CS598 Assignment2 Report

Mingwei Huang

March 30, 2022

1 Summary

This assignment is about SemEval 2021 task 6 (Dimitrov et al., 2021) and I chose the subtask 1 which can be considered as a multi-classification task, to detect what types of propaganda/fallacy are used in a text. The baseline model is implemented based on the text embedding generated from MPNet sentence transformer (Reimers and Gurevych, 2019) and two fully connected layers with batchnorm. I then tried make use some mentioned entity embedding from Zhang et al. (2019) in an end-to-end way and it turns out the performance is indifferent with using their original model. At last, I tried data augmentation using the tool from Ma (2019) and it does improve the generalization performance.

2 Result & Finding

Here is the result:

Model	Macro-F1 score	Micro-F1 score
Random	0.08802	0.06250
Baseline	0.13195	0.52510
Entity	0.16337	0.55224
Baseline w/ aug(1,1)	0.21350	0.55062
Baseline w/ aug(2,5)	0.23651	0.53507

The baseline does very good job on predicting Loaded Language, moderate (about 50% accuracy on labels with sufficient data) but extremely poorly on remaining labels (0% recall accuracy). After utilizing entity embedding, even though the numbers increase a little bit, the basic performance is not changing. So I partially contributed this phenomenon to lack of data, and tried data augmentation.

The data is augmented as two parts, the texts with rare labels and other texts. aug$(2, 5)$ means for each text with rare labels, there are 5 automatically augmented texts generated from it, and for other texts, 2 augmented texts are generated. Data augmentation does bring up improvement on the recall for rare labels (about 10% in average).

3 Comment & Suggestion

There are 20 categories of fallacies in total, and several of them have less than 5 sentences in the training dataset. Given such lack of data, I don't think it is even possible for human to learn those types without understanding the meaning of these fallacies. And I think in order to make a magnitude of improvement, we have to use the information from definitions of these fallacies.

The reason for using entity not working well can be the lack of data. Also, I don't think the entity embedding does not contain much symbolic meaning in the popular cultural or political context. The meaning

of an entity, such as TRUMP and BERNIE, is completely different in the Wiki from their meaning in Twitters.

I don't think the Micro F1 score means anything here. Almost half of the labels are Loaded Language, and 85 percent of the labels are from only 3 labels out of 20. And the most Loaded Language fallacy (I guess) can be detected by some syntactical features. So I only focus on how to improving the performance on predicting all labels even though majority of them do not have more than 10 occurrences in the training dataset.

I cannot understand the augmented text at all but it is just working. :(

4 Some Implementation Details

The batchnorm layer is not directly applied sequentially, which does not yield any improvement and I guess that it may be because of the presence of the dropout layers. The batchnorm of the first hidden layer is added to the output of the second hidden layers.

The end-to-end entity embedding is directly concatenated to the text embedding, and the performance is indifferent from using the text embedding to initialize a weight for entity embedding.

The loss function that I used is binary cross entropy with same weight for each class. And weirdly, after some threshold, the larger the training epoch is, the larger the loss is on the validation data, but it turns out to be better on generalization (better recall on rare labels) with a slight trade-off on precision.

Using entity information on augmented data is not implemented yet.

5 Miscellaneous

The random baseline Macro F1 score provided in the official website leaderboard is about 0.04 which is half of mine, maybe it is just because I am luckier. But it also shows that the performance on the macro F1 score is not very stable especially when the score is low (and the label distribution is extremely imbalanced).

References

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. SemEval-2021 Task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, 2021. URL `https://arxiv.org/abs/2105.09284`.

Edward Ma. Nlp augmentation, 2019. URL `https://github.com/makcedward/nlpaug`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `http://arxiv.org/abs/1908.10084`.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities, 2019. URL `https://arxiv.org/abs/1905.07129`.