Essay Scoring Tool by Employing RoBERTa Architecture

Majdi Beseiso*
Department of Computer Science, Al-Balqa Applied University, Salt-Jordan
bsaiso@bau.edu.jo

ABSTRACT

The automated essay scoring (AES) has significant importance in machine grading of student essays particularly in standardized exams like the Graduate Record Examination (GRE). However, some issues in AES have remained unsolved over the past several years. The current approaches have scrutinized AES from both classification and regression perspectives. This study discusses the cutting edge architectures such as RoBERTa, XLNet, and BERT and compares their automated essay scoring performance. ASAP, a publicly accessible dataset is used for this purpose. The obtained results indicate that the natural language understanding (NLU) model proposed in this paper depicts significantly improved performance than all the other existing approaches.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Natural language processing; Information extraction.

KEYWORDS

Automated essay scoring (AES), BERT, Deep learning, Essay, LSTM, Neural network, Rating criteria, RoBERTa

ACM Reference Format:
Majdi Beseiso. 2021. Essay Scoring Tool by Employing RoBERTa Architecture. In *International Conference on Data Science, E-learning and Information Systems 2021 (DATA'21), April 05–07, 2021, Ma'an, Jordan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3460620.3460630

1 INTRODUCTION

In AES, the writing tasks of individuals provided a specific prompt (a topic on which the essay has to be written by the students) are automatically graded. AES can facilitate organizations such as Educational Testing Service (ETS) in grading a large number of student assignments in less time and with more accuracy. Therefore, most of the standardized testing exams such as SAT, TOEFL, and GRE employ machine grading systems. Moreover, the standardized testing giants such as Pearson.org and ETS.org have their own AES systems for scoring thousands of student assignments in considerably less time.

The hand-crafted attributes are commonly used in the current AES approaches. Deep learning (DL) based schemes such as Recurrent neural networks (RNNs) are also used in some recent studies for increasing efficiency. Moreover, some papers have employed ensemble-based approaches to attain better performance. However, all these schemes lack the training of models based on a huge compilation of well-written English content that grabs the contextual meanings too. This could be a major concern in essay scoring systems since the candidates upon getting a specific prompt to start with can write the essays in different contexts. Moreover, the word embedding models based on the lookup table cannot capture the grammatical correctness of an essay, which is an important factor in determining the scores assigned to an essay.

Considering this, the use of the advanced language understanding-based models such as XLNet [1] and RoBERTa [2] is proposed in this paper to alleviate this issue. Transformers [3] are a base component of the architecture of these models that effectively utilize the self-attention concept. Since these models are previously trained in an unsupervised manner using a bulk of well-written English context, therefore, the available ASAP dataset is easily generalized by these models leading to exceptional results. Moreover, LSTMs are used over RoBERTa for preserving some recurring structures to erase the issue of the document length. In this manner, the entire content of the essay is maintained and the optimum results are achieved on the ASAP dataset.

2 RELATED WORKS

A DL-based strategy has been used in this paper considering the efficacy of neural network-based algorithms in several NL processing tasks. The majority of the past studies have employed LSTMs [4, 5] or a combination of CNNs and LSTMs [6, 7] as a linear or logistic regression approach. The proposed approach is different from the existing approaches since transformer-based architecture is used as the base of the network. Moreover, the problem based on the length of the document in BERT-based models is also addressed in the proposed approach since no data is dropped off or truncated. Contrary to the past HAN based networks, the pre-training part of the BERT models is considerably modified by emphasizing on the prediction of the next sentence. Similar work is presented in [4] in which native text coherence is modeled; however, instead of marking essays, the coherence options were mainly used to capture adversarial examples. Some works [8, 9] have employed the idea of document classification. However, the majority of such networks were unable to acquire satisfactory output since the sense of coherency is missing in understanding language. For solving this issue, BERT and its variants [10, 11] are also used in this paper for generating sentence embeddings and eventually utilizing these hierarchically. However, no significant improvement is witnessed on the downstream task compared to other models. This might be because even though BERT and its variants had been trained using

*Majdi Beseiso is an assistant professor at the Department of Computer Science, Al-Balqa Applied University, Salt, Jordan

DATA'21, April 05–07, 2021, Ma'an, Jordan

ACM ISBN 978-1-4503-8838-2/21/04…$15.00
https://doi.org/10.1145/3460620.3460630

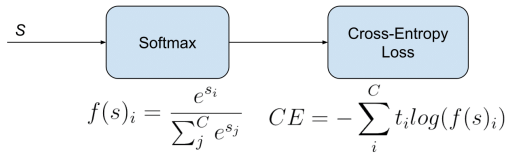

Figure 1: Minimized loss at each [MASK] token position

a considerably large volume of text in an unsupervised manner, however, the idea of pre-training is not sufficient enough for understanding the reasoning, fluency, and similar attributes that exist in a well-written essay.

