Bert Embedding and Scoring for Scientific Automatic Essay Grading

Abeer Abdulkarem and Anastasia Krivtsun
Department of Information Technologies and Management, Platov South-Russian State Polytechnic University (NPI), Prosveshchenie Str. 132, 346428 Novocherkassk, Russia
abeerabdulsalam15@gmail.com, anastasia.srstu@gmail.com

Keywords: Automatic Essay Grading, Word Embedding Techniques, BERT Techniques, Neural Network.

Abstract: The educational landscape is experiencing a surging demand for Automated Essay Grading (AEG), prompting the need for innovative solutions. This paper introduces a cutting-edge methodology that harnesses the power of Bidirectional Encoder Representation from Transformers (BERT) to embed and score essays in the scientific AEG domain. Tackling challenges such as Out-of-Vocabulary (OOV), BERT's contextual embedding proves instrumental. The study meticulously evaluates a hybrid architecture on a prototype incorporating non-English essay answers, establishing a benchmark against state-of-the-art studies. Beyond the expeditious grading of essays, particularly in scientific realms, this paper makes a substantial contribution to the ever-evolving field of educational technology. The AEG task revolves around the automation of essay response grading, where input data encompasses essay answers, and output data comprises assigned scores. The adopted mathematical model seamlessly integrates BERT for contextual embedding and subsequent scoring. The evaluation uncovers compelling results, underscoring the effectiveness of the proposed BERT-based model. The model's architecture, characterized by bidirectional layers and a dense output, encompasses a notable 2,243,401 parameters. Significantly, the Kappa Score achieved by the model impressively stands at 0.9725, highlighting its superiority over existing methodologies.

1 INTRODUCTION

In the dynamic landscape of educational assessment, the advent of Automatic Essay Grading (AEG) stands as a transformative force, significantly impacting the evaluation processes within esteemed institutions such as the Educational Testing Service (ETS) [1]. The imperatives of efficiently and accurately appraising a substantial volume of student assignments have driven the adoption of machine grading systems. Noteworthy, standardized examinations, including SAT, TOEFL, and GRE, have seamlessly integrated machine grading methodologies, and industry giants such as Pearson.org and ETS.org have pioneered the development of proprietary AEG systems, streamlining the assessment of a myriad of student essays with unprecedented efficiency [2].

Traditionally, AEG methodologies have leaned on manually crafted attributes, a paradigm increasingly complemented by the integration of Deep Learning (DL) techniques, exemplified by the utilization of Recurrent Neural Networks (RNNs). However, these approaches encounter a common limitation an inherent reliance on finite datasets for model training, hindering their capacity to comprehensively discern the nuanced contextual intricacies prevalent in well-articulated essays. Furthermore, conventional word embedding models based on lookup tables confront the formidable challenge of capturing grammatical correctness, an indispensable factor in ensuring precise essay scoring [3].

The earliest research efforts on AEG depicted were concentrating on determining structural features such as number of paragraphs, sentences, words, spelling and grammatical mistakes [4]. Apparently, the assessment of an essay answer using only the previously mentioned structural criteria would seem insufficient. It is necessary to examine the morphological and semantic aspects of the answer in order to give accurate scoring. Therefore, afterward researches have started to include lexical analysis, some sort of semantic analysis using external knowledge sources such as dictionaries, lexicons or ontology, or morphological. However, the emergence

of word embedding techniques which are based on neural network architectures has contributed toward a revolutionary progress in terms of the semantic analysis. Most text analysis/mining tasks such as document classification, topic modeling, question-answering and even AEG task have begun to take the advantage of word embedding due to its capability of determining the vector of a particular word in multiple dimensions where its lexical, semantic and syntactic aspects can be captured. One of the most commonly used datasets in essay grading is the Automatic Student Assessment Prize (ASAP).

