Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

A Study on a Deep Learning-Based Approach for Automated Scoring Solutions of Korean L1 Essays

Surim Kim, Sookki Choi*

Korea National University of Education

*Corresponding Author: agrement@knue.ac.kr

ABSTRACT

The study utilized 401 data points categorized into upper, middle, and lower levels. The model development process included five stages: 1) data purification and preprocessing, 2) embedding, 3) data segmentation and shape conversion, 4) model training, and 5) model performance evaluation. The research employed BERT, a pre-trained model, to develop the grading model. Performance evaluation of the trained model yielded an accuracy of 0.7377, precision of 0.6514, recall of 0.7377, and an F1 score of 0.6717. These results demonstrate a relatively high level of performance compared to previous studies on scoring Korean written and essay answers by L1 learners. As a result, the feasibility of developing an automatic scoring model using the BERT language model with small-scale data from a specific domain. Also, the model's performance shows some generalizability, but further exploration is needed for improvement. Limitations of the study include the use of writing samples graded without considering grade levels, limited test data, difficulties in processing unregistered tokens, and the inherent unexplainability of deep learning techniques. Further discussions and considerations on how to more effectively utilize artificial intelligence in Korean language education should continue. Ongoing research on automatic grading is necessary to provide accurate and detailed educational feedback to students. As automatic grading research in the field of Korean language education advances, it is expected that high-quality educational interventions for students will become possible in the future.

Keywords: Automated essay scoring, Korean essay scoring, L1 learner writing, deep learning, artificial intelligence automatic grading, AI automatic grading

Brain, Digital, & Learning
2024, Vol. 14, No. 4, 633-652
https://doi.org/10.31216/BDL.20240036

Received: December 11, 2024
Revised: December 29, 2024
Accepted: January 02, 2025

Introduction

As core competency assessment has emerged in the 2015 revised curriculum, there has been an emphasis on utilizing essay and written response assessments to measure students' critical thinking and integrated thinking skills, and to strengthen high-level analytical and problem-solving abilities (Park et al., 2022).

While discussions continue about introducing essay and written response assessments in future College Scholastic Ability Tests (CSAT), the "Elementary and Secondary School Curriculum

General Guidelines (Ministry of Education Notice No. 2022-33)" issued under the 2022 revised curriculum has established a foundation for expanding the proportion of essay and written assessments in schools. In particular, the "Korean Language Curriculum (Ministry of Education Notice No. 2022-33)" focused on evaluating real-life application competency, comprehension and inquiry abilities, and comprehensive thinking skills through essay and written assessments.

In fact, despite the educational value and advantages of essay and written assessments, particularly essay-type evaluations, they are still not actively utilized in educational settings due to the burden of grading and resulting problems. Regarding this, Song et al. (2016) pointed out that practical constraints such as time and cost, subjectivity in grading, consistency among graders, and grader fatigue may arise. Additionally, the lack of feedback that should be provided to students becomes problematic due to teachers' grading burden (Park et al., 2019).

AI-based automated scoring is gaining attention as a solution to alleviate assessment fairness issues and reduce the grading burden on human raters, namely teachers, while overcoming these practical constraints in grading situations (Kim, 2022; Lee et al., 2022). Generally, artificial intelligence (AI) assessment has advantages in terms of assessment speed, flexibility, and fairness (Pearson Languages, 2023). According to Lee (2022), AI assessment can be considered more fair compared to human rater assessment as it enables use without spatial or temporal constraints, rapid diagnosis of learner writing, and provision of consistent and standardized scoring.

The advantages of introducing essay and written response assessments include enabling the evaluation of critical, logical, and creative thinking and competencies, and expanding writing education beyond the existing reading-centered education in Korean language education. However, issues related to assessment fairness, shortage of grading personnel, and assessment burden remain significant concerns regarding the introduction of essay and written assessments.

Given the characteristics of the Korean language subject, which is directly connected to thinking and expression skills that form the foundation of essay and written assessments (Choi, 2023), more active research needs to be conducted on various automated scoring methods and scoring stabilization measures, including assessment fairness. Despite this necessity, automated scoring becomes more challenging compared to English-speaking regions due to the agglutinative characteristics of Korean, resulting in a relative lack of research on Korean essay automated scoring in the field of Korean language education.

While research on automated scoring methods for L1 learners' essays in Korea progresses slowly compared to overseas, this study was primarily conducted based on BERT, which is widely known as the most fundamental and powerful model. As can be seen in recent domestic research on essay automated scoring methods by Lee & Park (2022), Kim et al. (2023.11.), Kim et al. (2023), Yu et al. (2023), and Han et al. (2023), BERT is a powerful model being significantly used alongside the random forest algorithm, an ensemble technique.

In this study, we constructed an automated scoring model for Korean essays on a single topic written by L1 learners and verified the model's performance and potential for field application. Additionally, we proposed using the deep learning technique BERT, which is pre-trained on a large scale, among AI automated scoring algorithms as a solution to alleviate grading burden in school settings while resolving problems that occur in assessment situations. The BERT (Bidirectional Encoder Representations from Transformers) model is known to show high performance in various

natural language processing tasks as it can learn more accurate meanings and contextual characteristics through context-based bidirectional learning (Ravichandiran, 2021). Through this, we aimed to contribute to improving fairness and efficiency in writing education within Korean language education, and furthermore, to creating an educational environment where learner feedback can be smoothly provided in the future.

Background

Conceptual Framework and Development of Automated Essay Scoring

Automated Essay Scoring (AES) employs technology to evaluate and score written essays based on pre-defined criteria, such as grammar, lexical richness, coherence, syntax, and semantic relevance. AES addresses key challenges in human scoring, including subjectivity, inconsistency, and time inefficiency (Attali, 2013; Dikli, 2006). In recent years, AES has undergone significant advancements in artificial intelligence (AI), deep learning, and natural language processing (NLP), ensuring its relevance to modern educational contexts.