3 PROPOSED APPROACH

In this paper, a new strategy is presented that involves the integration of the BiLSTM module with the leading edge NLU model, i.e. RoBERTa. Here, RoBERTa is employed as the language encoder to elucidate the consistency in an essay. Moreover, the document length issue in RoBERTa is handled using the chronological nature of BiLSTMs.

RoBERTa is a huge language model that is pre-trained in an unsupervised way on 160 GB of data. Similar to BERT, a transformer is the base unit of RoBERTa architecture as well. Therefore, it understands the language more than the standard models such as LSTMs and RNNs. Facebook released RoBERTa for improving the performance exhibited by BERT. For attaining this objective, RoBERTa was trained using a bigger dataset compared to BERT. Moreover, the Next Sentence Prediction (NSP) was not included in the pre-training objective of RoBERTa. Hence, the pre-training of RoBERTa is done based on the Masked Language Model (MLM) objective only, which facilitates RoBERTa in effectively capturing the consistency in essays without caring about the next sentence.

The MLM objective involves the uniform selection of 15% of all the tokens in the complete text body. 80% of these selected tokens are then replaced with [MASK] token. Moreover, 10% of the chosen tokens are substituted with a random word while the rest of 10% of the selected tokens remain unchanged. At the time of pre-training, the model uses the cross-entropy loss over the whole vocabulary for predicting the words at the [MASK] positions.

In the AES problem, an algorithm has to grade an essay provided a specific prompt, i.e. the prompt is a part of the essay. Keeping that in consideration, MLM loss is employed here for handling this problem. The body of the essay and prompt are combined for getting a single block of text for each essay. This is followed by the fine-tuning of the pre-trained RoBERTa model based on this essay body, which is attained by reducing the MLM loss for 9 epochs throughout the complete dataset. The MLM loss is the absolute cross-entropy loss over the whole vocabulary for the [MASK] tokens. The loss reduced at each [MASK] token position is given by Figure 1

Here, the embedding representation is denoted by s_i, and $f(s)_i$ represents the softmax operation performed on the whole vocabulary.

Once the model is fine-tuned on the essay body, the ASAP dataset is split in an 80 to 20 ratio for training and testing. Though the splitting of data is done after the fine-tuning of the language model, the results will not be affected since the bias developed by the test set on the model weights is negligible because of the significantly large training set and the pre-training of the model on an even bigger dataset. For handling the problem of document length limitation in models like RoBERTa, the complete essay is initially divided into chunks of 512 Byte-Pair tokens [10]. After that, the first token of each chunk is used as the embedding representation for that chunk. When the embedding representation is attained for all the chunks, they are passed to BiLSTMs for maintaining a temporal sequence among these.

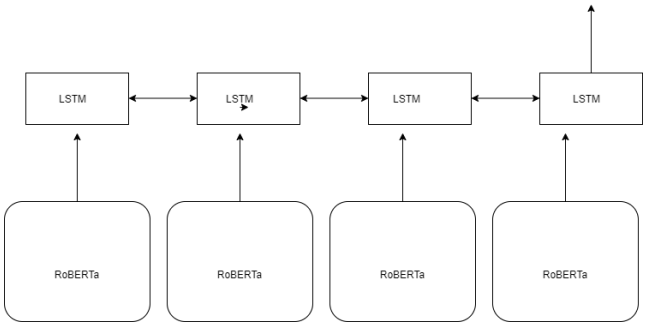

Figure 2: Schematic representation of complete architecture

The schematic representation of the complete architecture is shown in Figure 2. A chunk of the document (512 Byte-Pair Tokens length) is represented by each RoBERTa block, whose output is the first hidden state of every chunk. Finally, these representations are forwarded to the BiLSTMs where the output is taken from the final time step. The final embedding representation from the last LSTM time step is passed through a group of completely connected layers with a sigmoid activation function at the end for obtaining the final/predicted score. The sigmoid activation function is used since all the normalization problems can be effectively handled by it.

Lastly, the mean squared error (MSE) loss over the predicted scores and actual scores is minimized. The MSE loss is computed by the following equation.

$$\text{MSE} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} (\text{y}_\text{i} - \hat{\text{y}}_\text{i})^2$$

Here size of sample set is denoted by N, whereas the actual value and predicted value is denoted by y_i and $\hat{y}_i$ respectively.

In a nutshell, the following steps are performed for training this innovative neural network architecture.

- The RoBERTa language model is fine-tuned on the essay score from the ASAP dataset
- The essay is divided into chunks of 512 byte-pair tokens length followed by the application of the RoBERTa model on every chunk to obtain representation for all the chunks.
- The embedding representation of all the chunks is passed on to BiLSTMs for maintaining temporality. Then the sigmoid activation function and completely connected layers are employed to reduce the hidden state output of the last LSTM time step to eventually acquire a normalized score for the essay

4 EXPERIMENTAL SETUP

4.1 Datasets

The most commonly employed ASAP Dataset, which is released by HP in Kaggle competition for robustly grading essays, has been used in this study for automated essay scoring. The ASAP dataset contains eight sets of essays having around 12978 essays, and each set represents essays with particular attributes. The text is in ASCII form with an average essay length of 550 words and several gold annotations linked to it. Each of these essays also has a resolved score similar to exams like GRE and TOEFL in which an essay is graded by many human graders and a final score is determined.

