2022 2nd International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 24-26, 2022

Domain-Specific Hybrid BERT based System for Automatic Short Answer Grading

Jai Garg
Department of Applied Mathematics
Delhi Technological University
Delhi, India
jaigarg_2k18mc044@dtu.ac.in

Jatin Papreja
Department of Applied Mathematics
Delhi Technological University
Delhi, India
jatinpapreja_2k18mc049@dtu.ac.in

Kumar Apurva
Department of Applied Mathematics
Delhi Technological University
Delhi, India
kumarapurva_2k18mc058@dtu.ac.in

Dr. Goonjan Jain
Department of Applied Mathematics
Delhi Technological University
Delhi, India
goonjanjain@dtu.ac.in

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9847754

***Abstract*—Effective and efficient grading has been recognized as an important issue in any educational institution. In this study, a grading system involving BERT for Automatic Short Answer Grading (ASAG) is proposed. A BERT Regressor model is fine-tuned using a domain-specific ASAG dataset to achieve a baseline performance. In order to improve the final grading performance, an effective strategy is proposed involving careful integration of BERT Regressor model with Semantic Text Similarity. A set of experiments is conducted to test the performance of the proposed method. Two performance metrics namely: Pearson's Correlation Coefficient and Root Mean Squared Error are used for evaluation purposes. The results obtained highlights the usefulness of proposed system for domain specific ASAG tasks in real life.**

***Index Terms*—Automatic Short Answer Grading (ASAG), Semantic Text Similarity, Key-Response Similarity, Bidirectional Encoder Representation from Transformers (BERT), Masked and Permuted Pre-training for Language Understanding (MPNet)**

I. INTRODUCTION

Grading assignments and tests are an important part of any educational course. It is used as a method for assessing the level of understanding developed by an individual undertaking a particular course. No grading system is perfect but effective automation can provide numerous benefits on a large scale.

Generally, an assessment involves questions that can be classified into three categories on the basis of dependency on the reference key:

- *High Dependency*: involves questions such as MCQs, Fill-Ups and True/False.
- *Moderate Dependency*: involves questions having short answers generally from one-liners to a paragraph.
- *Low Dependency*: involves language based questions such as letter writing, essay writing and debate writing.

Modern Examinations generally involves Multiple Choice Questions due to ease of grading. In these questions, candidates are required to select one option out of some given options. It has already been shown that these types of questions are inadequate in accessing the caliber of students due to their closed ended nature [1].

Questions involving short answers can be a suitable replacement for the MCQ's if they can be graded efficiently. Short answers typically refer to one or two line responses given to a question in natural language involving free text. Short answer type questions generally require some additional information (key) along with the question for grading properly. Also, the response can be graded based on the context, making it a subjective decision rather than an objective one.

Traditionally, manual grading has been a preferred choice for assessment of student's responses. However, with increase in the number of students pursuing educational stream, workload of teachers and professors has increased many folds. Sometimes, manual grading may also introduce some bias or inconsistency due to various reasons such as involvement of more than one grader, writing style of a student and lack of appreciation of answers different from the reference key. Also, students are left deprived of accurate and timely feedback on their tests and assignments. Thus, a solution to this problem should involve an automated grading system that can provide faster results with relatively low level of bias and inconsistency while also enabling students to showcase their skills and knowledge. Automatic Short Answer Grading (ASAG) can be used to overcome these challenges. It involves a system to learn and understand high-level contextual meaning of reference key as well as a student's response for a particular question. Thus, it can be classified as Natural Language Processing task.

In this paper, the problem of Automatic Short Answer Grading is tackled using Bidirectional Encoder Representations from Transformers (BERT) on a domain-specific ASAG dataset. Mohler dataset [2] is used as ASAG dataset which contains questions and answers involving Data Structures.

BERT [3] is a transformer-based machine learning technique for Natural Language Processing (NLP) developed and pre-

978-1-6654-8407-7/22/$31.00 ©2022 IEEE

trained by Google. Unlabeled data extracted from the books with 800 million words and Wikipedia with 2,500 million words is used for pre-training. The original English-language BERT has two models:

1) *BERT Base*: 12 layers (transformer blocks), 12 attention heads, and 110 million parameters.
2) *BERT Large*: 24 layers (transformer blocks), 16 attention heads and, 340 million parameters.