Educational assessments are broadly categorized into selected-response (SR) and constructed-response (CR) items. SR items, such as multiple-choice questions, are easier to score objectively. CR items, including essays, require subjective and complex evaluation to assess higher-order thinking skills, such as analysis and synthesis (Isaacs et al., 2013; Nitko & Brookhart, 2007). AES was developed to address this complexity, offering a consistent, objective, and scalable solution for scoring CR items (Ramesh & Sanampudi, 2022).

Human scoring often faces challenges such as fatigue, subjectivity, and variability among raters. By contrast, AES systems use algorithmic methods to deliver objective, reproducible results while significantly reducing the time and labor required for evaluation (Powers et al., 2000). These characteristics make AES particularly appealing for large-scale assessments in educational institutions.

The origins of AES trace back to the 1960s with Ellis Page's Project Essay Grade (PEG). PEG used statistical methods to assess essay quality based on textual features, such as sentence length and punctuation (Dikli, 2006). While groundbreaking for its time, PEG faced criticism for focusing on superficial linguistic features, neglecting deeper semantic elements essential for meaningful evaluation (Attali, 2013). For example, test-takers could manipulate scores by inflating word counts or inserting unnecessary punctuation (Page, 2003). Nonetheless, PEG introduced foundational concepts, such as using "proxies" to measure linguistic fluency and diction, establishing a basis for subsequent AES developments (Dikli, 2006).

The 1990s marked significant progress in AES, fueled by advancements in NLP and machine learning. Systems like ETS's e-rater and the Intelligent Essay Assessor (IEA) emerged during this period, incorporating deeper linguistic analyses, such as syntax, semantics, and discourse structure evaluation (Burstein et al., 2013; Landauer et al., 1998).

E-rater, developed by Educational Testing Service (ETS), utilizes statistical and NLP techniques to evaluate grammar, organization, and content development. By incorporating semantic analysis, e-rater overcame many limitations of earlier systems like PEG (Burstein et al., 2013). Similarly, IEA employed Latent Semantic Analysis (LSA) to compare student

essays against domain-relevant text corpora, effectively evaluating coherence and relevance (Landauer et al., 1998).

Recent years have witnessed the integration of deep learning into AES, enabling the automatic extraction and learning of complex features from text data. Alikaniotis et al. (2016) introduced a bi-directional Long Short-Term Memory (LSTM) model with score-specific word embeddings (SSWE). This approach significantly enhanced essay scoring accuracy by capturing nuanced linguistic patterns and adapting to diverse textual inputs. Unlike earlier systems reliant on handcrafted features, the LSTM model demonstrated superior scalability and flexibility, making it a robust tool for essay evaluation.

Similarly, Taghipour and Ng (2016) developed a recurrent neural network (RNN)-based model that integrated convolutional layers to capture local dependencies and recurrent layers for long-term dependencies. This innovative design allowed the system to address persistent challenges such as assessing essay coherence and higher-order reasoning. The integration of convolutional and recurrent layers provided a multi-dimensional view of textual data, improving both accuracy and interpretability in automated essay scoring tasks.

Transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have redefined the AES landscape. These models leverage self-attention mechanisms to dynamically weigh the importance of words and phrases in context, enabling a holistic understanding of linguistic and contextual nuances. BERT, for example, employs bidirectional training to consider both preceding and succeeding word contexts, allowing for more precise evaluations of coherence and argumentation (Devlin et al., 2019). GPT, with its generative capabilities, excels in understanding creative and narrative constructs, making it particularly suited for evaluating essays with open-ended prompts.

Transformer-based models have achieved state-of-the-art performance in AES, excelling in tasks that require the evaluation of complex constructs such as creativity, argumentation, and rhetorical structure. For instance, these models have demonstrated their ability to align closely with human scoring by synthesizing semantic, syntactic, and discourse-level information (Zhang & Litman, 2018). Furthermore, advancements in pre-training techniques and fine-tuning have enabled these systems to adapt to specific domains, enhancing their generalizability across diverse essay topics and formats.

As deep learning models continue to evolve, they offer promising pathways for addressing the limitations of traditional AES systems. By combining the contextual depth of transformer-based architectures with the precision of earlier statistical methods, modern AES systems are poised to deliver more accurate, reliable, and equitable assessments.

From 2020 onward, AES research has emphasized explainability, fairness, and adaptability. The emergence of explainable AI (XAI) techniques has enabled educators to understand how AES systems derive scores, improving trust and transparency in automated assessments (Kumar & Boulanger, 2020). Additionally, hybrid models combining handcrafted and neural network-based features have demonstrated success in balancing interpretability and performance (Faseeh et al., 2024).

Innovations such as multilingual AES systems have also gained attention. Newer models now support essay evaluation in multiple languages, expanding the applicability of AES in diverse linguistic contexts (Horbach et al., 2024). Furthermore, adaptive learning platforms have integrated AES to provide personalized feedback, enabling students to address specific writing deficiencies and improve over iterative drafts (Nguyen & Dery, 2016).

Beyond scoring, AES has been widely adopted in educational settings through Automated Writing Evaluation (AWE) tools. Examples include Criterion, which integrates e-rater, and MY Access!, powered by IntelliMetric. These tools deliver real-time feedback on grammar, style, and organization, helping students refine their writing iteratively (Attali & Burstein, 2006; Dikli, 2006).

Emerging applications have focused on integrating AES into adaptive learning systems. These platforms use AES to recommend personalized exercises and resources based on individual student performance, fostering a holistic approach to writing improvement (Hussein et al., 2020).

