain et al., 2020). There were 49 groups in the first sub-task and 32 groups in the second.

Contextual word embeddings is currently one of the most popular ways to create a numeric representation of a document. It is capable of capturing the context of a word in a document, semantic and syntactic similarity, relation with other words, etc. Pre-trained models of this kind, such as (Liu et al., 2019; Yang et al., 2019; Devlin et al., 2018), can be fine-tuned to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. However, due to the fact that these models usually have a large number of parameters, over 300 million for the models mentioned above, they tend to overfit very quickly when used without any modification on small datasets. For example, BERT-large (Devlin et al., 2018) has over 30,000 times more parameters than the SemEval-2020-task-7-dataset has training examples. In this paper, we present a way to improve the performance of systems that are based upon language models (LM) or contextual token representation models by adding an L2-regularization term on the difference between the pre-trained LM and the fine-tuned model. For convenience, we'll use the term LM in this paper despite the fact some of the models we refer to are word representation models and not proper language models.

The paper is organized as follows: Section 2 describes the background for this task and an explanation of the dataset. Section 3 describes our system architecture and presents the L2-regularization loss for training the language model and our ensemble method. Section 4 describes our experimental setup, and Section 5 summarizes our results.

2 Background

The "Assessing Humor in Edited News Headlines" task, shared by SemEval 2020, consisted of two sub-tasks. The first sub-task was a regression problem in which we needed to predict how funny a micro change in news headline is. Micro change was defined as one of the following replacements: entity→noun,

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 981–985
Barcelona, Spain (Online), December 12, 2020.

Original Headline	Substitute	Grade
Kushner to visit **Mexico** following latest Trump tirades	therapist	2.8
4 **soldiers** killed in Nagorno-Karabakh fighting: Officials	rabbits	0.0

Table 1: Sample data-points from the competition data-set.

noun→noun and verb→verb. Each edited headline was scored by five judges, each of whom assigned a grade from 0 (not funny) to 3 (funny). The funniness level of each headline is the mean of its five funniness grades. Example of sample from the data is presented in Table 1. The second sub-task was a classification task of predicting the funnier of two edited headlines, which were derived from the same original headline. The sentences in the dataset were in English. In the first sub-task there were 12,071 samples in the training phase and 3,024 samples in the test set. In the second sub-task there were 11,736 samples in the training phase and 2,960 samples in the test set. As we will describe below, we didn't train a model for this sub-task, hence we used the training data just for validation.

3 System Overview

3.1 Architecture

The architecture of our system is made of two components. The first component is a pre-trained text representation model which is one of the following 3 models: BERT, XL-NET or Roberta (Liu et al., 2019; Yang et al., 2019; Devlin et al., 2018). These models yield both a representation of the tokens in the texts and a representation of the entire text (a CLS-token). The second component of our architecture is a classifier unit that receives the following inputs:

1. The CLS-token of the unmodified text (Input1).
2. The CLS-token of the modified text (Input2).
3. The first token of the replaced word from the unmodified text (Input3).
4. The first token of the replacement word from the modified text (Input4).

Our classifier has 4 dense linear layers, each of dimensionality <hidden LM size, 1> for the following terms we believe that:

1. Input2: can be trained to indicate how funny the modified text is.
2. Input2 minus Input1: can be trained to indicate how funny the replacement itself is.
3. Input4: carries an indication of how unexpected the replacement word is. If the replacing token is inconsistent with the rest of the headline, we expect to see it in the vector.
4. Input4 minus Input3: carries information about the possible relationships between the original word and replacement word in the context of the text.

The output of these layers is then averaged and fed to a sigmoid activation, scaled between 0 and 2.8, to obtain the model's results. The system's architecture is shown in Figure 1

3.2 L2-Regularization

Regularization is the process of adding information in order to prevent overfitting. In multiple areas of machine learning a regularization term is often added to the loss function during training. The underlined belief in doing so is that given two models that perform equally well on the training set, the "simpler" of the two will perform better on the test set. Unfortunately, "simpler" is not so simple to define. It is often the case that systems use the L2-norm of their models as a measure of complexity. However, when a system incorporates a pre-trained LM with hundreds of millions of parameters and then fits it on a

