Automated Essay Grading of Constructive Response Test Responses for Mechanical Engineering Students

Do Tien Dung
School of Environment and Society
Tokyo Institute of Technology
Meguro-ku, Tokyo, Japan
do.t.aa@m.titech.ac.jp

Hideki Mima
Promoting Organization for Future Creators
Kyushu University
Nishi-ku, Fukuoka, Japan
mima.hideki.483@m.kyushu-u.ac.jp

Farid Triawan
Department of Mechanical Engineering
Sampoerna University
Pancoran, Jakarta, Indonesia
farid.triawan@sampoernauniversity.ac.id

Jeffrey Scott Cross
School of Environment and Society
Tokyo Institute of Technology
Meguro-ku, Tokyo, Japan
cross.j.aa@m.titech.ac.jp

***Abstract*— Machine learning education related applications have increased with the appearance of large language models. While automatic essay grading (AEG) has been studied extensively in the past, most of these studies have focused on evaluating English competence instead of assessing knowledge competence in an engineering field. This study aimed to develop an AEG model to evaluate student's mechanical engineering Constructive Response Test (CRT) question responses which were instructor graded. Because of the small number of student responses (45), a synthesized set of responses was also generated by using text-to-text paraphrasing models. A neural network grading engine was built and trained to assess comprehension utilizing the Bidirectional Encoder Representation Transformer (BERT) and related models on student and synthesized responses. This study showed that the AEG based Natural Language Processing (NLP) model showed high accuracy and a higher degree of consistency in grading student responses compared to instructor-graded responses.**

Keywords—Automatic Essay Grading, Constructive Response Test, Mechanical Engineering

I. INTRODUCTION

In engineering education, evaluating students' knowledge using constructive response test (CRT) questions is critical for assessing their problem-solving skills, communication skills, and critical thinking abilities. These types of questions require students to solve real-world problems and effectively communicate their ideas and solutions ("Think-like-an - engineer") [1]. CRTs promote critical thinking and creativity, which are essential skills for success in the engineering profession. Compared to multiple-choice questions, constructive response tests provide a more accurate assessment of a student's understanding and ability to apply their knowledge in a real-world context. However, grading student's CRT responses requires a lot of time and effort for instructors compared to grading multiple choice question answers. The long time needed by instructors to grade standardized CRT responses is a barrier for large scale engineering testing in general.

Automatic essay grading (AEG) is a process in which software or algorithms are used on scored responses to evaluate and grade written responses. The technology behind AEG has been developed over several decades and is becoming increasingly sophisticated, with many different approaches and methods being used to evaluate essays. One of the key benefits of AEG is that it can provide quick and more consistent results and provides students with prompt feedback.

Several AEG models have been developed to evaluate English language proficiency, including Project Essay Grade [2], E-rater [3], and Intelligent Essay Assessor (IEA) [4]. These models use a variety of features such as grammar, coherence, organization, and word usage to score essays, but their performance and effectiveness may vary depending on the specific dataset and domain being evaluated.

In addition, large language models are pre-trained on massive amounts of text data and can be fine-tuned for specific question-answering tasks. Among these models, a widely recognized example is BERT, which is a pre-trained transformer-based model capable of being fine-tuned for tasks related to text similarity [5].

This research aims to explore the benefits of using an AEG system to assess Mechanical Engineering CRT student responses. The hypothesis of this research is that a finely tuned AEG model is more accurate and consistent than instructor graded responses. Furthermore, it can be used to assess students in the Mechanical Engineering major if provided with a sufficient and appropriate amount of data for training.

To achieve the research objectives, the following tasks were performed:

(1) Build and train a Neural Network Grading Engine model (Grading Model) using a student response dataset from the ME Tuning Item Test Bank CRT responses.

(2) Evaluate if the AEG CRT is more reliable and consistent in comparison with instructor-graded responses.

(3) Provide recommendations on which is the most efficient and accurate AEG model for CRT assessment.