Despite its advancements, AES faces challenges, including biases in training datasets, limitations in assessing creativity, and the "black-box" nature of deep learning models. Addressing these issues requires robust data collection practices and algorithmic safeguards to ensure fairness across diverse demographic groups (Hussein et al., 2019).

The future of AES lies in hybrid systems that combine the precision of handcrafted features with the adaptability of deep learning. Transformer-based architectures will likely play a pivotal role in advancing AES capabilities, particularly in evaluating critical thinking and argumentation (Zhang & Litman, 2018). Ethical considerations, such as ensuring inclusivity and minimizing bias, must remain central to AES research to promote equity and trust in automated scoring systems.

Automated Scoring for Korean L1 Essays

While automated scoring technology previously used traditional machine learning techniques, several deep learning-based writing automated scoring studies are being conducted as algorithms have significantly advanced recently. Deep learning is an artificial intelligence technique where machines learn patterns extracted from input data independently, with 'Deep' referring to a structure having hidden layers. Deep learning techniques are widely used in various fields for image classification, speech recognition and synthesis, and natural language processing (Lee, 2021). Major deep learning models include RNN (Recurrent Neural Network), CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory Network), GRU (Gated Recurrent Unit), BERT (Bidirectional Encoder Representations from Transformers), and GPT (Generative Pre-trained Transformer).

The concept of artificial neural networks was proposed as a model abstracting the information processing function of biological neurons (McCulloch & Pitts, 1943), where artificial neural networks operate similarly to biological neurons. Artificial neural networks are based on perceptron structure and form the foundation of deep learning. This deep learning technology is highly effective for automated scoring (Park et al., 2022), and particularly the BERT model announced by Google in 2018 can build models with relatively superior performance even with limited data through adjustment processes suitable for researchers' tasks and transfer learning (Sun et al., 2019; Dodge et al, 2020; Park et al., 2022).

Unlike Word2Vec, BERT is a context-based model that understands sentence context and generates word embeddings to match subsequent contexts (Ravichandiran, 2021), making it useful for relatively accurate understanding of word meanings and contextual characteristics. This BERT has proven superior embedding quality and has been proposed as a solution to address the lack of scoring data based on pre-training and fine-tuning (Devlin et al., 2019).

Additionally, being transformer-based, BERT is a model capable of reducing learning time compared to RNN. The transformer, which overcomes the information loss problem of existing seq2seq models, performs word embedding

tasks using attention models, where the decoder can make superior predictions by instantly extracting necessary information using vector information for all words in the input text. BERT uses only the encoder part of the transformer, predicting masked words through bidirectional learning enabled by randomly masking some words as part of the masked language model (Devlin et al., 2019).

Research on automated scoring models and systems using BERT is being conducted in the context of expanding utilization in educational settings for expanding writing assessment recently. For example, Kim et al. (2023.11.) explored optimal BERT-based models using 12,600 Korean essays written by elementary, middle, and high school students, provided by AIHUB. They found that Longformer showed the best performance (Accuracy: 0.770, Precision: 0.766, Recall: 0.770, F1-score: 0.768, QWK: 0.830) with BERT showing the second-best performance (Accuracy: 0.769, Precision: 0.766, Recall: 0.769, F1-score: 0.767, QWK: 0.826). Kim et al. (2023) proposed BERT and GPT-based system construction methods through scoring models predicting essay creativity scores for English and Korean writing using 3,766 essays. Korean writing assessment performance showed accuracy, precision, recall, and f1 scores of 0.629, 0.465, 0.445, 0.455 for BERT-based and 0.584, 0.336, 0.333, 0.334 for GPT-2-based models, indicating performance limitations due to data class imbalance despite these models being known for generally superior performance.

Additionally, research cases of automated scoring methods using deep learning techniques include studies developing and proposing human-machine collaborative scoring models for argumentative writing tasks using deep learning algorithms like CNN and BERT (Kim, 2022), and research developing PASTA-Ⅰ, an essay and written response automated scoring program using ELECTRA, another pre-trained language model, utilizing essay test data from the National Information Society Agency (NIA) (Lee et al., 2023).

Park & Lee (2022), who explored the feasibility of automated scoring for Korean essays, compared deep learning-based algorithms, including Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU), to identify the optimal algorithm for automated scoring. The study utilized essay data from 530 university students, evaluating four scoring areas (content, expression, structure, and grammar) and assessing each model's performance based on accuracy, precision, recall, and F1 scores. The results indicated that LSTM and GRU outperformed RNN, with GRU demonstrating a shorter training time, making it a suitable algorithm for large-scale data learning.

Through these research cases, we can see they focus on developing teaching and learning support tools or feedback provision programs while comparing algorithm performance, and automated scoring model research is exploring performance and possibilities with interest in whether they can be used as auxiliary tools to reduce teachers' scoring burden. This study aimed to present itself as a case that could expand into research directly beneficial to teachers and students by confirming how well the BERT algorithm performs for Korean L1 learners' essays and examining its potential for field application in terms of resolving assessment fairness and personnel issues in essay evaluation situations.

Meanwhile, research has been conducted on automated scoring methods using deep learning for scoring Korean writing answers written by learners whose native language is not Korean (Lee et al, 2022; Cho et al., 2021). Lee et al. (2022) developed a BERT-based model to automatically grade descriptive answers by Korean L2 learners at the Intermediate 1 level (TOPIK Level-3). Using 586 responses written between 2013 and 2017 on the topics "My

Hobby" and "Memorable Travel," the study demonstrated the model's strong performance, achieving 95.8% accuracy at epoch 40, and revealed a significant correlation between response length and scores. Additionally, Cho et al. (2021) confirmed the feasibility of automated scoring in Korean writing assessment by constructing an automatic score interval classification model using KoBERT and KoGPT2, achieving classification performance with average accuracies ranging from approximately 44% to 65%, even with small-scale data.