Word embedding is the task of processing a series of tokens/terms through a neural network architecture in which the output of such a network is the prediction of consequent term. During the neural network training, a vector embedding starts to be emerged in the hidden layer of the network. Such a vector would consist of multiple values that indicate the position of the word, its lexical and semantic perspective. Figure 1 shows the traditional word embedding architecture known as Word2Vec. Word2Vec is a technique in natural language processing (NLP) for obtaining vector representations of words [5].

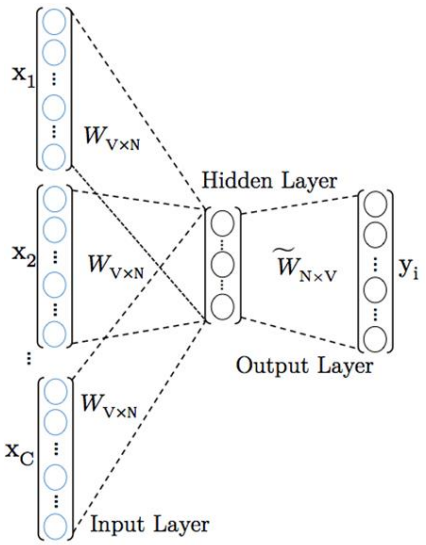

Figure 1: Word2Vec architecture.

A remarkable issue behind using Word2Vec is that it needs a large amount of text to train on in order to acquire better vector representation. For this purpose, have used a pretrained architectures of Word2Vec and GloVe for an AEG task. GloVe is also a very popular unsupervised algorithm for word embeddings that is also based on distributional hypothesis – "words that occur in similar contexts likely have similar meanings". Those architectures had been trained on vast amount of text and saved for future usage. After turning the template and students answers into vector representation, the authors have applied a regression analysis. The same ASAP dataset has been used and results of accuracy was 56% [6].

In the same regard, it have also utilized a pretrained word embedding model of GloVe in order to conduct AEG task. A mapping process has been performed to turn the answers' text into the GloVe representation. Then, a regression analysis has been conducted on ASAP dataset and the acquired accuracy was 77%. [7].

It also have been proposed an enhanced word embedding representation known as Bag of Super Word Embedding. This representation aims to provide embedding vectors based on significant terms. In other words, it focuses on specific important terms then, it gives relevant embedding to each term based on its relevancy to the pre-specified terms. Using a regression analysis based on ASAP dataset, the proposed method showed an accuracy of 78% [8].

ELMo is another popular word embedding framework. It was developed at Allen Institute for AI. It addresses the fact that the meanings of some words depend on the context. ELMo does not produce a fixed vector representation for a word. Instead, ELMo considers the entire sentence before generating embedding for each word in a sentence. This contextualized embedding framework produces vector representations of a word that depends on the context in which the word is actually used. ELMo uses a deep bi-directional language model giving the model a better understanding of not only the next words, but also the preceding ones [9].

Lastly, it have been proposed an AEG method based on Word2Vec representation and an advance prediction approach known as Gated Recurrent Unit (GRU). After replacing the answers' text into its Word2Vec representations, the proposed GRU has been applied to predict the score. Using ASAP dataset, accuracy result was 86% [10].

The state of the art in AEG task is mainly focusing on utilizing word embedding techniques. The majority of literature concentrated on traditional Word2Vec or pretrained Word2Vec and GloVe. However, these word embedding techniques suffer from a remarkable issue known as Out-of-Vocabulary (OOV).

This problem occurs when a word embedding model is trained on set of text tokens and later encounter a word that has no embedding vector on its model. In other words, such an unseen word would have no presence in the training text. In that case, researchers attempt to avoid this problem by normalizing a fixed vector (usually full of zeros) to substitute the absence of vector embedding for OOV words. In addition, the traditional word embedding techniques deal with text in a word-level rather than sentences or paragraphs. Since the AEG task is mainly depending on sentences or paragraphs answers thus, it would pose another problem.