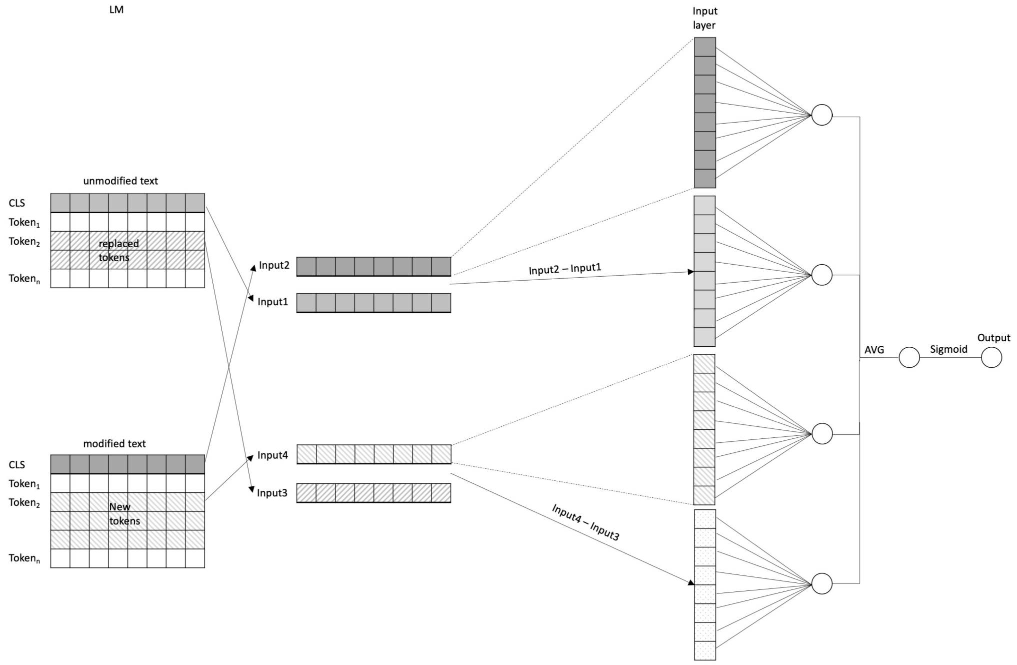

Figure 1: System's architecture

training set with less than ten thousand examples, the L2-norm of the entire model is meaningless. As an alternative, we use the change that our training process has caused to the weights of the LM as a measure of our model's complexity. Moreover, we incorporate the knowledge that the lower self-attention layers of transformer-based LMs often act in such a way that enables the model to perform dependency parsing tasks. We do so by setting the L2-regulerization coefficient to be small for the upper layers of the model and bigger the closer a layer is to the token embedding layer. Specifically, the results in this work were obtained by setting the L2-regulerization coefficient for each later as $C \cdot 1.2^{-N}$, where N is the distance of the layer from the embedding layer and C is a hyper-parameter that reflects the level of regularization. We combine the L2-regulerization with MSE loss function.

3.3 Freeze weigths

Our purpose is to use the dataset to fine-tune the LM to best fit our specific task. However, at the begining of our training, we start with a pre-trained LM and a randomly initialized model on top of it. While the weights of our model are random, it makes no sense to back-propagate our loss on the weights of the LM, and we therefore freeze those weights for the first epoch of training.

3.4 Ensemble

We ensemble models based of one of three language model: XL-Net(Yang et al., 2019), BERT(Devlin et al., 2018) and Roberta(Liu et al., 2019). To ensemble all the results, we separately average each language model and then use a weighted average of all three language models. The weigths of the language models are 0.5 for Roberta, 0.3 for XL-Net and 0.2 for BERT. Those weights yielded the best results on the validation set, with considertion to their individual performance.

4 Experimental setup

The competition dataset is described at (Hossain et al., 2019). For the first sub-task the dataset consists of 12,071 training samples and 3,024 test samples. Before the training phase we randomly set aside 1% of the training data for validation. The model is then trained for 15 epochs, and a version of the model

Model	RMSE
Baseline	0.574
Roberta	0.513
XL-Net	0.524
BERT	0.522
Ensemble	**0.507**

Table 2: Sub-Task 1 Results Comparing to Baseline.

Model	Accuracy
Baseline	49%
Roberta	65.7%
XL-Net	63.7%
BERT	63.5%
Ensemble	**65.8%**

Table 3: Sub-Task 2 Results Comparing to Baseline.

is saved after each epoch. The validation data that was set aside is used to select the best version of the model, the one that will be used for our ensemble. The results presented in this paper are derived from an ensemble of 90 models based on the three language models, 30 of each. The evaluation measurement for the first sub-task is Root Mean Squared Error (RMSE) on the overall test set.

For the second sub-task we used the predictions of the model from the first sub-task and didn't develop one specially for this sub-task. Hence, we used only the test set which consists of 2,960 samples. The measurement for this task is accuracy.

5 Results

At the first sub-task, the model achieved RMSE of 0.507 (2nd place). In Table 2 we present the result of the ensemble model in comparison to the baseline (all the predictions are the average score in the training set - 0.93) and to ensemble of each language model (30 models per each language model).

At the second sub-task, the model achieved accuracy of 65.8% (2nd place). In the Table 3 we present the result of the ensemble model in comparison to the baseline (the first sentence is always funnier) and to the ensemble of each language model.

As can be seen, the weighted ensemble of all the language models is superior in both sub-tasks. For a simpler model, Ensemble based on Roberta is not far behind.

6 Conclusion

In this paper, we described the system and experiments Amobee developed for task 7 in SemEval-2020: "Assessing Humor in Edited News Headlines", sub-tasks 1 and 2. We presented the L2-Regularization, a novel approach to train and fine-tune the language embedding model. This regularization, in addition to freezing the weights for the first epoch enables us to fit the language model to a given problem without losing the large amount of information that the algorithm already holds. In addition, the system uses an epoch-selection process for each model. The system consists of an ensemble of different language models - BERT, XL-NET and Roberta, and duplications from each LM, with a weighted average between the models. This system reached 2nd place for both sub-tasks with a RMSE of 0.507 for sub-task 1 and an accuracy score of 65.8% for sub-task 2. We plan to release the complete, fully trained version in the near future and to test the effect of L2-Regularization on different NLP tasks - such as topic classification, sentiment analysis, etc.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.