Korean essay automated scoring research types can be broadly categorized as 1) detailed research on scoring features, 2) focusing on potential use as auxiliary tools in educational settings, or 3) research on scoring model construction and validation. While research is relatively active in Korean L2 learner-targeted studies and computer engineering fields, there are only a few BERT utilization research cases available in automated scoring for evaluating essays written in Korean as L1, despite sufficient necessity.

Therefore, considering that it is mainly used along with random forest algorithms when utilizing small amounts of data, this study aimed to expand the possibilities of research in Korean language automated scoring by constructing an automated scoring model based on BERT, known to be useful even with small-scale data, and verifying its performance.

Method

Essay Database

The data used throughout this study, including model training and validation, is from data collected by Choi (2018). We aim to verify the level of performance as an automated scoring model targeting 401 writings. Considering potential negative impacts on the data learning process, 13 writings of excessively short length were excluded. The complete dataset consists of writings and scoring grades from 87 first-year middle school students, 101 second-year middle school students, 95 third-year middle school students, and 118 first-year high school students.

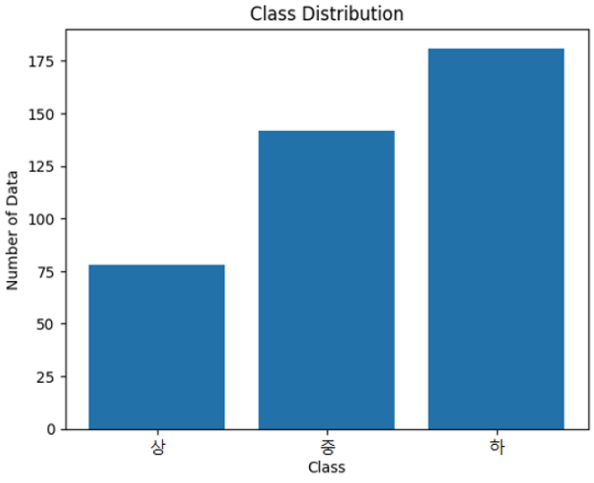

Fig. 1. Distribution of Number of Data by Class

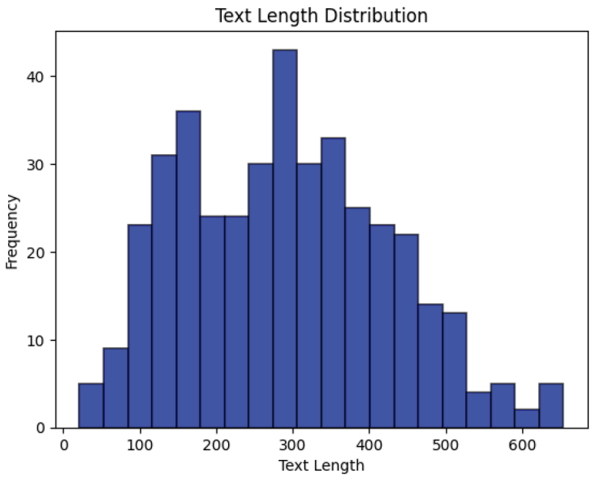

Fig. 2. Distribution of text data length

This study aimed to build a model that classifies levels into three categories: high, medium, and low. The data includes 78 writings at the high level, 142 at the medium level, and 181 at the low level. These represent approximately 19.45%, 35.41%, and 45.14% of the total data respectively, with high-level data being the most scarce. The distribution pattern of text data length varies greatly, ranging from about 25 to 650.

The score data collected alongside the essays used argumentative writing evaluation criteria developed by Park et al. (2019), scored using detailed evaluation standards across six areas: ‘analysis and utilization of argumentative materials’, ‘validity of argumentative structure’, ‘unity of writing’, ‘cohesion of writing’, ‘fluency and diversity of expression’, and ‘accuracy of grammar and citation’. The detailed evaluation criteria consisted of 22 items, with three Korean language teachers scoring each criterion as either 1 point or 0 points based on whether it met the requirements, structured to have a maximum score of 3 or 4 points per evaluation element. Writing evaluation workshops were conducted for student writings, and inter-rater reliability was checked by implementing common scoring among evaluators for over 35% of the writings in each group, confirming a high Krippendorff α value of 0.93 (Kim et al., 2020:191).

The Procedure of BERT Application

In the data cleaning and preprocessing stage, first, all elements (characters) except Korean consonants, vowels, Korean characters, and spaces were removed from input sentences and replaced with spaces to clean the sentences. Then, after retrieving the input values corresponding to the original data’s text and text-specific scores, each text’s sentences were tokenized using the pre-trained BertTokenizer. The length measurement target becomes the number of subword-unit tokens tokenized through BertTokenizer. Generally, BERT uses WordPiece Tokenizer, which is based on a Subword Tokenization algorithm that breaks given text into subword units.

In the subsequent embedding stage, variables are constructed as values corresponding to input_ids, attention_mask, and token_type_ids are added for each text. First, input_ids represent the token sequence converted to indices, while attention_mask indicates whether it is a padding token (Lee, 2021). If it corresponds to a padding token, the attention_mask value appears as 0; if not, it appears as 1. Finally, token_type_ids contain BERT’s segment information, which is a token sequence that varies according to document order - for example, the token sequence of the first document becomes 0, and that of the second document becomes 1 (Lee, 2021).

