Automatic Grading System Using Sentence-BERT Network

Ifeanyi G. Ndukwe[1(✉)], Chukwudi E. Amadi[2(✉)], Larian M. Nkomo[1(✉)], and Ben K. Daniel[1(✉)]

[1] University of Otago, Dunedin, New Zealand
{glory.ndukwe,larian.nkomo,ben.daniel}@otago.ac.nz
[2] Federal University of Technology, Owerri, Nigeria
emmanuel.amadi@futo.edu.ng

Abstract. The integration of digital learning technologies into higher education enhances students' learning by providing opportunities such as online examinations. However, many online examinations tend to have multiple-choice questions, as the marking of text-based questions can be a tedious task for academic staff, especially in large classes. In this study, we utilised SBERT, a pre-trained neural network language model to perform automatic grading of three variations of short answer questions on an Introduction to Networking Computer Science subject. A sample of 228 near-graduation Information Science students from one research-intensive tertiary institution in West African participated in this study. The course instructor manually rated short answers provided by the participants, using a scoring rubric and awarded scores ranging from 0 to 5. Some of the manually graded students' answers were randomly selected and used as a training set to fine-tune the neural network language model. Then quadratic-weighted kappa (QWKappa) was used to test the agreement level between the ratings generated by the human rater compared with that of the language model, on three variations of questions, including description, comparison and listing. Further, the accuracy of this model was tested on the same questions. Overall results showed that the level of the inter-rater agreement was good on the three variety of questions. Also, the accuracy measures showed that the model performed very well on the comparison and description questions compared to the listing question.

Keywords: Neural network · Natural language processing · Similarity · Short answer grading · BERT

1 Introduction

Language model pre-training and transfer learning have led to significant performance increase in NLP tasks [2,3]. The deployment of self-training methods such as Embeddings from Language Model (ELMo) [6], Generative Pre-trained Transformer (GPT) [8], Bidirectional Encoder Representations from Transformers

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 224–227, 2020.
https://doi.org/10.1007/978-3-030-52240-7_41

(BERT) language model [3], cross-lingual language model (XLM) [4] and XLNet [10] resulted in significant gains in performance. Thus, enabling researchers to smash several benchmarks with minimal task-specific fine-tuning and providing the rest of the NLP community with pre-trained models that could be easily fine-tuned and applied to generate the state-of-the-art results (with fewer data and less computation time). Consequently, generating sentence encoder models that are already trained on a large corpus and subsequently transferred to other tasks. For instance, Conneau et al. [1] showed how universal sentence representations trained using data from the Stanford Natural Language Inference datasets can consistently outperform unsupervised bag-of-words models such as word2vec-SkipGram and unigram term frequency-inverse document frequency (TFIDF) model.

Reimers et al. [9], argued that even though the BERT [3] and RoBERTa [5] language model have laid down new state-of-the-art sentence-pair regression tasks, such as semantic textual similarity, which allow all sentences to be fed into the network, the resulting computing costs overhead is massive. In their work, they proposed Sentence-Bidirectional Encoder Representations (SBERT), as a solution to reduce this bottleneck. SBERT modifies the BERT network using a combination of siamese and triplet networks to derive semantically meaningful embedding of sentences. This adjustment allows BERT to be used for some new tasks which previously did not apply to BERT, such as large-scale semantic similarity comparison, clustering, and information retrieval via semantic search. In this study, we utilised the SBERT language model to perform automatic grading of students short answer questions.

2 Method and Procedures

In this research, three variations of short answer questions, including description, comparison and listing, on an Introduction to Networking Computer Science subject (see Table 1), were administered to a sample of 228 near-graduation Information Science students from one research-intensive tertiary institution in West African. These questions were designed and administered by the course tutor, using the online Google form. Then the course instructor manually rated short answers provided by the students using a scoring rubric. The result of each answer scored one of the six possible ratings, 0, 1, 2, 3, 4, 5.