It is pretrained on two tasks namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

- *Masked Language Modeling (MLM)*: BERT is designed as a deeply bidirectional model. The network effectively captures information from both the right and left context of a token starting from the first layer and all the way through to the last layer.
- *Next Sentence Prediction (NSP)*: BERT is also trained on the task of Next Sentence Prediction for tasks that require an understanding of the relationship between sentences.

As a result of the training process, BERT learns contextual embeddings for words. After pretraining, it can be finetuned with fewer resources on smaller datasets to optimize its performance on specific tasks.

The contributions of this paper are summarized as follows:

- Building an Automatic Short Answer Grading System involving a baseline fine-tuned BERT Regressor model.
- Generating Semantic Text Similarity using a modified BERT model fine-tuned for Question Answering and a Sentence Transformer model.
- Integrating BERT Regressor model with Semantic Text Similarity to further improve the performance.
- Finally, creating a model which is capable of grading a student's response for a particular question when provided with a reference key.

Thus, the system should be able to learn and understand high-level contextual meaning of both reference key as well as student response and compare them to generate a final grade in a specified numerical range.

The paper is further organized as follows: Section II contains the related work. Section III includes the approach proposed for the ASAG task. Experimental Setup including Dataset, Architecture and System Specifications is discussed in Section IV. Section V contains the results obtained from the proposed method and its comparison with different existing methods, followed by conclusion in Section VI.

II. RELATED WORK

Studies on Automatic grading has gained a lot of momentum in recent years. The need to have a reliable ASAG system is at its peak right now. The studies have been fueled further by the onset of pandemic and shift of teaching practices to online mode. Authors from around the world have proposed various techniques to solve the ASAG task. Developing an effective as well as a reliable ASAG system remains the focal center in the studies being carried out in recent years.

The study of automatic grading system was initiated by Page [4] for grading essays using computers. Since then, various contribution in this domain have been made by many authors around the world. Burrows et al. [5] researched about the existing ASAG system and classified them into 5 categories based on timeline and the techniques used. They researched about 35 ASAG systems from 1996 to 2015 which were the flag bearers of development in the field of automatic grading. Mohler et al. [2] proposed an ASAG system based on lexical semantic similarity measures. They showed that a combination with machine learning techniques is more useful in accurately predicting grades as compared to that in isolation. They also studied dependency graphs of response and key and further researched over their alignment to gain more information.

Recently, Ramachandran et al. [6] proposed a word-order graph-based study to achieve the ASAG task. They rely on finding important patterns from rubric texts and responses from highly graded students. They also explored semantic metrics to find out the synonyms to represent replacement words. Sultan et al. [7] proposed an ASAG system based on feature extraction. The features extracted included Text similarity between reference answer and student response, question demoting, term weighting and length ratio. Based on these features, regression and classification models were built according to the requirement of dataset. Wang et al. [8] introduced ml-BERT method for ASAG task. They combined BERT with meta-learning to help in initialization of the BERT parameters in a specific target subject domain using unlabeled data, thus leveraging the limited labeled training data for the grading task. Sung et al. [9] studied about improving BERT by supplementing data from resources belonging to a particular domain for ASAG task. They used multi-domain resources as datasets to perform fine-tuning. Tulu et al. [10] presented an ASAG system using sense vectors obtained from SemSpace algorithm and LSTM combined with Manhattan Vectorial Similarity. Sense embeddings of Synsets corresponding to each word in Student's answers or reference answers are given as input into parallel LSTM architecture. Finally, Manhattan Similarity is found between the text embedding of student's response and reference answer.

With the help of these prior studies, a hybrid approach consisting of fine-tuned BERT Regressor model followed by integration of key-response semantic text similarity for a particular question is proposed.

III. METHODOLOGY

The proposed ASAG system is responsible for generating a numerical grade (G) when provided with a Question (Q), a Key (K) and a Response (R).

$$(Question, Key, Response) \rightarrow Grade \tag{1}$$

The proposed system involves a hybrid approach consisting of three fundamental steps as shown in Fig. 1. In the first step (A), BERT Regressor model is fine-tuned using a domain specific ASAG dataset. It involves a BERT model followed by

a regressor model. Second step (B) involves determining the similarity between key and response for a particular question. A modified Question-Answering BERT model along with cosine similarity is proposed for this purpose. In the third step (C), the similarity characteristic obtained in second step is combined with BERT Regressor model to provide final grading results. This section provides a detailed explanation of the three mentioned steps.