The BERT model has two pre-training methods: MLM (Masked Language Modeling) and NSP (Next Sentence Prediction). Among these, when performing the latter NSP task, segment information is input during the process of classifying whether two input documents have continuity or not (Lee, 2021). The BERT model, in performing the task of grade classification, that is, document (sentence) category classification, returns probability values for each category that a document belongs to, where one student’s writing is treated as one document and categories are classified based on the entire document. Accordingly, it becomes a BERT model that takes two documents as input and determines whether the tokens belonging to the document are from the first document or the second document.

Finally, in the data splitting and format conversion stage, we performed the task of dividing data into training data and test data, followed by converting the training data into a format suitable for input into the BERT model. Table 1 below summarizes the values corresponding to the input_ids, attention_mask, and token_type_ids variables of writing No. 2 after going through the embedding process, along with the values output when decoded. In the decoding output values,

which are conversions to text, [UNK] represents unregistered tokens, [CLS] appearing at the beginning of each sentence serves as a token that distinguishes it from the previous sentence, [SEP] appears at the end of the last sentence of the writing indicating its conclusion, and the subsequent [PAD] tokens are padding tokens where remaining embeddings are all processed as 0 to match the predetermined number of embeddings.

Table 1. input_ids, attention_mask, token_type_ids, and Sequence Decoding Values for Article 2(continued)

input_ids of Article 2	[101 9448 10954 33768 11489 8922 16605 12453 13767 9758 33768 22333 10954 10459 63185 10622 9248 10247 24982 9309 19105 80001 8992 119274 12965 21711 16139 119 9599 24891 32158 9248 10247 24982 9309 19105 9758 33768 22333 10954 10459 9340 119216 11018 9960 101814 122 19105 29935 8865 10739 9323 24017 14102 119 113 9768 60469 131 8935 18623 49515 113 10158 114 119 9448 10954 33768 8857 16605 10459 9949 48549 17138 11882 8857 16605 9328 79544 117 9751 22333 10954 30005 20309 25486 17196 9672 90861 20309 117 10413 22028 114 32775 9408 12030 117 9929 25549 117 9328 18227 117 9666 12092 9121 30085 110149 9663 12508 37819 9248 10247 24982 9309 19105 9758 33768 22333 27056 91621 12178 9405 61250 22879 9434 12030 10530 100876 9414 14423 17022 8843 118983 21614 9744 68773 10622 9322 24098 9983 12945 9744 97005 9322 12508 45593 117 9356 20309 9744 37712 19105 9322 14153 22143 119 8924 56710 28911 9599 119229 9751 22333 10954 22879 9670 25387 14801 117 9626 29683 14801 9434 55635 8898 41521 80001 9074 9389 17342 119210 24098 9546 118881 10622 9137 17655 8843 16605 11882 9953 25242 20625 9625 18622 108280 12424 89093 9251 16985 21711 23969 9966 18778 117 9251 10622 8972 36240 119136 10739 84703 11664 11506 119 9638 118866 35979 103279 100 84703 12508 9290 70146 100 117 100 80956 9966 69448 9952 14867 9521 54780 8870 12030 12508 9287 17342 12424 100 31613 23969 30050 8924 48387 9951 21611 18392 9400 11925 119 9144 9670 25387 17022 9309 119063 13374 9405 11664 61964 16439 9903 24989 61964 9365 52560 13374 9340 119216 11513 9663 48599 94710 9952 54141 17342 12092 9340 119216 11513 9663 12508 37819 30050 9663 12508 37819 8870 117 41099 110463 9835 9339 10622 9689 14867 100 8924 118627 10892 9952 14867 9521 24683 9966 18778 10739 17196 16439 100 9233 8924 8932 14863 10530 9994 31503 18108 8946 118772 11664 52292 17655 11018 89093 9523 10622 24190 119 23289 100 9405 14863 11261 25805 9091 16985 119153 8932 88168 9672 28000 108436 16139 119 100 33292 9251 10739 60030 9655 118673 9339 10622 20625 9519 96279 9689 12508 12092 45593 25805 11018 8924 56710 9712 10622 9521 9955 27487 60362 9356 63671 39218 119 9604 26344 9340 119216 11287 9641 54305 9109 10530 8924 93834 9460 119081 12178 27487 63783 8924 9641 10739 25805 9641 12965 16439 12508 9523 14153 9247 16985 21711 16139 119 102 0]

Table 1. input_ids, attention_mask, token_type_ids, and Sequence Decoding Values for Article 2

attention_mask of Article 2	[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
token_type_ids of Article 2	[0 0]
Sequence-to-Text Return for Article 2	[CLS] 소년법에서 규정하고 있는 촉법소년의 기준을 만 14세 미만보다 낮추어야 한다. 요즈음 만 14세 미만 촉법소년의 범죄는 해마다 1만여 건이 발생한다. (출처 : 김진희 (2011). 소년법 개정의 필요성과 개정 방향, 청소년보호연구 제27호, 31쪽) 하지만 살인, 폭행, 방화, 절도 등 여러 일을 저지른 만 14세 미만 촉법소년에 해당하는 사람들은 성인에 비해 상대적으로 가벼운 처벌을 받으며 형사 처분을 받지 않고, 보호 처분만 받게 된다. 그런데 요즘 청소년들은 정신적, 육체적 성장이 과거보다 더 빨라졌으며 어렸을 때부터 가정과 학교 또는 유치원에서 하지 말아야 하는 행동, 말을 끊임없이 배우고 있다. 이럼에도 불구하고 [UNK](미등록 토큰) 배우지 못해서 [UNK], [UNK] 이런 행동을 하면 안 되는 것인지 몰라서 [UNK] 라고 하는 것은 그저 핑계일 뿐이다. 또 정신적으로 미숙하여 사고력이나 판단력이 부족하여 범죄를 저질렀다 하더라도 범죄를 저지른 것은 저지른 것, 이에 따른 큰 벌을 주면 [UNK] 그것은 하면 안되는 행동이구나 [UNK] 를 그 기회에 확실히 깨닫고 다음부터는 하지 않을 것이다. 그리고 [UNK] 사회로 다시 돌아올 기회를 제공해야 한다. [UNK] 라는 말이 있는데 잠깐 벌을 또는 아예 주지도 않고 다시는 그런 짓을 안 할 것이라는 보장은 없다. 우리는 범죄가 일어난 뒤에 그것을 수습하는 것이 아닌 그 일이 다시 일어나지 않게 막아야 한다. [SEP] [PAD]