II. LITERATURE REVIEW

A. Automatic Essay Grading Models

AEG has a rich history dating back to the 1960s when researchers began developing computer programs that could score essays. Early AEG systems were rule-based models, relying on a set of predetermined rules and algorithms to evaluate essays. The first person to come up with this idea was Ellis B. Page with Project Essay Grader [6].

Other models, such as Support Vector Machines (SVMs), grade essays by establishing a hyperplane that maximizes the margin between classes of grade-related data [7]. Latent Semantic Analysis (LSA) analyzes essays based on their semantic content and predict their grades accurately [8]. Since 2016, the neutral network has become a popular method with impressive results, as demonstrated by studies [9], [10], [11]. Regarding large language models, BERT (Bidirectional Encoder Representations from Transformers) [5] showed similar results to many other neutral networks AEG but gives better results when input responses are preprocessed by removing stopwords [12].

B. Data Augmentation

Data augmentation is a technique used to avoid overfitting and improve model generalization by modifying the original data while preserving its content. Common methods include Synonym Replacement, Random Insertion, Random Deletion, Random Swap, and Back-Translation [13]. In addition, using the text generative learning model as automatic data augmentation achieves superior performance to the manual methods [14]. One of the text-generative models, ChatGPT, is said to be very promising for generating fresh and innovative sample data [15].

C. Data Preprocessing

Data preparation is an important phase before applying any machine learning algorithms [16]. Some popular data preprocessing methods are removing stopwords, lemmatization, and part of speech (POS) identification.

III. Methodology

A. Mechanical Engineering Questions

The CRT used in this research is the Wind Power Generation set of questions for bachelor students majoring in Mechanical Engineering from the ME Testbank [17]. The selected questions include 3 component questions. Each component question identifies the skills on which it assesses the student within the Mechanical Engineering Competency Framework. From there, a grading rubric that includes student response requirements, and grading scales (Fig. 1) is designed to aid the human instructor in the grading process. Each question has 3 rating levels including 0 (unsatisfactory), 1 (satisfactory), and 2 (highly satisfactory).

B. Paticipants

Participants in the study included 5 ME instructors and 45 bachelor students majoring in Mechanical Engineering at 3 universities in Indonesia (15 students for each university) denoted **S, I**, and **U** institutions in this study.

Students answered questions using an online learning management system via a computer within 1 hour. At the end of the test, the student responses were graded by instructors at the student's institution. **S** and **U** had 2 instructors, while **I** had only one instructor. Therefore, at **S** and **U** institutions, each instructor gives an independent score based on scoring rubrics, if the scores for a response are different, the two instructors would discuss the grade with each other to give the final score for the response.

Q1(1)	Viewpoint	(0 points)	Level 0 (0 points)	Level 1 (1 point)	Level 2 (2 points)	Comments
		Either no answer is provided or the answer makes no sense.	The answer is unsatisfactory because it has failed to address many of the question's key points.	The answer is satisfactory because a majority of the question's key points have been properly addressed.	The answer is highly satisfactory because all or nearly all key points have been properly addressed.	
① EGS2	• The length of the answer is within the range of approximately 50 to 100 words. • The answer is logically organized.					
④ EA2	• The answer states that the wind power generation turbine has a lower moment of inertia about its rotor axis than the traditional windmill. • The answer provides reasons why wind power generation turbines have a lower moment of inertia (e.g., narrower, hollow, more tapered and fewer rotor blades composed of lighter materials). • The answer states that the wind power generation turbine's lower moment of inertia has the advantage of making it more adaptable to changes in blade rotation speed and wind velocity. Describing other advantages may also be acceptable.					

Fig. 1. Scoring guidance rubric example.