A. BERT Regressor ASAG Model

A BERT base model is combined with a custom artificial neural network based regression model for the ASAG task. The architecture of BERT Regressor model is discussed in Section IV. The model is trained using a domain-specific ASAG dataset to generate the grades. In this study, objective is to generate numerical grading in a particular range. Thus, a regression model is used.

Initially, a sequence of tokens is generated for each key-response pair using BERT tokenizer. This sequence includes a classification token [CLS] at the start of sequence and a separator token [SEP] in between key and response as shown in Fig. 2. This sequence of tokens is then passed on as an input to the BERT Regressor model for training purposes. The training task involves predicting the grade in a particular range for an input sequence of tokens.

The BERT model generates a high dimensional embedding of each input sequence which is further fed to the regressor model in order to generate a grade for the key-response pair.

B. Determining Key-Response Similarity

The semantic similarity between a key and response is one of the most important features on the basis of which grading should be performed. More the semantic similarity between key and response for a particular question, higher should be the numerical grade awarded.

Evaluating similarity between a key-response pair is an important task and involves caution. A response might contain keywords present in the key, but can still lack contextual similarity. Another case may arise, where the length of response is relatively longer than the reference key but semantically correct. Thus, simple truncation of a response from an extreme end would result in an inaccurate representation of original response. In order to overcome these challenges, a filtered response is constructed from original response which is then used for finding similarity with the key.

In order to construct the filtered response, a fine-tuned BERT model for Question-Answering is used. A typical Question-Answering BERT model works by generating a start and stop value for each word in the response. The model selects the word with maximum start value as the starting word of the answer. Similarly, word with maximum stop value is chosen as the ending word while ensuring that the starting word comes before the ending word. This method of extracting answers from the response can lead to very short filtered responses and thus, result in a loss of information as compared

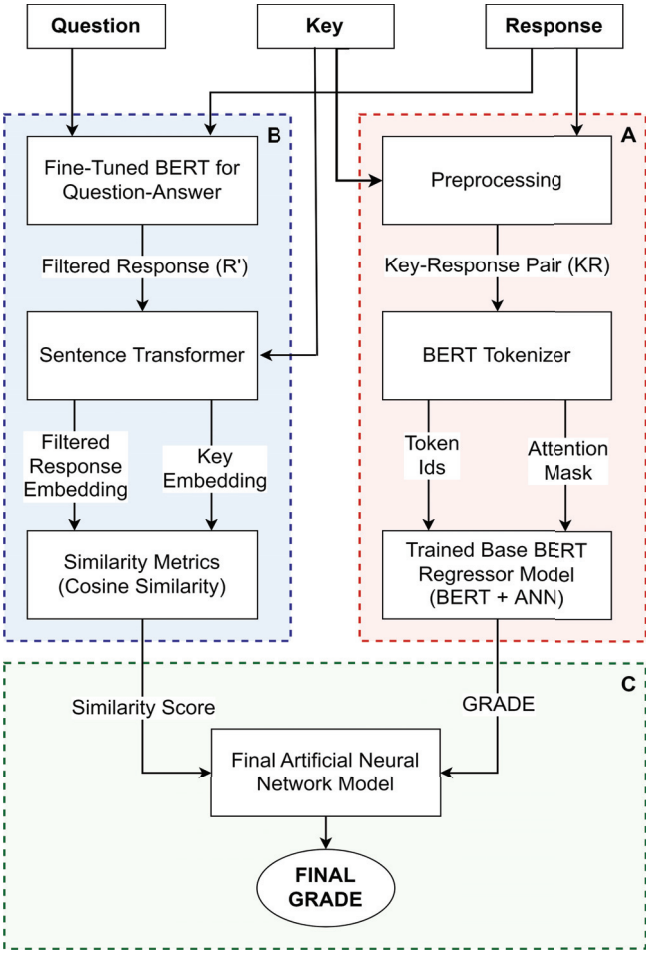

Fig. 1. Model Approach Flowchart.

Question: What is the experimental approach for measuring the running time of an algorithm?

Key: Implement the algorithm and measure the physical running time.

