Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Performance Enhancement of Automatic Short Answer Grading (Asag) Using Deep Learning

Rupal Chaudhari[1]*, Manish Patel[2], Ankur Goswami[3]

[1]Assistant. Professor, Department of Computer Engineering, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India.

[2]Associate Professor, Department of Information Technology, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India.

[3]Assistant. Professor, Department of Computer Engineering, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India.

***Abstract*:**

There has been a considerable increase in the number of blended learning courses, which has prompted interest in the question of how appropriate and beneficial automated assessment may be in such a circumstance. This is despite the fact that the impacts of the pandemic have not even been taken into consideration. In this study, we investigate the automated short answer grading (ASAG) scenario, which is a scenario that includes the application of machine learning, more specifically deep learning techniques, to grade student responses that have severe length limits. Because it offers a basis for thinking about answers that are more conversational or open-ended, ASAG continues to be one of the most active areas of study in the field of natural language processing (NLP), despite the fact that it has been investigated for more than half a century. One of the most fundamental issues that ASAG faces is the absence of sufficient training data, which includes, among other things, information that is tagged and information that is relevant to a domain. The purpose of this project is to investigate these questions using a variety of deep learning approaches. To be more specific, there is a comprehensive analysis of deep learning models, the curation of datasets, and evaluation criteria for ASAG tasks. This research comes to a close by examining the development of guidelines for educators, with the goal of enhancing the utility of ASAG research materials.

Keywords: *Automatic Short Answer Grading, Deep Learning, Automated Assessment*

I. Introduction

The amount of effort that is required to correctly mark student assessments is increasing in blended learning situations. This is due to the large student cohorts, the requirement for more comprehensive and prompt feedback, and the constraints that are placed on teaching resources. In this study, we investigate how Natural Language Processing (NLP) can assist in reducing the amount of effort associated with grading and give students with comments that are insightful.[1] Automatic interpretation of student replies is the primary focus of this research, with the goal of reducing disparities in the distribution of marks and ensuring that the final result is distributed fairly. Free text answer tests are still considered to be among the most effective techniques of grading because of the effectiveness with which they display tacit knowledge and validate talents. This is the case despite the problems that are associated with them.[2]

The use of automated assessment grading should, among other things, make the roles of teachers more manageable.[3] Because many forms of evaluation may require the application of specialist grading approaches that adhere to specific rubrics or grading criteria, natural language processing (NLP) may be proposed as a helpful analytical tool. This is especially true in the domains of science and engineering.[4] When compared to

manually evaluating student work, adopting an automated process can save instructors time. This is because any automation effort must have a well-defined and repeatable grading mechanism, and it must also have sufficient training data accessible.[5]

Page was the first person to discuss the use of computers to automatically grade natural human responses. It is important to note that his Project Essay Grade (PEG) system utilised a wide variety of natural language processing techniques.[6] The score for the essay was determined by the algorithm based on a number of criteria, including the number of words, the length of the response, and the tags for parts of speech. For the purpose of forecasting the outcome, multiple linear regression was then utilised.[7] In order to carry out surface feature analysis, the PEG system made use of syntactical similarity metrics. Since that time, a big number of researchers working in natural language processing have developed an interest in the machine grading of natural answers. An argument that the majority of the progress that has been made up to this point has been accomplished in the automatic scoring of brief human responses is a realistic conclusion to reach.[8]

Due to the fact that essays and short responses are categorized as descriptive and free text answers, it is essential to differentiate between the two in order to create solutions that are both effective and exact.[9]

Every piece of writing that fits the requirements listed below is considered to be a brief answer:

- Students are required to respond to a specific question using language that is natural to them.
- It is required that the length of an answer be restricted to between one sentence and one paragraph.[10]
- As part of their response, students are required to demonstrate the external knowledge that they have acquired as a result of their comprehension and that is not specifically stated within the question.[11]
- It is not appropriate for a response grade to be determined by subjective writing quality concerns; rather, it should be based on objective content quality criteria.
- Responses in natural language should be able to be clearly constrained depending on the syntax of the question that has been posed to them.[12]

The Automatic Short Answer Grading (ASAG) system compares student responses to one or more reference solutions for a specific topic and utilises machine learning (ML) to provide a mark [13]. This system was developed to reduce the amount of work that teachers and teaching assistants have to do. Despite the fact that automatic short answer grading is not a new concept, the state of the art for rapidly evaluating solutions that employ natural language responses has just evolved. This is despite the fact that the technology already existed. An initial interpretation of ASAG was that it was either a classification or regression task. In the classification task, a response is either recorded as correct or incorrect, whereas in the regression work, a score is assigned to the response. Additionally, the bulk of the patterns and text similarity algorithms that were utilised in those studies were created manually. This was also the case for the majority of the methods. In recent years, researchers have begun to apply deep learning approaches for ASAG. This is due to the fact that these techniques have been demonstrated to be effective in a wide variety of NLP domains and tasks.