C. Data Augmentation

The T5 Paraphraser model is designed for text-to-text transformations and has been fine-tuned using the ChatGPT paraphrases corpus. It stands among the top-performing paraphrasing models for short texts. The data synthesis process utilizing the T5 model involves several steps. Initially, a baseline classification model is created by fine-tuning BERT on the master data. This classification model helps filter out irrelevant data. Next, the student's responses are segmented into smaller parts, each containing a single sentence. The T5 model then paraphrases each of these sentences independently. These paraphrased sentences are eventually combined to form complete synthesized data sets. For every score level, the process generates 500 synthesized essays, from which the baseline model selects and ranks the top 300 essays for each score level.

ChatGPT, a text generation large language model, has garnered significant attention due to its ability to provide answers across various domains. Previous research has explored utilizing ChatGPT as a data augmentation tool by directly requesting responses in its prompts. However, this approach gives incorrect augmentation results deal to a lack of domain knowledge [15]. Therefore, this study proposes a different approach, using fine-tuned ChatGPT model to synthesize data. The master data undergoes processing to be converted into JSON format, comprising two parts: the "prompt", which includes the essay score, and the "completion", which represents the corresponding essay. Once the fine-tuning process is completed, the model generates 300 synthesized essays for each scoring level.

D. Data Preprocessing

The research made use of the stopword list, WordNetLemmatizer, and POS TAG from the Natural Language Toolkit. Part of the Speech Tag would be added to the word by an underscore.

E. Model Training

Student responses are used for each question are used data augmentation process to generate an appropriate amount (900 synthesized responses) (300 synthesized data for each scoring level). This 900 synthesized data then goes through data preprocessing. Processed data is fine-tuned with a pre-trained model (BERT, distilBERT, or RoBERTa) using 5-fold cross-validation to create an AEG model for that component question.

F. Model Evaluation

The assessment grades provided independently by four instructors from **S** and **U** (each having two instructors) universities serve as a metric for measuring inter-rater reliability between human instructors. Additionally, the final scores from all three universities, alongside the predictions made by the model, are utilized to quantify the inter-rater reliability between the computer and human instructor. The computation of inter-rater reliability relies on the application of Cohen's kappa score or Quadratic Weight Kappa (QWK, κ) as indicated by Eq. (1).

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_{ij}m_{ij}} \quad (1)$$

where κ is quadratic weight kappa score, k is the number of data samples, ω is a weighted matrix, m is the confusion matrix calculated by chances, and x is the confusion matrix.

QWK scores are utilized for assessing the level of concordance between the scores provided by human instructors and those predicted by the models. The results of these evaluation methods range between 0 and 1, with 1 indicating a complete alignment between the system's predicted scores and the human instructor's scores, while 0 signifies a random correspondence between the system's predicted scores and the human instructor's scores. When interpreting the κ value, a value of less than 0.4 is indicative of poor agreement, κ between 0.4 and 0.75 suggests fair to good agreement, and κ greater than 0.75 implies excellent agreement (Cheon, 2015).

IV. Results and Discussion

A. Student responses and scores data

A total of 45 data points were received (15 from each university). For each question, invalid responses, such as non-attempt responses and non-English responses, were discarded. This leaves with three valid responses for analysis, resulting in between 29 and 37 usable data points (Table I). Of the three institutions, **I** university provided the lowest amount of valid data, mainly because some students did not answer the selected component questions, and two students responded not in English. The score distribution for the three component questions appears to be relatively even, though not perfectly balanced. There is not a significant gap between the number of students scoring at different levels.

TABLE I. Student Response Analysis

Component Question	Valid responses by score level by instructor			Valid responses by University instructor			Valid responses number
	0	1	2	S	I	U	
1	12	11	14	15	8	14	37
2	13	15	8	13	8	15	36
3	8	10	11	13	7	9	29

B. Data Augmentation Comparison

In Table II, QWK is calculated for each synthesized model. The findings indicate that the T5 model outperforms the fine-tuned ChatGPT. The T5 model outperformed the anticipated accuracy benchmarks, surpassing 0.8 for accuracy and 0.75 for the QWK index. This signifies that the fine-tuned AEG model, using data synthesized by the T5 model, demonstrates outstanding performance, and closely resembles the human-grade final assessment. Fine-tuned ChatGPT yielded only moderate performance, achieving a score of 0.57.