Student Response: running an algorithm on a specific set of data

Key-Response Pair: [CLS] Implement the algorithm and measure the physical running time. [SEP] running an algorithm on a specific set of data. [SEP]

Token IDs of Key-Response Pair without padding:
[101, 10408, 1996, 9896, 1998, 5468, 1996, 3558, 2770, 2051, 1012, 102, 2770, 2019, 9896, 2006, 1037, 3563, 2275, 1997, 2951, 102]

Fig. 2. Example of Key-Response Pair and Token ID's.

to original response. Eventually, this can result in a lower grade for even a very good response.

To overcome this problem of potential information loss, a modified technique is proposed for selecting starting and ending words. In the proposed method, starting and ending words are chosen in such a manner that length of the answer (filtered response) obtained is at least as long as the key provided. Hence, the generated filtered response contains majority of the important information required for grading purposes while eliminating any irrelevant text as shown in Fig. 3.

Question: What is the difference between a circular linked list and a basic linked list?

Key: The last element in a circular linked list points to the head of the list.

Student Response: The circular linked list's tail points to the head, whereas the basic linked list's tail points to a NULL.

Filtered Response: tail points to the head, whereas the basic linked list's tail points to a null

Key Embeddings:
[0.01993977 -0.1645347 -0.01536680 ... -0.03414393 0.01878243 -0.04168871]

Filtered Response Embeddings:
[0.01826572 -0.1475896 0.012160487 ... -0.04462539 0.04133997 -0.01269365]

Cosine Similarity Score: 0.723192

Fig. 3. Example of Filtered Response and Similarity Score.

The filtered response generated contains only the important section extracted out from a response and thus, ensures that quality is preferred over quantity. This also helps in partially achieving the task of question demoting. Filtered response and its corresponding key is fed to the Masked and Permuted Pre-training for Language Understanding (MPNet) [11] model to generate two embeddings respectively. These embeddings are then combined using a similarity metric to generate a similarity score. In the proposed approach, cosine similarity is used as similarity metric, which represents the cosine of the angle between the two embeddings when drawn in the corresponding embedding space. Equation (2) is used to calculate the cosine similarity where A and B represents the two embeddings.

$$Similarity\ Score\ (A, B) = \frac{A \cdot B}{|A||B|} \tag{2}$$

Overall, this method is used to generate a similarity score between the important information extracted from the response and the key.

C. Integration of Similarity Score with BERT Regressor

The similarity score is integrated with the BERT Regressor model using an artificial neural network. The network is then trained in order to predict the final grade. Any score outside the permitted range is rounded off to the nearest score within the range. Using similarity score as a separate feature, results in performance improvement of the proposed system as discussed in Section V.

IV. EXPERIMENTAL SETUP

A. Dataset

The proposed system is evaluated using Mohler ASAG dataset [2]. It involves generating a numerical score in range 0 to 5 for a student's response to a particular question on the basis of a reference key provided. The dataset consists of 80 questions from ten assignments and two exams based on Data Structures. It contains 2,273 student responses. The average grade of two different human graders has been considered as the final grade. The entire dataset has been divided into training and testing set with 0.2 being the test size.

B. Model Architecture

In BERT Regressor, BERT-base-uncased model is used for generation of embeddings. It generates 768-dimensional embedding vector for each input sequence of tokens. Artificial neural network-based regression model contains an input layer, 2 hidden layers, an output layer along with dropout layers to prevent overfitting as shown in Fig. 4. Rectified Linear Unit (ReLU) is used as the activation function.

Key-Response similarity method uses BERT-large-uncased model involving whole word masking and finetuning on SQuAD dataset as the Question-Answering BERT model. MPNet model is used in order to generate embeddings for filtered response and key before the application of cosine similarity. Neural network used for integration of similarity score with BERT Regressor model contains 2 hidden layers with ReLU activation function and an output layer with a linear activation function as shown in Fig. 5.