The purpose of this study is to investigate a number of deep learning approaches that are utilised for ASAG by making use of public datasets that have been extensively cited and that compare a number of different ASAG evaluation criteria. At the same time that we provide a framework for educators who are looking for a reliable ASAG evaluation method for their students in this area, we also provide a benchmark analysis of the methods that are currently prevalent.

II. Data Sets

The majority of natural language processing (NLP) activities, such as text classification, named entity recognition, and sentiment analysis, have multiple established standard datasets that are used to evaluate the efficacy of novel methodologies and tactics. A few examples of these activities are text classification, named entity recognition, and sentiment analysis. Nevertheless, the lack of appropriate datasets presents a significant

challenge for the work that ASAG is doing. For the purpose of evaluating the effectiveness of various ASAG systems, the literature has made use of a number of benchmark datasets that are available to the general public. The scope of this study was limited to the datasets that were utilised for the purpose of benchmarking deep learning-based systems. There is a broad range of variation in the topics, sizes, and grading systems of these databases. It has been brought to people's attention that an assortment of well-known datasets, including the ASAP and SemEval-2013 datasets, were made available through competitions.[14]

Mohler's Dataset

According to Mohler and Mihalcea (2009), this dataset was made accessible in 2009 and is based on the assignments that were given in a data structures course that was taken by undergraduate students at the University of Texas. Three tasks, each consisting of seven questions, have been given to the thirty students who are enrolled in the class. As a consequence of this, the dataset contains 630 replies supplied by students. Each of these responses is graded by two human teachers in their own right on a scale that ranges from 0 to 5, with 0 representing an absolutely inaccurate response and 5 representing the correct response. Figure 1 presents a sample taken from the dataset that Mohler has created.[15]

Mohler's dataset- Assignment 1

Question: What is the role of a prototype program in problem solving?

Reference Answer: To simulate the behavior of portions of the desired software product.

Student Answer	**Average Marks**
A prototype program simulates the behaviors of portions of the desired software product to allow for error checking.	**4**

Figure 1: Example from Mohler's dataset.

2011 saw the publication of an enlarged dataset by the developers of the ASAG job assignment. In the year 2011, Mohler et al. Included in this enhanced edition are the replies that students provided to 10 different tasks as well as two different exam papers.[16] On each assignment, there are between four and seven questions, and on each exam paper, there are 10 questions. Due to the fact that there are 81 questions and 20 answers for each question, there are a total of 1620 question-answer pairs. Each response is evaluated by two different markers, and the ultimate score is determined by taking the average of the ratings that each of them received. This dataset is available to the general public.

ASAP-SAS Dataset

Quick Response for the Prize for the Automated Student Assessment programme Kaggle was the platform where the Hawlett Foundation first presented the score corpus. Less than fifty words are included in the replies, which were provided by students in grades eight through ten throughout the United States. A total of ten prompts, one for each question that is included in the dataset. The responses are rated on two distinct scales, ranging from 0 to 2, and there are a total of 17204 responses included. Moreover, the marking rubric for each and every prompt is provided in this dataset as well. The example that is displayed in Figure 2 is taken from the ASAP dataset.[17]

ASAP-SAS dataset- Prompt 3

Question: Explain how pandas in China are similar to koalas in Australia and how they both are different from pythons. Support your response with information from the article.

Scoring Rubric
2-point response: The response demonstrates: an exploration or development of the ideas presented in the text a strong conceptual understanding by the inclusion of specific relevant information from the text an extension of ideas that may include extensive and/or insightful inferences, connections between ideas in the text, and references to prior knowledge and/or experiences

0-point response: The response demonstrates: limited or no exploration or development of ideas presented in the text limited or no understanding of the text, may be illogical, vague, or irrelevant possible incomplete or limited inferences, connections between ideas in the text, or references to prior knowledge and/or experiences

Student Answers:
2-point response: According to the story both Pandas and Koalas eat only one type of food they are both specialists. Pythons are generalists meaning they can find food anywhere and eat many different kinds of food.