Throughout the entire process of data training and performance evaluation in this study, a Tesla T4 GPU was used in a development environment with 12.7GB of system RAM. Ubuntu 22.04.3 LTS was used as the operating system (OS), with Transformer 4.35.2 library version, and Python 3.10.12 was used as the development language. Finally, TensorFlow 2.15.0 and its corresponding version of Keras library (2.15.0) were used as the framework for the development language and libraries.

For fine-tuning work, hyperparameters were set to epochs=5, Adam optimizer, batch_size=2, max_length=512, and learning rate=1e-5. The fine-tuning performed in this study can be described as the process of adding layers or adjusting weights to enable the task of classifying document grades according to the three categories - advanced level, intermediate level, and basic level - that were previously established for the student essay data being scored.

Additionally, to address the issue of varying output sum totals occurring in grade classification problems, we underwent a normalization process using the softmax function to fairly compare these totals (Park, 2019). Through this softmax function applied to the classifier in the output layer, the probability of each data corresponding to advanced level, intermediate level, and basic level is returned, and the level (class) corresponding to the highest probability is output as the predicted value.

Results and Discussion

Result of Application BERT

We constructed a BERT model trained on the previously cleaned data and verified its performance. The validation data, like the training data, belongs to one of the three categories - advanced, intermediate, or basic - and the overall pattern of agreement between actual values and predicted values is shown in Fig 3. In Fig 3, darker blue indicates greater agreement between the actual values of the data and the model's predicted values. Through this confusion matrix graph, we can intuitively grasp the agreement between the predicted values and actual values when the scoring model performs scoring on test data.

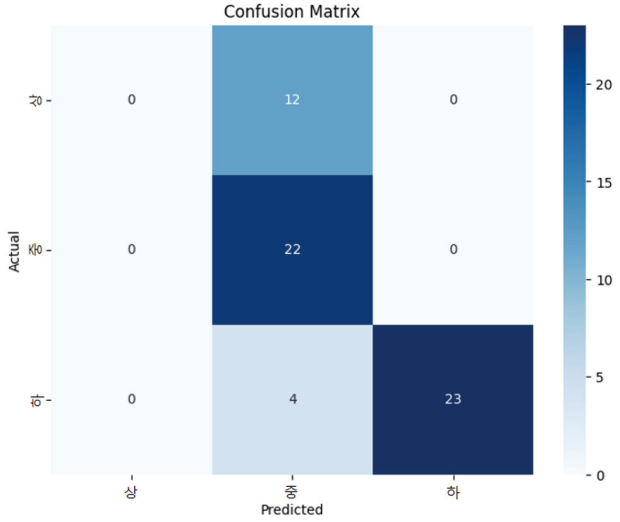

Fig. 3. Confusion matrix for model scoring results for test data

First, for intermediate and basic level test data, there was relatively high agreement between actual values and predicted values. In contrast, for advanced level test data, except for one data point, we observed that predictions tended to be at the intermediate level, one level below advanced. Meanwhile, for intermediate level test data, all 22 data points showed agreement between actual and predicted values. Finally, for basic level test data, 23 out of 27 data points had matching predicted-actual values, achieving an accuracy rate of about 85%. The normalized confusion matrix showing the pattern of actual-predicted values by classification class for the scoring model is presented in Table 3, based on the basic form of confusion matrix shown in Table 2.

Table 2. Basic form of confusion matrix

		Predicted	
		Positive	**Negative**
Actual	**Positive**	**True Positives**	False Negatives
	Negative	False Positives	**True Negatives**

Table 3. Normalized confusion matrix of model grading results for test data

Actual \ Predicted	Advanced	Intermediate	Basic	total
Advanced	0.00	1.00	0.00	1
Intermediate	0.00	1.00	0.00	1
Basic	0.00	0.15	0.85	1

The above normalized confusion matrix table calculates probabilities by considering the total number of essays at each level as the whole for each level in this study, so the sum of probabilities for each level equals 1. Through this table, we can intuitively confirm through normalized figures not only which category the model's predicted data actually belongs to but also what the proportions are. Taking the 'intermediate level' row under 'Actual' as an example, we can see the figures for: first, when writing predicted as advanced level is actually intermediate level; second, when writing predicted as intermediate level is actually intermediate level; and third, when writing predicted as basic level is actually intermediate level.

Due to using small-scale data, this study set 85% as training data and 15% as test data, with 20% of the training data used as validation data. This study aimed to verify the performance of the trained model by checking how well it predicts when using test data not previously used in training, rather than the training data. The performance evaluation results for the 61 test data points across various performance metrics can be seen in the table below.

Table 4. BERT model performance (overall)

Test data	61		
Accuracy	Precision	Recall	f1-score
0.7377	0.6514	0.7377	0.6717

The scoring model in this study shows an Accuracy of 0.7377, Precision of 0.6514, Recall of 0.7377, and F1 Score of 0.6717. Here, accuracy is the value obtained by dividing the number of data points where predicted values match actual values by the total number of data points; precision is the proportion of data actually belonging to a category among data predicted for each category based on predicted value categories; recall is the proportion predicted for each category among total data in each category when based on actual value categories; and f1-score is the harmonic mean of precision and recall.