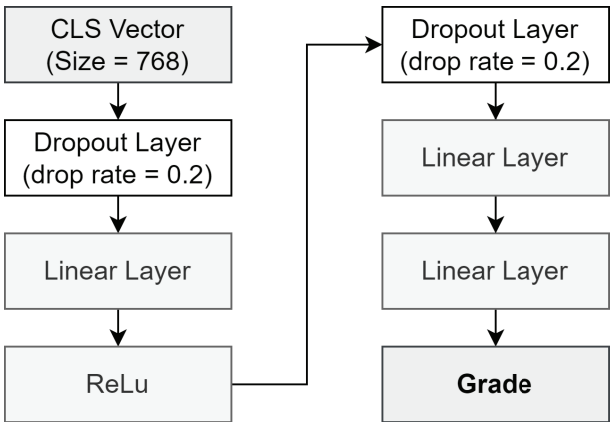

Fig. 4. BERT Regressor ANN architecture.

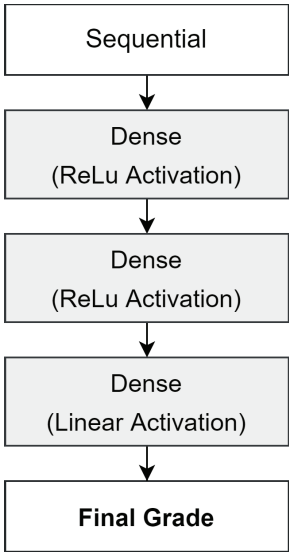

Fig. 5. ANN architecture for integration phase.

C. Parameters

BERT Regressor model is trained using AdamW (Adam with weight decay) optimizer with a learning rate of $5x10^{-5}$ and epsilon being 10^{-8}. Number of epochs is chosen to be 10 with a batch size of 32. Mean Squared Error (MSE) is used as the loss function accompanied by a linear scheduler with warmup. All other parameters retain their default values.

D. System Specifications

Google Colab is used for developing the proposed model with 2 vCPUs, GPU Tesla K80 having compute capability of 3.7 and 12 GB RAM.

V. RESULTS

A. Comparative Study

The proposed model as well as base model (BERT Regressor) is compared to the existing systems based on its performance on the test set. Pearson's Correlation Coefficient (R) and Root Mean Squared Error (RMSE) are used as two evaluation metrics. These are well known metrics for evaluation of any ASAG system as determined by various other studies. A higher Pearson's Correlation Coefficient and a lower RMSE value is desired for a good ASAG system. The results obtained using proposed approach along with other existing models are shown in Table 1. The proposed model gives Pearson's Correlation Coefficient as 0.777 and RMSE value as 0.732 on the test set.

Base model gives 0.760 as Pearson's Correlation Coefficient and 0.753 as RMSE. A clear improvement (about 2-3%) over the base model is observed in both evaluation metrics upon integration with the similarity score.

Hence, similarity score results in a more robust grading due to better understanding of the key-response pair while reducing the biasness of base model towards higher scores.

TABLE I
COMPARISON OF PROPOSED SYSTEM WITH EXISTING ASAG SYSTEMS

Models	**Pearson's R**	**RMSE**
Final Proposed Model (BERT Regressor + Similarity Score)	0.777	0.732
Base Model (BERT Regressor)	0.760	0.753
Tulu et al. (2021) [10]	0.949	0.040
Sultan et al. (2016) [7]	0.630	0.850
Tf-Idf (2016) [7]	0.320	1.020
Ramachandran et al. (2015) [6]	0.610	0.860
Mohler et al. (2011) [2]	0.518	0.978

The Proposed model performs better than most of the existing models except MaLSTM model proposed by Tulu et al [10]. The reason being that their LSTM model is trained on each assignment separately rather than on the entire dataset in one go. Thus, it narrows down the domain to each sub-topic. The difference in results between the proposed model and MaLSTM model can be justified since the proposed model is developed keeping domain generality as the primary objective.

B. Sample Data Analysis

In order to provide a more comprehensive overview of the results obtained, a random sample of six student responses are extracted from the test set. Fig. 6 shows the actual and the predicted grades of the random sample. Predicted scores almost match actual scores in majority of the questions. The RMSE for the random sample of questions comes out to be 0.32. Fig. 7 shows the result of a response to a particular question along with a key. The score given through manual grading is 5.0 and the score given by the proposed model comes out to be 4.979. This result validates the semantic similarity present between the response and the reference key.