4.2 Metrics for performance evaluation

The Quadratic Weighted Kappa (QWK) score [12] is used for assessing the efficiency of the proposed strategy. The QWK metric is mainly used due to its scoring techniques and robustness to essay scenarios. Its value ranges between 0 and 1, where 0 stands for no agreement between raters, and 1 denotes complete agreement. In this study, the QWK score between the resolved essay scores and the automated essay scores is calculated separately for every set of essays. The obtained kappa values are put through Fisher Transformation and then the mean is calculated.

It is hypothesized that a well-written essay should maintain consistency between sentences. This consistency can be ensured by calculating the similarity score between a sentence and its adjoining sentences such that neighboring sentences should be more coherent compared to the farther ones. The approach presented in [13] is followed in this paper to achieve this consistency. For this purpose, the similarity of a sentence (I) with its adjoining sentence is calculated first, which is denoted by sim2. After that, the similarity between I and all the other sentences is computed, which is represented by sim all. The cosine similarity has been chosen as the similarity function.

5 RESULTS

The proposed technique is compared with the recently introduced DL techniques and feature extraction-based pipelines for drawing a clear idea about its effectiveness. A comprehensive comparison of the proposed model with different models such as feature-based baseline (involving POS tagging and dependency parsing features), TSLF, NLI-DCM-BCA, NLI-DCM-BCA+features, BERT-HAN, and BLSTM is given in Table 1.

Table 1: Comparison of proposed approach with its counterparts

Model	Average QWK
Feature based baseline	0.74
TSLF	0.76
NLI-DM-BCA + Features	0.77
NLI-DM-BCA	0.73
BERT-HAN	0.68
BLSTM	0.66
RoBERTa + BLSTM (Proposed Approach)	0.80

6 CONCLUSION AND FUTURE WORK

There is a strong need for transparency in the AES system since it is employed in grading several standardized exams. The available techniques involve DL-based approaches, classic NLP pipelines, and the combination of these two [14]. However, the coherency problem in student-written essays is not effectively addressed by any of these existing techniques. Considering this, RoBERTa, a cutting edge deep learning approach is used in the proposed technique for handling this problem. The proposed approach depicted an exceptional performance compared to the existing approaches in tackling the coherency issue along with keeping the other attributes intact. The proposed method significantly surpasses the available approaches that prove its effectiveness in solving coherency issues in AES. This work may be extended further for automatically solving the issue of bad-faith essay submission, which might be attained by conducting comprehensive research in the domain of automated essay generation.

REFERENCES

[1] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le.2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Advances in neural information processing systems. 5754-5764. arXiv: 1906.08237. Retrieved from https://arxiv.org/abs/1906.08237

[2] 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.arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc. NY,United States, 6000-6010. arXiv:1706.03762. Retrieved from https://arxiv.org/abs/1706.03762

[4] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe.2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)(New Orleans, Louisiana, June 2018). ACL, 263-271. https://doi.org/10.18653/v1/N18-1024

[5] Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang.2018. Automatic Essay Scoring Incorporating Rating Schema via Reinforcement Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Brussels, Belgium, October-November, 2018). ACL, 791-797. https://doi.org/10.18653/v1/D18-1090

[6] Kaveh Taghipour and Hwee T. Ng.2016. A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1882–1891. https://doi.org/10.18653/v1/d16-1193

[7] Huy V. Nguyen and Diane J. Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In The Thirty-Second AAAI Conference on Artificial Intelligence, 5892–5899.

[8] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. ACL, 1480–1489. https://doi.org/10.18653/v1/N16-1174

[9] Farah Nadeem and Mari Ostendorf.2018. Estimating linguistic complexity for science texts. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (New Orleans, Louisiana, June 2018).ACL, 45–55. https://doi.org/10.18653/v1/W18-0505

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

[11] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (New Orleans, Louisiana, June 2018). ACL, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[12] David Vaughn and Derek Justice. 2015. On the direct maximization of quadratic weighted kappa. arXiv:1509.07107. Retrieved from https://arxiv.org/abs/1509.07107

[13] Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated Essay Scoring with Discourse-Aware Neural Models. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (Florence, Italy), 484-493. https://doi.org/10.18653/v1/W19-4450

[14] Majdi Beseiso and Saleh Al-Zahrani. 2020. An Empirical Analysis of BERT Embedding for Automated Essay Scoring. International Journal of Advanced Computer Science and Applications(IJACSA), 11(10), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111027