Meanwhile, performance metrics by classification category can be seen in the table below. At the advanced level, accuracy, precision, recall, and F1 score all showed 0.00, while at the intermediate level, they measured 1.00, 0.58, 1.00, and 0.73 respectively, and at the basic level, they measured 0.85, 1.00, 0.85, and 0.92 respectively.

For the advanced level category, the precision of 0.00 means that 0% of the total data predicted by the model as advanced level actually corresponded to advanced level. This figure appeared similarly in recall and f1 score, and due to the scarcity of essays corresponding to the advanced level category, it can be seen as showing significantly degraded performance compared to the other two categories, intermediate and basic levels. The precision for intermediate and basic levels is 0.58 and 1.00 respectively, meaning that 58% and 100% of the total data matched their actual categories based on the categories predicted by the model for each level. For recall, with values of 1.00 and 0.85 for intermediate and basic levels respectively, this can be interpreted as 100% and 85% of actual intermediate and basic level samples being correctly predicted.

Table 5. BERT model performance(by class)

Class	Precision	Recall	f1-score	Test data
Advanced	0.00	0.00	0.00	12
Intermediate	0.58	1.00	0.73	22
Basic	1.00	0.85	0.92	27
Macro avg	0.53	0.62	0.55	-
Weighted avg	0.65	0.74	0.67	-

Macro avg and weighted avg are comprehensive indicators of performance for each class. The macro avg (macro average) is calculated by giving equal weight to all classes, producing values by calculating precision, recall, and f1-score for each class and then finding their average. In this study, the macro avg shows precision of 0.53, recall of 0.62, and f1-score of 0.55.

On the other hand, the weighted avg (weighted average) of precision, recall, and F1-score is calculated by reflecting weights assigned according to each class's size. In this case, classification classes with relatively more data have greater weight, so ultimately the weighted avg values are more influenced by larger classes (Park & Choi, 2023). In this study, the weighted avg values show precision, recall, and f1-score of 0.65, 0.74, and 0.67 respectively, and unlike the macro avg values examined earlier, this can be interpreted as showing somewhat higher performance due to being most influenced by the basic level classification class, which has a larger size.

In conclusion, considering the overall data of this study's scoring model, despite the considerably high frequency of

intermediate and basic level categories, the overall model performance appeared low (about 57%) due to the extremely low performance in the advanced level category.

Comparing this study's scoring model performance with performance values presented in several studies shows the following*: Lee & Park (2022) showed somewhat poor performance where predicted scores concentrated on certain scores among three classification interval scores in an imbalanced data situation. While they reported their KoBERT-based scoring model had accuracy of 35%~41%, precision of 12%~17%, recall of 35%~41%, and f1 score of 18%~24%, this study's scoring model shows significantly higher levels with accuracy of 0.7377, precision of 0.6514, recall of 0.7377, and f1 score of 0.6717. Meanwhile, Cho et al. (2021), who also used small-scale data, confirmed that the classification performance using the average accuracy of classification models based on KoBERT and KoGPT2 was about 44% to 65%. Based on these figures, we can see that this study's scoring model shows relatively high performance with an accuracy value of 0.7377.

However, this is a lower performance level compared to the highest accuracy of 95.80 in Lee et al. (2022), who constructed a BERT-based Wscore model, indicating potential for expansion into research to improve model performance in the future.

Discussion

The performance of the BERT-based automated essay scoring (AES) model demonstrated promising results, particularly for intermediate and basic-level essays, achieving accuracy rates of 100% and 85%, respectively. These results align with prior studies highlighting BERT's capability to process structured language data effectively (Devlin et al., 2019; Horbach et al., 2024). However, the model's failure to classify advanced-level essays accurately, as indicated by an F1 score of 0.00, underscores the limitations posed by data imbalance and insufficient training examples. Similar challenges have been observed in studies on AES systems, where imbalanced datasets often lead to reduced performance for less-represented categories (Horbach et al., 2024; Hussein et al., 2019).

The significant impact of data imbalance on the model's performance highlights the necessity for balanced datasets in AES development. Strategies such as data augmentation, synthetic data generation, and transfer learning have been proposed as effective solutions to mitigate this issue (Horbach et al., 2024). Furthermore, Zhang and Litman (2018) emphasized the importance of leveraging transformer-based models to enhance the evaluation of complex constructs like coherence and argumentation. These strategies could be instrumental in improving the model's performance, particularly for advanced-level texts that require nuanced linguistic understanding.

Comparing the performance of the current model with international benchmarks reveals both strengths and areas for improvement. While the model's overall accuracy of 73.77% surpasses several earlier AES implementations, such as KoBERT-based systems (Lee & Park, 2022) and IEA (Landauer et al., 2003), it falls short of the superior performance demonstrated by e-rater (Burstein et al., 2013) and Wscore (Lee et al., 2022). The integration of advanced architectures like Longformer (Beltagy et al., 2020) or ELECTRA (Clark et al., 2020) could address these limitations by enhancing

* Regarding all performance comparisons below, while exact comparisons are not possible as these are separate studies, comparisons were made using the most compatible metrics available.

the model's ability to process longer and more complex texts, thereby improving its applicability across diverse educational contexts.

Transformer-based architectures have redefined the capabilities of AES systems, with models like BERT and GPT offering significant advancements in processing contextual and semantic information. While BERT excels at bidirectional contextual learning, enabling precise evaluations of coherence and argumentation (Devlin et al., 2019), GPT's generative capabilities make it well-suited for assessing creative and narrative constructs (Alikaniotis et al., 2016; Taghipour & Ng, 2016). The integration of these approaches, combined with domain-specific fine-tuning, as suggested by Horbach et al. (2024), could further enhance the adaptability and effectiveness of AES models in Korean educational settings.