1-point response: Pandas in China and Koalas in Australia are both specialists. Pandas eat nothing but bamboo. The Koalas eat exclusively eucalyptus leaves. They both stick to one main type of food.

0-point response: Chinas panda bears only eat bamboo and koala bears only eat eucalyptus leaves, but pythons are able to live in more than one area.

Figure 2: Example from ASAP dataset.

SRA Dataset

This dataset was made available to the public in 2013 by the SemEval (Semantic Evaluation) workshop. Unseen Domain (UD) and Beetle (Dzikovska et al., UA) are the two subsets that make up this collection. SciEntBank (SEB) is the first subset. The UQ dataset contains questions that are inside the domain but have not yet been seen, the UA dataset contains answers to questions that are within the training dataset, and the UD dataset contains questions and answers that are not within the domain. With that being said, the Beetle test set is comprised of the Unseen Questions (UQ) and Unseen Answers (UA) subgroups.[18]

SRA- Beetle Dataset: 3-way

Question: Explain why you got a voltage reading of 1.5 for terminal 1 and the positive terminal

Reference Answers

1: Terminal 1 and the positive terminal are separated by the gap

2: Terminal 1 and the positive terminal are not connected

Student Answer	**Label**
Because there is a gap between terminal one and the positive battery terminal	**correct**
The terminal is connected to a positive circuit	**contradictory**
It was separated by a gap.	**incorrect**

Figure 3: Example from SRA dataset.

III. Evaluation Metrics

Different evaluation criteria are utilised, and these criteria vary according to whether an ASAG system is intended as a classification model or a regression model. The purpose of this section is to provide an overview of numerous performance indicators that are commonly used for evaluating ASAG models.[19]

Pearson's r Correlation

Utilising this method, one can determine the degree of correlation that exists between two numerical variables. This method assigns values that range from -1 to 1, where a value of 1 indicates a positive correlation, a value of 0 indicates that there is no connection, and a value of -1 indicates a negative correlation.[20] Equation 1 is used to compute it.:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{1=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad \ldots\ldots(1)$$

Where for two distributions X and Y, x_i and y_i are the i^{th} value of distributions and $\bar{x}$ and $\bar{y}$ are the mean values for both distributions respectively

When it comes to the ASAG job, one of the correlation metrics that is utilised the most frequently is Pearson's correlation. This particular metric is utilised to compare the marks that instructors assign with the marks that are anticipated.[21]

Root Mean Square Error

In order to determine the error value that exists between the predicted values and the observed values, a separate measure is utilised. RMSE is the score that is calculated by Equation (2).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_p - x_o)^2} \quad \ldots\ldots\ldots(2)$$

Where x_p is predicted value and x_o value that is noticed is the value. A performance that is superior is shown by a lower RMSE number.
A performance that is superior is shown by a lower RMSE number.

F1 Score

Precision and recall are both components of this evaluation metric for categorization, which combines the two. It goes from 1 (the best) to 0 (the worst) and is calculated by taking the weighted average of precision and recall information.

$$F1 = \frac{2*(Precision*Recall)}{(Precision+recall)} \quad \ldots..(3)$$

One way to quantify precision is as the proportion of the model's accurate predictions to the total number of forecasts taken into account.

$$Precision = \frac{TP}{TP+FP} \quad \ldots..(4)$$

Recall: it is the ration of correct predictions made by the model to the actual labels.

$$Recall = \frac{TP}{TP+FN} \quad \ldots\ldots(5)$$

This equation is used to determine the average F1 score for each class. It is referred to as the large-scale average F1.