On the other hand, most of the researches in AEG task have utilized datasets that are related to second language tests such as the ASAP. These datasets contain essay answers related to general topics. Yet, not all exams would have general topic essay answers. Some subjects especially related to science require a domain specific answer to a particular question. Customizing the AEG task to a particular scientific domain of interest would facilitate mastering its questions and answers.

2 RESEARCH METHODOLOGY

The methodology adopted for the Automated Essay Grading (AEG) task encompasses a two-fold approach aimed at addressing critical challenges in the existing landscape. This section provides a comprehensive overview of the proposed methodology's key components and phases,

A crucial facet in the landscape of Automated Essay Grading (AEG) research is the selection of datasets, a factor that significantly influences the efficacy and generalizability of proposed methodologies. Notably, the majority of AEG studies have traditionally gravitated towards datasets associated with second language tests, exemplified by widely used benchmarks like ASAP.

While these datasets offer valuable insights and facilitate a comprehensive understanding of AEG challenges, they predominantly comprise essay answers related to general topics. However, the diverse array of academic disciplines implies that not all examinations yield responses aligned with generalized themes. In particular, scientific subjects demand a domain-specific approach, where questions are tailored to assess knowledge and comprehension within a particular scientific domain.

Customizing the AEG task to a specific scientific domain holds the promise of enhancing the relevance and accuracy of grading. This involves formulating questions within the chosen scientific domain, constructing model answers, and subsequently testing students.

The manual assessment of student responses by teachers provides labeled scores, which undergo statistical analysis to derive average scores. The culmination of this process yields a meticulously curated dataset comprising questions, model answers, and labeled student responses.

In this subsection, we scrutinize the prevalent practice of relying on generic datasets and emphasize the need to tailor AEG research to specific scientific domains. The ensuing sections detail the proposed methodology's dual phases, addressing the OOV challenge and sentence-level embedding problems using Bidirectional Encoder Representation from Transformers (BERT) on benchmark datasets like ASAP and a newly developed scientific dataset.

BERT was released with two versions: BERT-base and BERT-large, and each has a cased and uncased iteration (there is also a Chinese BERT for Chinese and a Multilingual BERT that was originally trained on 102 different languages). BERT-base has twelve layers (Transformer blocks), each with twelve self-attention heads and 768 hidden neurons. BERT-base consists of approximately 110 million parameters: approximately 24 million from the embeddings, 85 million from the transformers, and one million from the pooler (BERT-large, comparatively, has roughly 340 million parameters). BERT-large is over three times as large, but it has not been shown to outperform BERT-base by an equally significant margin. BERT-base’s architecture can be seen in Figure 2, where twelve encoders are stacked sequentially. Each encoder is a transformer with its own attention heads. It is also important to note that only the encoder portion of the transformer (shown) is included in BERT’s architecture, as BERT is not a generative model and does not implement a decoder [11].

The application of Bidirectional Encoder Representation from Transformers (BERT) in Automated Essay Grading (AEG) represents a pivotal aspect of this study, leveraging cutting-edge techniques for enhanced performance. BERT's model

architecture, inspired by Vaswani et al., stands as a beacon of innovation, purposefully designed to navigate the intricacies of scientific text analysis [12].

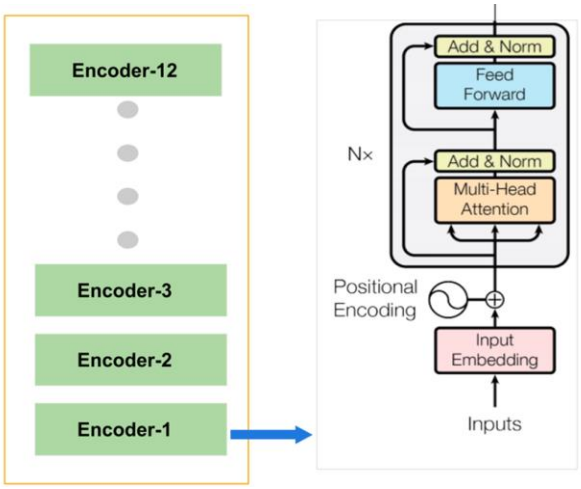

Figure 2: BERT architecture.