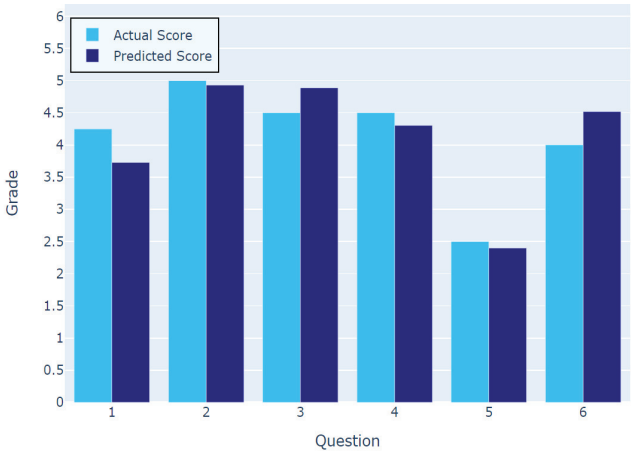

Fig. 6. Actual v/s Predicted Grade Comparison for a sample set.

Question: What does a function signature include?

Key: The name of the function and the types of the parameters.

Student Response: the function's name and parameters

Score Given By Manual Grading: 5.0

Score Given By the proposed approach: 4.9794846

Fig. 7. Example of Grading using the proposed system.

VI. CONCLUSION AND FUTURE WORK

In this paper, a 'Domain-Specific Hybrid BERT based System for Automatic Short Answer Grading' is developed. The system relies on understanding high-level contextual meaning of Question, Key and Response to perform automatic grading. It involves integration of similarity score with a base BERT Regressor model. The performance of the model is analysed using Pearson's Correlation Coefficient and Root

Mean Squared Error. Mohler's ASAG Dataset is used for experimental analysis which is widely known for benchmarking ASAG systems. The proposed model performs better than the genuinely eligible models considered for comparison by providing a higher value of Pearson's Correlation Coefficient and a lower value of RMSE. The results obtained using the proposed model highlights the usefulness of the approach in real life.

As a future study, the performance can be further improved by selecting more suitable and performance centric models with more computational power for the purpose of Question Answering, Sentence Similarity and BERT-Regressor model. Experimenting with different similarity metrics and parameter tuning of proposed models may also lead to performance gains. Implementation of classification version of proposed method for datasets such as SemEval [12] can be useful.

REFERENCES

[1] Y. Oksuz and E. Demir, "Comparison of open ended questions and multiple choice tests in terms of psychometric features and student performance," Hacettepe Univ. J. Edu., vol. 34, no. 1, pp. 259–282, 2019, doi: 10.16986/HUJE.2018040550.

[2] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol., 2011, pp. 752– 762.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[4] E. B. Page, "The imminence of grading essays by computers," Phi Delta Kappan, vol. 47, no. 5, pp. 238–243, 1966.

[5] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," Int. J. Artif. Intell. Edu., vol. 25, no. 1, pp. 60–117, Mar. 2015, doi: 10.1007/s40593-014-0026-8.

[6] L. Ramachandran, J. Cheng, and P. Foltz, "Identifying patterns for short answer scoring using graph-based lexico-semantic text matching," Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015.

[7] M. A. Sultan, C. Salazar, and T. Sumner, "Fast and easy short answer grading with high accuracy," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol., 2016, pp. 1070–1075, doi: 10.18653/v1/N16-1123.

[8] W. Zichao, S. L. Andrew, E. W. Andrew, P. Grimaldi, and R. G. Baraniuk, "A meta-learning augmented bidirectional transformer model for automatic short answer grading," in Proc. 12th Int. Conf. Educ. Data Mining (EDM), 2019, pp. 1–4.

[9] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, and R. Arora, "Pretraining BERT on domain resources for short answer grading," in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 6073–6077, doi: 10.18653/v1/D19-1628.

[10] C. N. Tulu, O. Ozkaya and U. Orhan, "Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM," in IEEE Access, vol. 9, pp. 19270-19280, 2021, doi: 10.1109/ACCESS.2021.3054346.

[11] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu, "MPNet: Masked and Permuted Pre-training for Language Understanding", Proceedings of NeurIPS (2020), pp. 16857-16867.

[12] R. Nielsen, M. Dzikovska, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, P. Clark, I. Dagan, and H. Dang, "SemEval-2013 Task 7: The joint student response analysis and 8th recognizing textual entailment challenge," Assoc. Comput. Linguistics, Atlanta, GA, USA, Tech. Rep., 2013, pp. 263–274.