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=0}^{N}(F1\ score)_i \quad \text{……..(6)}$$

The harmonic mean of the recall and precision for each class is represented by the micro-average F1, which is pronounced as F1.

$$\text{Micro-F1} = \frac{2*(Micro\ Precision*Micro\ Recall)}{(Micro\ Precision+Micro\ Recall)} \quad (7)$$

$$\text{Micro Precision} = \frac{\sum TP_i}{\sum TP_i + \sum FP_i} \quad (8)$$

$$\text{Micro Recall} = \frac{\sum TP_i}{\sum TP_i + \sum FN_i} \quad (9)$$

Quadratic Weighted Kappa

Obtaining the inter-rater agreement between the scores that were predicted and those that were forecasted is one way to compute this statistic.

$$k = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \quad \text{…….(10)}$$

Where, p_0 is the observed agreement and p_e refers to the consent that is expected. The value of this variable is 1 when the expected and projected scores are identical; when they are not identical, the value is 0; and so on.

IV. Deep Learning Approaches For Asag

Measurement of text similarity has been a traditional use of machine learning techniques for a long time. These techniques have demonstrated their potential in a variety of natural language processing applications, including automatic scoring, machine translation, and text summarization. In earlier iterations of the ASAG models, feature engineering and fundamental text vectorization techniques were combined with conventional text similarity techniques, such as string-based and corpus-based similarity standards. Nevertheless, the manual generation of features through the use of regular expressions is a difficult process that requires thorough understanding of the subject matter. Additionally, due to the fact that these systems are trained on high-dimensional sparse vectors, they require a significant amount of processing resources. When it comes to tasks that are based on text similarity, such as Automatic Essay Scoring and Automatic Short Answer Grading, deep learning architectures have become increasingly popular in 2016. This is in contrast to traditional machine learning techniques, which have been somewhat less popular. In addition, a number of authors looked into the application of deep learning strategies to word embedding models. Early Neural Network-based, Attention-based, and Transformer-based ASAG designs are the three categories that are utilised in this research to organise the work that has been done in this sector respectively.[22]

Attention-based Approaches

One of the most significant advancements that has taken place in the field of Deep Learning over the past 10 years is the introduction of the attention mechanism. Since its inception, natural language processing (NLP) has undergone a number of significant advancements, one of which is the implementation of transformers in various programmes, such as Google's BERT (Bidirectional Encoder Representations from Transformers). Attention is designed to construct the context vectors that are required by the decoders by taking into account all of the intermediate encoder states. This is the primary objective of attention. Through the process of extracting semantic information from student and model responses, Liu et al. (2019) presented an attention-based strategy that does away with the need for feature engineering. Through the utilisation of transformer layers and a

multiway attention mechanism, this method provides a more comprehensive comprehension of the semantic relationship that exists between the words that make up a sequence. During the testing of this system, a K–12 dataset was utilised. Within the scope of a paper that was published in 2019, the authors described yet another attention-based deep model for ASAG. For the purpose of learning phrase vector representations of student and model responses, this model makes use of BiRNN with LSTM and attention mechanism in conjunction with pre-trained word embeddings. The technique showed a 10% gain in performance compared to baseline models when it was examined on datasets containing K–12 data examples. This study was quite similar to the one described above. The researchers took advantage of a BiLSTM neural network in order to incorporate attention mechanisms and Google Word Vector (GWV)3 into this design model. Within the realm of brief response scoring, this model produces an actual average QWK value of 0.70, which is distinct from the results of a great number other baseline models.[23]

V. Discussion

For the majority of ASAG systems, it is necessary to have a "deep" learning model that consists of numerous training layers and a training phase that is performed on a corpus that is sufficiently representative of a wide cross section of both great and poor sample responses. There are a number of factors that will determine the effectiveness of an ASAG system. Some of these factors include the efficiency of the system training phase, the processing time allotted to infer the final scores, the incorporation of deep layer fine-tuning on specific questions during the training phase, scalability, and the ability to regenerate the results on datasets that are comparable but distinct. Deep learning models have been shown to perform better than standard feature engineering-based ASAG systems in terms of accuracy, semantic similarity, computational cost, and generalizability. This has been proved through tests. According to the models that were evaluated in this work, attention-based and BERT-based models perform better than alternatives for the ASAG task on the Mohler, ASAP-SAS, and SRA datasets. Furthermore, with the current improvements in the transformer and pre-trained language model literature, it is possible to attain remarkable performance in terms of efficacy. It has been established that transfer learning mechanisms, for instance, can reduce the requirement for large-scale training data while still preserving domain specificity.[24]

