Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Evaluating Deep Transfer Learning Models for Assessing Text Readability for ESL Learners

Yo EHARA[a*]
[a]*Department of Technology, Tokyo Gakugei University, Japan*
*ehara@u-gakugei.ac.jp

Abstract: Assessing the readability of texts is a basic task in educating English-as-a-second-language (ESL) learners. As the manual evaluation of readability requires considerable human effort and is costly, methods for automatically assessing readability are needed. In natural language processing, automatic readability assessment is considered a text classification task. Recently, the predictive performance of text classification methods has significantly improved owing to the development of deep transfer learning. In transfer learning text classification, a large unlabeled corpus is used for pre-training, following which fine-tuning with training data, i.e., pairs of texts and their labels manually annotated, is performed. The predictive performance of these methods depends on the pre-trained models and fine-tuning parameters. In previous studies, however, experiments were typically conducted using one pre-trained model with few fixed fine-tuning parameters because testing different models and parameters resulted in technical difficulties, such as insufficient availability of GPU memory. In this study, we compared various pre-trained models on various settings using an NVIDIA A100 unit with 80GiB of GPU memory. We found that using many epochs, considering many tokens, and using large models are key to achieving excellent accuracy.

Keywords: Readability, automatic assessment, natural language processing, second-language learning

1. Introduction

Assessing the readability of texts written by native speakers for second-language learners is essential in language education. For instance, this process is used to select texts in daily language classes. Notably, conducting readability assessments manually is quite costly. To address this limitation, ideally, we must gather reliable human assessors and have them read and evaluate texts. However, as such an undertaking would also be quite costly, we must develop automatic readability assessors using natural language processing (NLP).

Automatic readability assessment (ARA) is considered a text classification process in NLP (Vajjala & Lučić, 2018). The OneStopEnglish dataset (Vajjala & Lučić, 2018) is among the most reliable datasets for benchmarking ARA as a text classification task. In this dataset, professional language teachers read articles acquired from The Guardian newspaper and evaluate their readability for ESL learners.

In recent years, text classification has been an area where deep transfer learning techniques such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2019) have significantly enhanced predictive performance. Further, deep transfer learning techniques have been reported to improve ARA performance (Martinc, Pollak, & Robnik-Šikonja, 2021; Vajjala & Lučić, 2018). Text classification using deep transfer learning techniques can be divided into two stages. First, the model acquires basic patterns, such as the grammar of the text, from a large amount of raw texts written by native speakers. This stage is called *pre-training*. Thereafter, the pre-trained model is trained with manually created training data (pairs of texts and their labels manually annotated) for text classification. This stage is called *fine-tuning*. Various types of pre-trained models have been distributed. However, many previous studies conducted pre-training using only one type of model. For example, Martinc et al. (2021) used only one type of model, namely, “bert-base-uncased.” This tendency may be due to technical limitations such as insufficient GPU memory or training time.

Herein, we report the previously unreported performance of ARAs on the OneStopEnglish dataset, one of the most reliable datasets, with various pre-training models. For pre-training, the SciBERT model pre-trained on a large number of scientific papers (Beltagy, Lo, & Cohan, 2019) and the other models pre-trained on Wikipedia articles were compared. We also compared different fine-tuning parameters that were fixed in previous studies. The results indicated that the SciBERT achieved an accuracy of 0.991, suggesting the effectiveness of using scientific papers for pre-training and the influence of the size of pre-training models.

2. Experiments

We used the OneStopEnglish dataset (Vajjala & Lučić, 2018). The dataset has 567 texts, and each text is annotated with a three-point scale readability of *elementary*, *intermediate*, and *advanced*. We first randomly split the 567 texts into five folds: three folds with 114 texts and two folds with 113 texts. In the experiments, one-fold was used for the test, and the remaining four were used for training and validation. We followed the experiment settings of recent reports (Ehara, 2021; Martinc et al., 2021). Throughout the experiments, we used the Adam optimizer with a learning rate of 0.00001. The *maximum length* is the number of tokens to consider in each text. In other words, each model truncates each text to this number of tokens for classification. BERT can process up to 512 tokens. Although different models work well with different numbers of epochs, most models show best performance within 20 epochs of fine-tuning. Hence, for fair comparison, we report the best accuracy observed in 20 epochs of training. In the four folds, three were used for training and one was used for validation.

