EFFECT OF TRAINING FRAGMENT LENGTH ON THE PERFORMANCE OF TRANSFORMERS IN TEXT COMPLEXITY CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

With the myriad practical applications of text complexity classification, it is important to optimize the training text fragment size for performance. We experiment with fine-tuning pre-trained BERT models to classify the complexity of Russian school text using different fragment sizes for training.

1 INTRODUCTION

Our aim here is to quantify and evaluate the effect of text fragment length in training on the performance of different transformer-based models (Vaswani et al., 2017). For this purpose, we have used a Russian School textbook corpus that we divide into fragments with different sizes to get training datasets, then train 3 models on these datasets of different fragment lengths using the regression task of quantifying text complexity. Our aim is to test the following null hypothesis: $H_0 = \{$Training fragment size is correlated with the model's MAE$\}$

2 RELATED WORK

The assessment of text complexity is one of the relevant tasks that are considered in Natural Language Processing. It is sometimes referred to as "text readability" or "text acceptability" and has many practical applications, like curating texts for school textbooks Fitzgerald et al. (2015), and picking the adequate pieces of text for Second language courses Kurdi (2020). Multiple studies have used handcrafted morphosyntactic features to predict the text complexity Kurdi (2020), Cuzzocrea et al. (2019). After the appearance of transformer-based language models like BERT Devlin et al. (2018), many studies attempted to assess its performance in the complexity classification task compared to handcrafted features (Iavarone et al. (2021), Alaparthi et al. (2022)). Martinc et al. (2021) ran a similar experiment to ours, comparing different models in the text complexity classification task, and found that BERT performed better on one dataset, and HAN (Yang et al., 2016) performed better on the 2 other used datasets. Iavarone et al. (2021) found that the usage of explicit language features performed better than BERT in a 7-class complexity classification task, and attributed this to the small size of the dataset, which is not enough to train a relatively large model like BERT.

3 METHODOLOGY

The experiment consists of parsing a corpus of Russian school textbooks into text fragments that constitute the dataset instances for fine-tuning a pre-trained BERT model on classifying the corresponding school grade of the presented text. We experiment with multiple maximum allowed fragment lengthes (ranging from 50 to 510), and 3 BERT models, all pretrained as language models for the Russian language: `rubert-base-cased` , `rubert-base-cased-sentence` , and `xlm-roberta-large-en-ru` . All models were acquired from the HuggingFace platform using the Python `transformers` library.

We consider the complexity classification task as a regression task, assuming that higher grades are correlated with more complex text passages. We assigned a complexity value $C(x)$ to each text

Table 1: Correlation between fragment size and MAE

Model	CE all	CE > 100
RuBERT	-0.35244494	-0.11844252
RuBERT-s	-0.43573392	-0.30136838
XLMR	-0.49109714	-0.41563056

passage, with x the school grade. There are 11 grades in the Russian school system (numbered from 1 to 11) and also "advanced" grades (taught in pioneer schools). We choose $C(x) = x$ for regular grades, and $C(x \text{ advanced}) = x + 0.5$ for advanced grades e.g. $C(10 \text{ advanced}) = 10.5$.

For the evaluation, we have used MAE and the Pearson Correlation Coefficient, we also mapped the output to bins and calculated the F1 score. We also tracked the training time.

4 Results

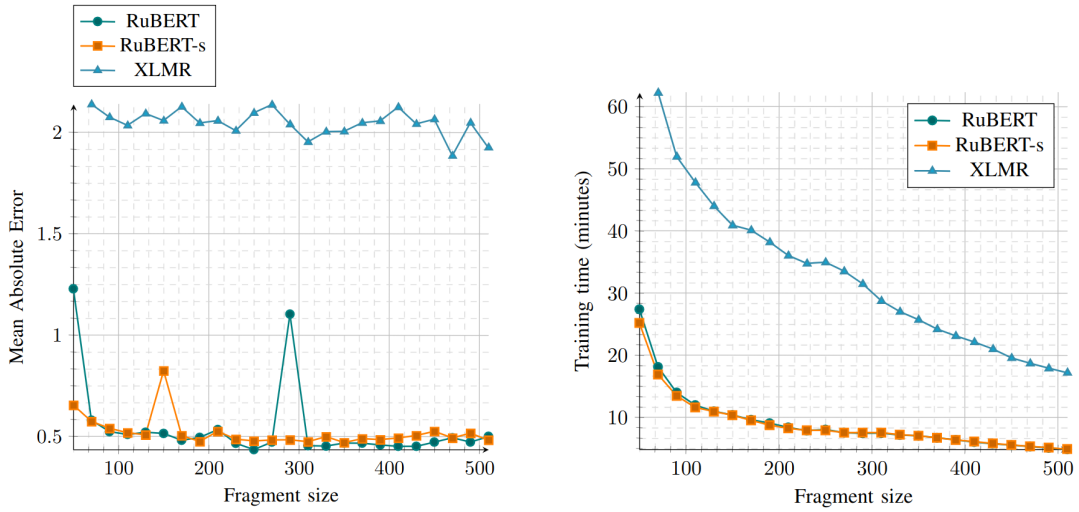

Figure 1: Left: MAE, Right: Model training time

In figure 1, we illustrate the main results of the experiment. XLMR performed worse than the 2 other models, but for all models, the MAE is does not vary significantly when changing the fragment size. The performance gets slightly worse for the smallest fragment sizes, and there are 2 unexplained performance drops at about the 150 and 300 fragment sizes: We hypothesize that they are spurious, but further experiments are needed. For the training time, we noticed that it decreases with bigger fragment lengthes (because there are fewer instances in the dataset). To answer our hypothesis, we have calculated the correlation coefficient between the fragment length and the MAE for all models, either for all fragment lengthes and also for fragment lengthes larger than 100 (table 1). Using a correlation threshold of ± 0.6, there is no correlation between fragment size and MAE, therefore we reject the null hypothesis. However, there is a weak correlation when including smaller fragment sizes.

5 Conclusion

We experimented with fine-tuning BERT models on text complexity classification using different text fragment sizes. The results suggest that the MAE of the task is not affected by the fragment length, except for the smallest fragment sizes (less than 100 tokens), which implies that one can freely choose a fragment size for training without hindering the model performance. The training time is reduced with bigger fragment sizes without affecting performance.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Varun Sai Alaparthi, Ajay Abhaysing Pawar, CM Suneera, and Jay Prakash. Rating ease of readability using transformers. In *2022 14th International Conference on Computer and Automation Engineering (ICCAE)*, pp. 117–121. IEEE, 2022.

Alfredo Cuzzocrea, Giosué Lo Bosco, Giovanni Pilato, and Daniele Schicchi. Multi-class text complexity evaluation via deep neural networks. In *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 313–322. Springer, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jill Fitzgerald, Jeff Elmore, Heather Koons, Elfrieda H Hiebert, Kimberly Bowen, Eleanor E Sanford-Moore, and A Jackson Stenner. Important text characteristics for early-grades text complexity. *Journal of Educational Psychology*, 107(1):4, 2015.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. Sentence complexity in context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 186–199, 2021.

M Zakaria Kurdi. Text complexity classification based on linguistic information: Application to intelligent tutoring of esl. *Journal of Data Mining & Digital Humanities*, 2020, 2020.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489, 2016.