The methodology of this study consists of two main phases as shown in Figure 3.

The first phase represents the application of the proposed BERT embedding that is intended to solve both OOV and sentence-level embedding problem for the AEG task. For this purpose, the benchmark dataset of ASAP that has been widely examined by the literature, will be used in this phase. After applying the BERT embedding and scoring, an evaluation task will take a place in order to assess the automatic scoring produced by the proposed BERT method and comparing its results against the state-of-the-art. Once, the proposed method demonstrated superior results in terms of accuracy compared to the literature, the second phase will take a place.

The second phase represents the development of new scientific dataset for the AEG task. To do so, there are multiple procedures will be conducted. First, a set questions in specific domain of science will be initiated. Then, a model answer for each question will be formulated. Consequentially, the questions will be given to students in order to be tested. The tested answers produced by the students will be assessed by different teachers in order to give a manual score for each essay answer. The scores given by the teachers will be undergoing statistical analysis in order to take the average score. Lastly, a dataset of questions, model answers, and tested and labeled/scored answers.

Finally, the proposed BERT embedding and scoring will be applied on the dataset. The results of BERT scoring will be compared against teachers' scorings.

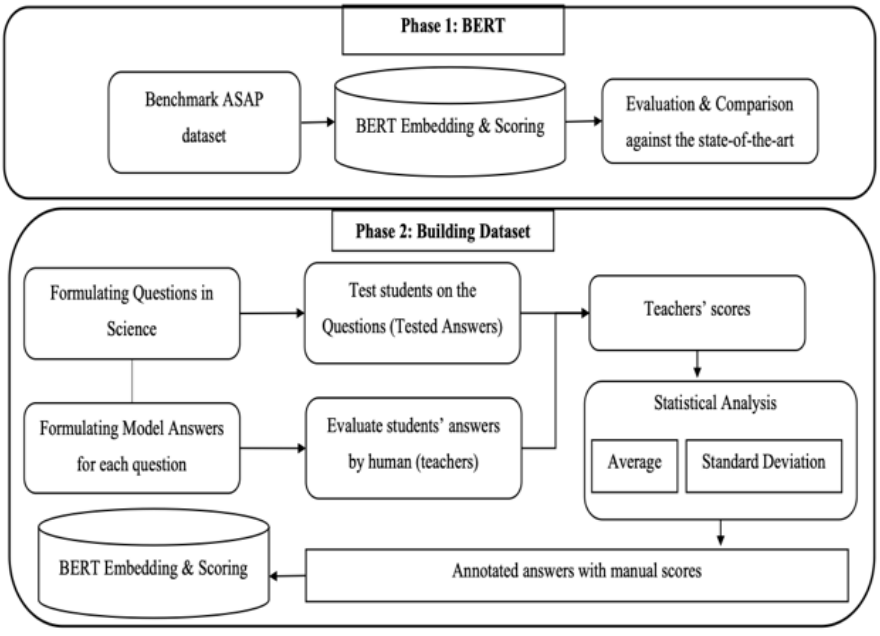

Figure 3: Methodology.