Each pre-training model is identified using a name, such as "bert-base-uncased." Because of space limitations, although we cannot show all detailed settings of the pre-training models, we list the identifiers of the models compared: **bert-(base/large)-(cased/uncased)**, **bert-large-(cased/uncased)-whole-word-masking**, and **allenai/scibert_scivocab_(cased/uncased)**. The details of each model can be found at https://huggingface.co/models. Briefly, all models except for SciBERT were pre-trained using 3.3B tokens from Wikipedia and Wikibooks, whereas SciBERT was pre-trained using 3.17B tokens from scientific papers (Beltagy et al., 2019).

Table 1 lists the experimental results. We compared three settings, namely, **Max. 128 tkn., Max. 512 tkn.**, and **Max. 512 tkn. half-train**. **Max. 128 tkn.** uses only the first 128 tokens of each text. **Max. 512 tkn.** uses the first 512 tokens of each text. **Max. 512 tkn. half-train** uses the first 512 tokens of each text, but the number of texts used for training is halved. In Table 1, we have abbreviated **whole-word-masking** in the names as **wwm**.

The average number of epochs corresponding to the best accuracy is written within "(" and ")", and the numbers are rounded as integers. From Table 1, we can easily observe the following novel findings.

1. Three epochs are not sufficient to achieve the best performance in all cells.
2. The maximum length significantly influences the accuracy.
3. The number of texts used for fine-tuning affects accuracy but not as much as the maximum length does.
4. Uncased models tend to achieve better performance than cased models.

Because we could not obtain the test sets used by Martinc et al. (2021), we could not directly compare our results with theirs. However, we confirmed similar scores in their settings. In their study, they fixed the number of epochs to 3 and achieved 0.647 using bert-base-uncased. In our experiments, bert-base-uncased at the third epoch resulted in 0.632. Ehara (2021) reported 0.92 using bert-large-cased-wwm with 128 maximum tokens. We observed 0.850 in this setting because our training/test split was different from the one that Ehara (2021) used. Overall, our best score, 0.991, is likely to have outperformed these previously reported accuracies.

Fine-tuning large models may require a large GPU memory. For example, to train **bert-large-uncased-wwm** with 512 maximum tokens, approximately 73GiB of GPU memory was required. Hence, we conducted all experiments using NVIDIA A100 80GiB. The requirement of a large GPU memory is presumably one of the reasons why this type of comparison has not been extensively conducted.

Table 1. *Best Accuracy Score of Each Method in 20 Epochs*

Model	Max. 128 tkn.	Max. 512 tkn.	Max. 512 tkn. half-train
bert-base-cased	0.850 (19^{th})	0.982 (14^{th})	0.947 (20^{th})
bert-base-uncased	0.912 (16^{th})	0.956 (12^{th})	0.938 (17^{th})
bert-large-cased	0.859 (13^{th})	0.956 (15^{th})	0.982 (10^{th})
bert-large-uncased	0.885 (15^{th})	0.982 (13^{th})	0.982 (17^{th})
bert-large-cased-wwm	0.850 (16^{th})	0.973 (10^{th})	0.956 (16^{th})
bert-large-uncased-wwm	0.903 (20^{th})	0.982 (6^{th})	0.947 (14^{th})
scibert_scivocab_cased	0.780 (4^{th})	0.982 (8^{th})	0.956 (11^{th})
scibert_scivocab_uncased	0.842 (7^{th})	0.991 (9^{th})	0.964 (11^{th})

3. Discussion

Table 1 shows that the use of many epochs, consideration of many tokens, and use of large models are key to achieving excellent accuracy. Further, the results show that the use of SciBERT slightly improves the accuracy for the **Max. 512 tkn.** setting. This implies that SciBERT is suitable for the OneStopEnglish dataset in this setting, which is a novel, previously unreported observation (Ehara, 2021; Martinc et al., 2021). The limitation of our study is that we used the OneStopEnglish dataset for our experiments. Whether SciBERT is also beneficial for other readability datasets remains an open question.

4. Conclusion

We compared previously insufficiently studied experiment settings and identified key parameters that influence the assessment accuracy scores, namely, the maximum length of tokens used in each text, the number of epochs, the number of the data used for fine-tuning, and the selection of the pre-training model. We used one of the most reliable evaluation datasets in this study, and in future, we plan to investigate the other datasets.

Acknowledgements

This work was supported by JST ACT-X Grant Number JPMJAX2006 in Japan. We used the AIST ABCI infrastructure and RIKEN miniRaiden system for computational resources. We appreciate the valuable comments from the anonymous reviewers.

References

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615-3620.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171-4186.

Ehara, Y. (2021). LURAT: a lightweight unsupervised automatic readability assessment toolkit for second language learners. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 806-814.

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. Computational Linguistics, *47*(1), 141-179.

Vajjala, S., & Lučić, I. (2018). OneStopEnglish corpus: a new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, 297-304.