Furthermore, it has been documented that transfer learning mechanisms can also provide effective generalisation adaptations. The findings presented in Table 1 provide evidence that simple LSTM-based models are computationally efficient and produced satisfactory outcomes when applied to Mohler's dataset. Through the utilisation of a distinct training file for each and every question, it was proved that it was possible to get the highest possible Pearson score (with this corpus) of 0.94. On the other hand, when the system is trained with a single training file that includes all of the questions, responses, and reference answers, the Pearson correlation drops to 0.15, which is a considerable decrease from the previous value of 0. It would appear that when a significant number of Out-Of-Vocabulary (OOV) terms are present in the dataset, an ASAG system demonstrates a significantly lower success rate and trains significantly more slowly. This is the case when the system is extended to a large dataset, which is accompanied by an increase in the context training set. In light of this, it is abundantly evident that the concept of utilising LSTM algorithms in this particular use case is negatively impacted. In this particular use instance, The syntactic and semantic links between words are preserved through the utilisation of these approaches, which make use of the embeddings associated with student and reference answers. Table 2 demonstrates that an attention-based model performs better than an LSTM when applied to the ASAP-SAS dataset with respect to performance. In addition, the vast majority of attention-based models that have been evaluated with the K–12 dataset have demonstrated even more favorable results.[25]

This provides useful and implementable insights regarding the process of data curation for subsequent experiments including mixed learning. For the purpose of further improving performance for the ASAG use case, transformer-based models, which can be shown in Table 3 and which have been trained and tested on the SemEval dataset, have consistently been utilised. The vast majority of these models have utilised BERT and the fine-tuning process that is associated with its variations. Additionally, it is pretrained on Wikipedia and a

substantial book corpus in order to acquire the knowledge necessary to comprehend the contextual link between phrases, which is essential for text similarity-based tasks such as ASAG. Despite the fact that the majority of these systems are capable of producing good results, the generalised performance of these systems is still questionable when it comes to the case of entirely new student cohorts. These students may provide answers on new question banks that are significantly different from the (limited) datasets that have been considered in this work.

Table 1: Performance scores for Mohler's Dataset

Model	**Approach**	**Evaluation Score**	
		Pearson's Correlation	**RMSE**
(Kumar, Chakrabarti, & Roy, 2017)	Neural Network based	0.649	0.830
(Hassan, A, & El- Ramly, 2018)	Neural Network based	0.569	0.797
(Prabhudesai & Duong, 2019)	Neural Network based	0.655	0.883
(Gonna &Fahmy, 2020)	Neural Network based	0.63	0.91
(Tulu, Ozkaya, & Orhan, 2021)	Neural Network based	0.949'	0.040

Table 2: Scores of performances for the ASAP-SAS Dataset

Model	**Approach**	**Evaluation Score**
		QWK k
(Riordan, Horbach, Cahill, Zesch, & Lee, 2017)	NN	0.743
(Xia, Guan, Liu, Cao, & Luo, 2020)	Attention Based	0.70

Table 3: Ratings of performance for the Seme Val dataset

		Evaluation Score				
Model	**Approach**	**Pearson's Correlation**	**RATSE**	**Macro-Fl**	**Weighted-Fl**	**Accuracy**

(Kumar et al, ,2017)	NN	0.554	0.758	-	-	-
(Riordan, Hotbach, Cahill, Zesch. & Lee, 2017)	NN	-	-	-	0.791	-
(Saha et al.2019)	NN	-	-	0.798	0.803	-
(Gomm & Fatany, 2020)	NN	-	-	-	0.58	-
(Sum& Diumecha. & Mukhi, 2019)	Transformer based	-	-	0.720	0.758	-
(Carpus & Filighera, 2020)	Transformer based	-	-	0.791	0.197	0 797
(Lun, Zhu, Tang, & Yang,2020)	Transformer based	-	-	0.522	0.826	0.827
(Ghavidel Zouag , Desmarais, J Eisenstein 2020)	Transformer based	-	-	0.007	0.723	0.726

VI. Conclusion

The purpose of this research is to provide a concise overview of the most cutting-edge deep learning-based techniques to Automatic Short Answer Grading and to discuss the feasibility of implementing these approaches in educational settings. Additionally, we discussed the primary limitations of the benchmark datasets and assessment measures that are currently available to the public. Our research has shown that models that are based on transformers and transfer learning, which was just recently accepted, perform significantly better than models that were based on neural networks, which were previously used for ASAG. In spite of this, there is still something that needs to be done in order to investigate the utilisation of the most modern transformer-based models, such as GPT-2, GPT-3, T5, and XLNET. As a result of this research, a number of potential future routes have been identified, each of which has the potential to present an obstacle to universal implementation. The adoption of ASAG systems has been made easier by bringing attention to the many interests of stakeholders, the requirement of new norms for the curation of datasets, and the pressing requirement for interfaces that are more user-friendly.