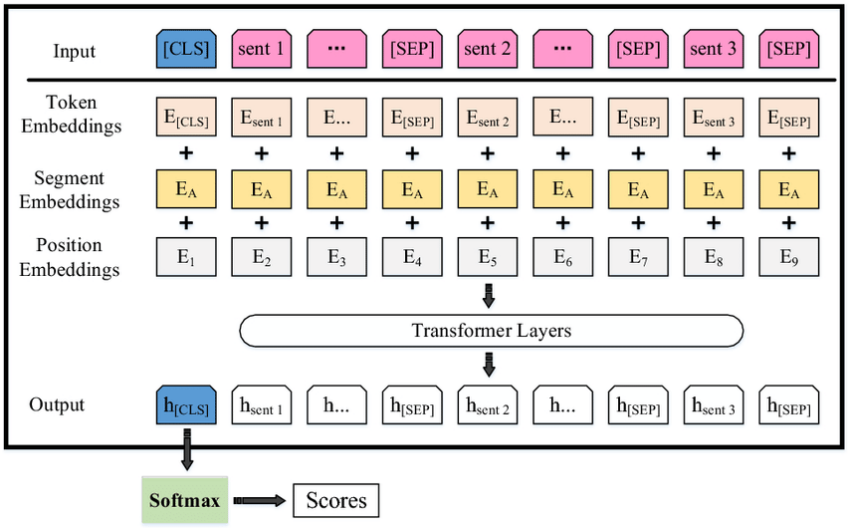

Figure 4: BERT's Input Representation Strategy.

To fully harness the capabilities of Bidirectional Encoder Representation from Transformers (BERT) in Automated Essay Grading (AEG), it's imperative to delve into the intricacies of BERT's architecture. BERT represents a paradigm shift in natural language processing, offering a sophisticated framework that aligns seamlessly with the nuances of scientific text analysis. BERT stands tall with its multi-layer bidirectional Transformer encoder. Departing from conventional models, this architecture introduces Transformers with bidirectional self-attention, a key element in capturing contextual dependencies with unparalleled sophistication. The versatility of BERT shines through in its adept handling of both single sentences and sentence pairs, providing a nuanced representation ideal for the complexities of scientific essays [13].

Notably, BERT employs Transformers with bidirectional self-attention, a departure from conventional models. This departure is significant, as it enables BERT to capture contextual dependencies with unparalleled sophistication.

BERT's prowess in handling a spectrum of downstream tasks is underpinned by its versatile input representation. It seamlessly accommodates both single sentences and sentence pairs, thereby accommodating the complex nature of scientific essays. Leveraging WordPiece embeddings with a 30,000-token vocabulary, BERT employs a special classification token ([CLS]) at the sequence's outset. The final hidden state corresponding to this token serves as the aggregate sequence representation for classification tasks [14].

A notable feature is the unambiguous representation of sentence pairs. Employing a special token ([SEP]) to separate sentences and adding a learned embedding to distinguish sentence A from sentence B BERT orchestrates a comprehensive input representation strategy, exemplified in Figure 4.

Pre-training, the inaugural phase in BERT's journey, unfolds over extensive unlabeled data using two unsupervised tasks. The first task involves Masked Language Modeling (MLM). Unlike traditional language models constrained by left-to-right or right-to-left conditioning, BERT employs a masked token approach. A percentage of input tokens are randomly masked, and the model predicts these masked tokens, paving the way for deep bidirectional understanding [11].

The second task, Next Sentence Prediction (NSP), addresses the intricacies of sentence relationships. Generating a binary next sentence prediction task, BERT establishes a foundational understanding of the contextual interplay between sentences. A noteworthy aspect is the synergy between these pre-training tasks, showcasing BERT's commitment to bidirectionality while mitigating mismatches between pre-training and fine-tuning [15].

BERT consists of two main architectures; language modeling and fine-tuning. The first architecture aims at process an answer sentences where each sentence is represented by its tokens along with two tags of ‘CLS’ and ‘SEP’ which refer

to the beginning and ending of a sentence respectively as shown in Figure 5. The output of this architecture is the language modeling where BERT would have the ability to understand the answer text.

The second architecture of BERT contains the fine-tuning where the processed answer in the previous architecture will be processed as an input to such an architecture. Fine-tuning, a streamlined process enabled by BERT's self-attention mechanism, unfolds seamlessly across a spectrum of downstream tasks. The model's inherent bidirectionality proves pivotal, allowing it to encode and comprehend text pairs efficiently.