In order to generate the reference answers that was used as a training dataset to fine-tune our language model, we wrote a python code that randomly selected a maximum of ten distinct student answers for each rating scores. In order words, the code randomly selected 50% of distinct answers that got a particular rating score, and if the total count was greater than ten, it used the top ten selected answers. Otherwise, if the total count was less than ten, it used all the selected answers. The code also appended the standard answer provided by the course tutor to the list of randomly selected answers with ratings of 5. Then the SBERT language model was adapted and used to predict the rating scores. This model mainly functions by performing a search through all the reference answers used to

Table 1. Three variations of questions requiring short answers.

Type	Question
Description	What is server virtualization?
Comparison	Differentiate with examples between IPV4 and IPV6
Listing	Outline at least 5 networking devices that will be used for the integration process

fine-tune the model, in order to determine the one that has the closest similarity, for each provided answer to predict a rating score.

We used the quadratic-weighted kappa (QWKappa) [7,11] for assessing the agreement among the grades assigned by the different raters. Instead of the traditional Cohen's Kappa, we adopt QWKappa, because the former can capture the order information of the scores. For illustration purpose, suppose a response can have scores of up to 3 ratings (0, 1, 2), the first-rater scores a response as 0, the second-rater scores the same response as 1, and the third-rater scores the response as 2. While both the second and third raters disagree with the first-rater, it is clear that the second-rater is more similar than the third-rater. That difference cannot be captured by the traditional Cohen's Kappa, while QWKappa can. Finally, we applied Precision, Recall, and F1 measures to test the accuracy performance of this language model in predicting each variety of the question.

3 Result

Our language model achieved an average score of 0.70 on the QWKappa metric for all the question types, which is considered an outstanding score. The language model predicted comparison questions with the highest accuracy, followed by the description questions. In contrast, the listing question has the lowest accuracy, this result is not surprising, given the nature of the question; "Outline at least five networking devices that will be used for the integration process." The standard answer provided by the instructor had about ten items, and for any five correct answers, the students get a rating of five. Hence, there were several permutations in the answers students provided, and this may have contributed to the low prediction performance of this model on the listing question.

4 Summary

This study found that a pre-trained neural network model can be fine-tuned using minimal reference answers to predict the rating scores of a variety of questions type with reasonable accuracy. Results suggest that students can obtain timely feedback, and lecturers can reduce their workload by utilising automatic

grading systems for their students' short type answers. Furthermore, from the observed result on comparing the performance based on the precision, recall, F1-score and accuracy, we conclude that the prediction for the comparison and description answer type outperformed the listing answer type. Although the listing answer type obtained a reasonable score, there is an indication that answers that have several permutations do not necessarily perform well in this model. Hence, cases of complex questions of this nature may not work with an optimal level of accuracy. Despite the success demonstrated, a significant limitation is that this model does not provide reasoning and explanation capabilities, such as feedback to each learner with his particular mistakes. Also, in an online course with low instructor workload, the consistency of grading is very critical for learner's progress. This model generated some false-negative scores, which can be particularly severe due to confusion and time-wasting trying to find out the wrong mistakes in the answers, whereas they were correct. In future research, we hope to validate our proposed model by testing the model performance with standard datasets. We also aspire to see how this model performs in other subject areas. Other interesting aspects may also be explored further, including comparing the performance of this model with other existing models and providing a hint to correct answers.

References

1. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
2. Devlin, J., Chang, M.W.: Open sourcing BERT: state-of-the-art pre-training for natural language processing. Google AI Blog, 2 November 2018
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)
5. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
6. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
7. Automated Student Assessment Prize: The Hewlett Foundation: Short Answer Scoring. https://www.kaggle.com/c/asap-sas. Accessed 01 Mar 2020
8. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning. Technical report, OpenAI (2018)
9. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
10. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems. pp. 5754–5764 (2019)
11. Zhang, L., Huang, Y., Yang, X., Yu, S., Zhuang, F.: An automatic short-answer grading model for semi-open-ended questions. Interact. Learn. Environ. 1–14 (2019). Taylor & Francis