References

[1] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2023). Learning phrase representations using RNN encoder-decoder for statistical machine translation. Preprint retrieved from http://arxiv.org/abs/1406.1078

[2] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2022). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. Preprint retrieved from http://arxiv.org/abs/1810.04805

[3] Drolia, S., Rupani, S., Agarwal, P., & Singh, A. (2022). Automated essay rater using natural language processing. International Journal of Computer Applications, 163(10), 44–46.

[4] Eisenstein, J. (2020). Introduction to natural language processing. Adaptive Computation and Machine Learning series. MIT Press, London. https://books.google.com.co/books?id=72yuDwAAQBAJ

[5] Gomaa, W. H., & Fahmy, A. A. (2021). Tapping into the power of automatic scoring. In The Eleventh International Conference on Language Engineering, Egyptian Society of Language Engineering (ESOLEC).

[6] Gong, T., & Yao, X. (2020). An attention-based deep model for automatic short answer score. International Journal of Computer Science and Software Engineering, 8(6), 127-132.

[7] Hassan, S., A, A., & El-Ramly, M. (2021). Automatic Short Answer Scoring based on Paragraph Embeddings. International Journal of Advanced Computer Science and Applications, 9(10).

[8] Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451.

[9] Kumar, S., Chakrabarti, S., & Roy, S. (2022). Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading. Paper presented at the IJCAI.

[10] Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G. Y., & Liu,Z. (2019). Automatic short answer grading via multiway attention networks. Paper presented at the International conference on artificial intelligence in education.

[11] Lun, J., Zhu, J., Tang, Y., & Yang, M. (2020). Multiple data augmentation strategies for improving performance on automatic short answer scoring. Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

[12] Madnani, N., & Cahill, A. (2021). Automated scoring: Beyond natural language processing. Paper presented at the Proceedings of the 27th International Conference on Computational Linguistics.

[13] Mohler, M., Bunescu, R., & Mihalcea, R. (2021). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. Paper presented at the Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies.

[14] Mohler, M., & Mihalcea, R. (2020). Text-to-text semantic similarity for automatic short answer grading. Paper presented at the Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2020).

[15] Page, E. B. (2021). The Imminence of... Grading Essays by Computer. The Phi Delta Kappan, 47(5), 238-243.

[16] Prabhudesai, A., & Duong, T. N. B. (2021, 10-13 Dec. 2021). Automatic Short Answer Grading using Siamese Bidirectional LSTM Based Regression. Paper presented at the 2019 IEEE International Conference on Engineering, Technology and Education (TALE).

[17] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[18] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

[19] Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. (2017). Investigating neural architectures for short answer scoring. Paper presented at the Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications.

[20] Saha, S., Dhamecha, T. I., Marvaniya, S., Foltz, P., Sindhgatta, R., & Sengupta, B. (2019). Joint multi-domain learning for automatic short answer grading. arXiv preprint arXiv:1902.09183.

[21] Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., & Sengupta, B. (2023). Sentence level or token level features for automatic short answer grading?: Use both. Paper presented at the International conference on artificial intelligence in education.

[22] Sasi, Nair, D., & Paul. (2020). Comparative Evaluation of Pretrained Transfer Learning Models on Automatic Short Answer Grading. arXiv pre-print server.

[23] Sultan, M. A., Salazar, C., & Sumner, T. (2021). Fast and easy short answer grading with high accuracy. Paper presented at the Proceedings of the 2020 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[24] Sung, C., Dhamecha, T. I., & Mukhi, N. (2023). Improving short answer grading using transformer-based pre- training. Paper presented at the International Conference on Artificial Intelligence in Education.

[25] Surya, K., Gayakwad, E., & Nallakaruppan, M. (2022). Deep learning for short answer scoring. Int. J. Recent. Technol. Eng.(IJRTE), 7(6).