TABLE II. Data Augmentation Comparison

Synthesized model	Accuracy	QWK
T5	0.95 (Excellent)	0.95 (Excellent)
Fine-tuned ChatGPT	0.67 (Medium)	0.57 (Good)

C. Data Preprocessing Comparison

Using stopwords removal and lemmatization alone did not affect the AEG model's performance. Utilizing part-of-speech (POS) tags increased the overall accuracy score but reduced the QWK score, as demonstrated in Table III. This indicates that while the POS method achieved more accurate results, it also led to more substantial errors, such as misclassifying some level 0 responses to level 2. However, when combining lemmatization with the part-of-speech method, both accuracy and QWK scores surpassed other approaches.

Hence, the data structure simplification methods did not positively impact the BERT model's performance. On the contrary, preprocessing methods such as POS, and POS + Lemmatization enhanced the model accuracy and increased QWK value.

TABLE III. Data Preprocessing Comparison

Preprocessing Method	Accuracy	QWK
No Preprocessing	0.92 (Excellent)	0.94 (Excellent)
Remove Stopwords	0.92 (Excellent)	0.94 (Excellent)
Lemmatization	0.92 (Excellent)	0.94 (Excellent)
POS	**0.95 (Excellent)**	0.90 (Excellent)
Lemmatization + POS	**0.95 (Excellent)**	**0.96 (Excellent)**

D. Instructor and Model inter-rater reliability comparison

Across all three questions, the Kappa Score (κ) consistently showed greater agreement between the human final score and the model prediction score compared to the score between the two human instructors. Instances of disagreement between the two instructors grade were relatively frequent. For Question 1.1, there were 8 cases out of 29; for Question 2.1, there were 10 cases out of 28; and for Question 2.2, there were 6 cases out of 22. Conversely, the model had significantly lower disagreement with 2 out of 37, 2 out of 36, and 1 out of 29 for the respective questions, as shown in Table IV. This highlights the model's superior consistency over instructor graded responses.

TABLE IV. INSTRUCTOR VS MODEL INTER-RATER RELIABILITY COMPARISON

Question	Instructor 1 vs Instructor 2		Real grade vs Model prediction	
	Accuracy	κ	Accuracy	κ
1	21/29	0.5944 (Good)	35/37	0.9181 (Excellent)
2	18/28	0.4285 (Good)	34/36	0.9144 (Excellent)
3	16/22	0.5975 (Good)	28/29	0.9479 (Excellent)

E. Training, Prediction Time, and Accuracy

Regarding training time efficiency shown in Table V, the distilBERT model demonstrates faster training compared to the other two models. It achieves 53% to 60% of the BERT's training time. On the other hand, RoBERTa exhibits a similar training time to BERT (98% to 106%).

DistillBERT gave the most consistent prediction results e.g. from 40% to 89% compared to when using BERT to make the prediction. In between these two models is the RoBERTa which was 55% to 94% longer than the BERT.

All three models across all three component questions achieve high accuracy levels (93.64% - 96.95%). The difference in accuracy between the models is not significant. distilBERT performs about 1% less than BERT, while RoBERTa achieves approximately 0.2% higher accuracy than BERT.

TABLE V. TRAINING TIME, PREDICITON TIME, ACCURACY COMPARISON

Model	Training Time	Prediction Time	Accuracy
BERT	2.50 - 2.83 s/response	0.54 - 1.05 s/response	94.44% - 96.87%
DistilBERT	**1.42 - 1.58 s/response**	**0.39 - 0.48 s/response**	93.64% - 95.63%
RoBERTa	2.45 - 2.91 s/response	0.51 - 0.58 s/response	**94.78% - 96.95%**