In the scientific automatic essay grading context, this adaptability is invaluable. Fine-tuning involves plugging in task-specific inputs and outputs, thereby configuring BERT for tasks ranging from question answering to text classification. The bidirectional cross attention facilitated by BERT's self-attention mechanism elegantly unifies the encoding of text pairs, offering a versatile solution for the demands of scientific essay grading.

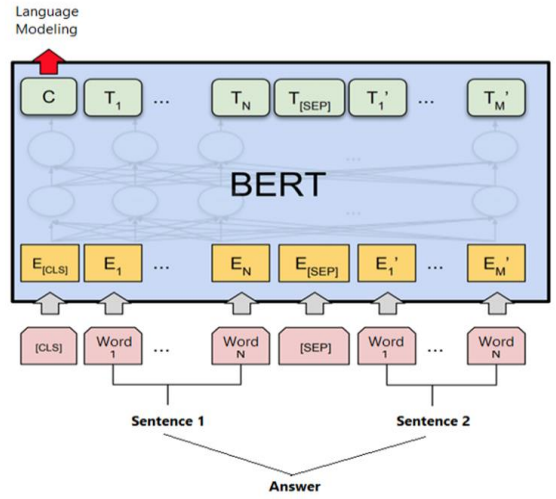

Figure 5: BERT language modeling.

The output of this architecture is the answer score prediction as shown in Figure 6. Afterwards., the evaluation metrics that intended to examine the performance of the proposed architecture will be applied. The common accuracy metric used for AEG task is the Quadratic Weighted Kappa (QWK) [11].

Such a metric is intended to calculate the agreement between human score and automatic score. It can be computed as follows:

$$QWK = \frac{p_o - p_e}{1 - p_e}, \quad (1)$$

where p_o is the observed agreement and p_e is the agreement by chance between human and automatic rater.

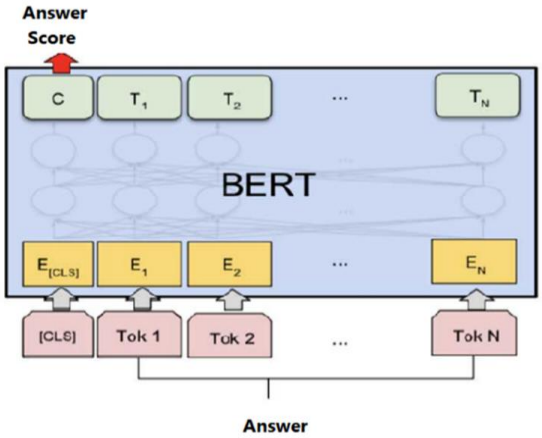

Figure 6: BERT fine-tuning.

Since this study is aiming to conduct the experiments on ASAP dataset which is a benchmark that contain scored answers by multiple teachers. Therefore, the evaluation would take a straightforward style where the score generated by the proposed BERT architecture will be compared directly with the original or actual score within the dataset.

After evaluating the scores produced by the proposed hybrid architecture, it is necessary to consider the state-of-the-art studies in the comparison. This is to point out the actual novelty or improvement depicted by the proposed hybrid architecture. Hence, all the state-of-the-art studies that have considered BERT embedding, bidirectional LSTM or CNN will be considered.

This evaluation paradigm will take a place on the proposed prototype where a collection of non English essay answers will be acquired. Consequentially, the proposed BERT architecture will be applied on such essays. Lastly, three teachers will be asked to score these answers and comparison between their scores and the proposed hybrid architecture's scores will be accommodated.

The model's architecture, characterized by bidirectional layers and a dense output, encompasses a notable 2,243,401 parameters. Significantly, the

Kappa Score achieved by the model impressively stands at 0.9725, highlighting its superiority over existing methodologies.

3 CONCLUSIONS

One of the significant processes within the education is the assessment of student's performance. Oral and MCQ examinations are the easiest parts to be automated where a web/mobile portal can be developed and facilitate the communications and exam conducting. However, the assessment of essay answers is still representing the most challenging task to be turned into a fully-automatic task.