V. CONCLUSION

This study compared several techniques for generating CRT responses and AES with instructor graded responses in the field of mechanical engineering. The phases encompass data augmentation, data preprocessing, and data training. The results indicate that the T5 model surpasses the finely tuned ChatGPT in terms of data synthesis. The most optimal performance in Data Preprocessing is attained through a fusion of Lemmatization and Part of Speech techniques. When contemplating the AEG model, distilBERT is recommended due to its energy-efficient training and prediction procedures. However, for inquiries necessitating utmost precision, the investigation proposes the utilization of the RoBERTa model. The application of AEG to evaluate CRTs responses in the Mechanical Engineering realm with a restricted number of data samples is feasible. Furthermore, the research assesses the inter-rater reliability between human instructors and the model which reveals that Automated Essay Grading demonstrates greater consistency compared to instructors graded responses. The authors believe this approach can also use for AEG of CRTs in other fields.

ACKNOWLEDGMENT

We would like to express our gratitude to the 45 student participants and 5 instructors from three universities who dedicated a significant amount of time to participate in the test. We also extend our thanks to the members of the Cross Laboratory at Tokyo Tech for their support, as well as to the Japan ME Tuning Item Test Bank for providing question rubrics, scoring guidance, and assisting in the collection of data samples.

REFERENCES

[1] J. S. Cross *et al.*, "Development of a Mechanical Engineering Test Item Bank to promote learning outcomes-based education in Japanese and Indonesian higher education institutions," *Tuning Journal for Higher Education*, vol. 5, no. 1, pp. 41–73, Nov. 2017.

[2] Page, E. B. Project Essay Grade: PEG. In M. D. Shermis and J. Burstein (Eds.), "Automated essay scoring: A cross-disciplinary perspective", (pp.4354). Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[3] Yigal Attali and J. Burstein, "Automated Essay Scoring With e-rater® V.2," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, Feb. 2006.

[4] P. W. Foltz, L. A. Streeter, K. E. Lochbaum, and T. K. Landauer, "Implementation and applications of the intelligent essay assessor," in *Handbook of automated essay evaluation*, M. Shermis and J. Burstein, Eds. New York, NY, USA: Routledge, 2013, pp. 68–88 .

[5] 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

[6] E. B. Page, "The imminence of... grading essays by computer," *The Phi Delta Kappan*, vol. 47, no. 5, pp. 238-243, 1966.

[7] Z. Ke, H. Inamdar, H. Lin, and V. Ng, "Give me more feedback II: Annotating thesis strength and related attributes in student essays," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3994-4004, July 2019.

[8] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259-284, 1998.

[9] A. Shehab, M. Elhoseny, and A. E. Hassanien, "A hybrid scheme for automated essay grading based on LVQ and NLP techniques," in *2016 12th International Computer Engineering Conference (ICENCO)*, pp. 65-70, IEEE, December 2016.

[10] S. K. Kopparapu and A. De, "Automatic ranking of essays using structural and semantic features," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 519-523, IEEE, September 2016.

[11] F. Dong and Y. Zhang, "Automatic features for essay scoring–an empirical study," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1072-1077, November 2016.

[12] P. U. Rodriguez, A. Jafari, and C. M. Ormerod, "Language models and automated essay scoring," 2019, *arXiv:1909.09482*. [Online]. Available: http://arxiv.org/abs/1909.09482

[13] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of Big Data*, vol. 8, pp. 1-34, 2021.

[14] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, ..., and N. Zwerdling, "Do not have enough data? Deep learning to the rescue!" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 7383-7390, April 2020.

[15] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, ..., and X. Li, "AugGPT: Leveraging ChatGPT for Text Data Augmentation," 2023, *arXiv:2302.13007*. [Online]. Available: http://arxiv.org/2302.13007

[16] V. Kalra and R. Aggarwal, "Importance of Text Data Preprocessing & Implementation in RapidMiner," in *Proceedings of the First International Conference on Information Technology and Knowledge Management*, vol. 14, pp. 71-75, 2017.

[17] "Mechanical engineering university student competence test bank," ME test bank. https://www.me-testbank.org/ (accessed Aug. 20, 2023).