The automation of essay answer grading requires a wide range of textual analysis including the lexical, morphological, semantic and syntactic aspects to train the computer to give a grade for a particular essay answer. Improving the AEG task would be a significant contribution to the educational process especially in the online manner where plenty of time spent by the instructor/teacher to give the grade will be saved.

In this research a new embedding technique based on Bidirectional Encoder Representation from Transformers (BERT) were proposed to overcome the OOV and sentence-level embedding problems in AEG task. The proposed method, incorporating BERT's contextual embedding, outperforms traditional approaches. Notably, the model architecture, featuring bidirectional layers and a dense output, comprises 2,243,401 parameters. The Kappa Score of 0.9725 attests to the model's exceptional performance.

The automation of essay grading, leveraging BERT's contextual embedding, stands as a breakthrough. The model, trained to analyze lexical, morphological, semantic, and syntactic aspects, significantly contributes to the educational process.

Secondly, this research study also curates a new question answering dataset in the science domain with the assistance of manual/human labelling scores as far as annotation is concerned. Improving the AEG task would be a significant contribution to the educational process especially in the online manner where plenty of time spent by the instructor/teacher to give the grade will be saved.

REFERENCES

[1] F. Li, X. Xi, Z. Cui, D. Li, and W. Zeng, "Automatic Essay Scoring Method Based on Multi-Scale Features," Applied Sciences (Switzerland), vol. 13, no. 11, Jun. 2023, doi: 10.3390/app13116775.

[2] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic, "Nougat: Neural Optical Understanding for Academic Documents," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.13418.

[3] F. Nadeem, H. Nguyen, Y. Liu, and M. Ostendorf, "Automated Essay Scoring with Discourse-Aware Neural Models," 2019, [Online]. Available: www.smashwords.com.

[4] M.V. Koroteev, "BERT: A Review of Applications in Natural Language Processing and Understanding."

[5] Y. Farag, H. Yannakoudakis, and T. Briscoe, "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.06898.

[6] Y.-J. Jong, Y.-J. Kim, and O.-C. Ri, "Improving Performance of Automated Essay Scoring by using back-translation essays and adjusted scores." [Online]. Available: https://github.com/j-y-j-109/asap-back-translation.

[7] H. Zhang and D. Litman, "Co-Attention Based Neural Network for Source-Dependent Essay Scoring," Aug. 2019, doi: 10.18653/v1/W18-0549.

[8] C. T. Lim, C. H. Bong, W. S. Wong, and N. K. Lee, "A comprehensive review of automated essay scoring (Aes) research and development," Pertanika Journal of Science and Technology, vol. 29, no. 3. Universiti Putra Malaysia Press, pp. 1875-1899, 2021, doi: 10.47836/pjst.29.3.27.

[9] J. Liu, Y. Xu, and Y. Zhu, "Automated Essay Scoring based on Two-Stage Learning," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1901.07744.

[10] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," Artif Intell Rev, vol. 55, no. 3, pp. 2495–2527, Mar. 2022, doi: 10.1007/s10462-021-10068-2.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018, arXiv e-prints, page arXiv:1810.04805.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, et. al., "Attention Is All You Need," 2017, arXiv:1706.03762.

[13] K. Taghipour and H. T. Ng, "A Neural Approach to Automated Essay Scoring," [Online]. Available: https://www.kaggle.com/c/asap-aes.

[14] B.F. Dhini, A.S. Girsang, U.U. Sufandi, and H. Kurniawati, "Automatic essay scoring for discussion forum in online learning based on semantic and keyword similarities," Asian Association of Open Universities Journal, Dec. 2023, doi: 10.1108/AAOUJ-02-2023-0027.

[15] E. Mayfield and A.W. Black, "Should You Fine-Tune BERT for Automated Essay Scoring?," 2020, [Online]. Available: https://course.fast.ai.