Finally, the educational implications of this research are significant. The high accuracy observed for intermediate and basic levels suggests that the BERT-based AES model could effectively support formative assessments, alleviating the grading burden on teachers. However, the model's limitations at the advanced level necessitate the adoption of Explainable AI (XAI) techniques to improve transparency and trustworthiness in the scoring process (Kumar & Boulanger, 2020). Additionally, integrating real-time feedback systems, as proposed by Han et al. (2024), could provide students with actionable insights to refine their writing skills. Such developments would not only enhance the reliability of AES systems but also promote their adoption in educational environments by addressing fairness, efficiency, and usability.

Conclusions

The growing focus on fairness and efficiency in educational assessments has heightened the demand for Automated Essay Scoring (AES) technologies. Specifically, AES systems for Korean essays hold the potential to provide personalized feedback to learners while alleviating the grading burden on teachers. Such technologies also enhance assessment consistency and reliability, thereby improving fairness among learners. As essay-based constructed-response assessments play an increasingly significant role in Korean education, the need for efficient and reliable AES systems becomes more pronounced. This study employed the BERT model to develop an AES system for Korean L1 learners' essays, demonstrating its efficacy even with small-scale data and exploring its educational applications.

The key findings of this study are as follows:

First, the study successfully developed an AES model using the BERT framework, tailored for Korean L1 learners. Despite the constraints of limited data, the model exhibited high performance, particularly in intermediate and beginner-level essays. This confirms the robustness of BERT in delivering reliable results even with small datasets.

Second, performance evaluation results showed an overall accuracy of 73.77%, precision of 65.14%, recall of 73.77%, and an F1 score of 67.17%. These results underscore the model's effectiveness compared to similar AES models from previous studies, particularly in handling data at intermediate and beginner levels.

Third, the model's performance deteriorated for advanced-level data due to a lack of sufficient training examples in this category. This resulted in an F1 score of 0.00 for advanced essays, highlighting the critical impact of data imbalance on the learning and evaluation processes.

Fourth, the study partially confirmed the model's generalizability but emphasized the necessity of incorporating larger datasets, including advanced-level essays. By leveraging diverse topics and writing styles, the model's generalizability and educational applicability could be significantly expanded.

The successful application of the BERT-based AES model to Korean essay scoring is a promising advancement. The high accuracy and recall rates achieved in intermediate and beginner-level essays demonstrate the model's capability to effectively evaluate essays that adhere to basic writing rules and logical structures. These findings align with previous studies where BERT demonstrated strong performance with languages possessing relatively simpler grammatical structures, such as English (Devlin et al., 2019).

However, the model's poor performance on advanced-level essays underscores the dependency of AES systems on the quality and diversity of training data. Advanced-level writing often involves more creative and complex reasoning, which the model failed to adequately learn and evaluate due to limited exposure. This result emphasizes the importance of balanced datasets that adequately represent different learner levels when developing AES systems for practical educational use.

Additionally, while the study confirmed the feasibility of BERT in small-scale datasets, it lacked comparative analysis with other state-of-the-art deep learning models. Advanced architectures like RoBERTa, Longformer, or DeBERTa could address BERT's limitations, particularly in handling long-text essays or intricate sentence structures. Future research should focus on comparative analysis and optimization of various models to enhance the reliability of AES systems further.

Finally, the potential of AES technology to evolve beyond score generation into a tool for providing actionable feedback is significant. This advancement could play a crucial role in helping learners continuously improve their writing skills by identifying strengths and areas for improvement.

BERT, as a Bidirectional Encoder Representations from Transformers model, represents a breakthrough in natural language processing due to its ability to process contextual information in both forward and backward directions. This dual-context approach allows BERT to excel in capturing intricate relationships within sentences, making it particularly suitable for tasks requiring complex sentence structures and logical flow, such as essay scoring (Devlin et al., 2019). However, several challenges remain in applying BERT to AES.

First, data imbalance poses a significant limitation to model performance. The lack of sufficient advanced-level data in this study led to substantially reduced predictive accuracy for this category. Addressing this issue requires strategies such as synthetic data generation, transfer learning, or data augmentation techniques. Horbach et al. (2024) demonstrated the efficacy of multilingual dataset construction in resolving similar challenges, providing a reference for building large-scale Korean AES datasets encompassing diverse learner levels and topics. Incorporating detailed evaluation criteria for logical structure and creative thinking within these datasets could further enhance reliability.

Second, lack of interpretability in deep learning models can hinder the adoption and trustworthiness of AES systems. The "black-box" nature of these models reduces transparency in scoring processes, which is a critical concern in educational contexts (Kumar & Boulanger, 2020). Introducing Explainable AI (XAI) into AES systems could visually represent how scores are assigned, thereby fostering trust and comprehension among educators and learners. For

instance, integrating XAI into the BERT-based AES model used in this study could provide visual feedback to learners on their writing strengths and weaknesses.

Third, leveraging advancements in deep learning architectures is a crucial avenue for enhancing AES performance. Models such as RoBERTa and Longformer outperform BERT in processing long-form texts, offering significant advantages in computational efficiency and optimized attention mechanisms. Future studies should integrate these models into Korean AES systems to determine the most effective architecture for handling extended essay inputs. This approach could significantly improve the precision and timeliness of learner feedback systems.

Fourth, development of real-time feedback systems maximizes the educational potential of AES. Attali & Burstein (2006) highlighted that automated scoring systems could provide instantaneous feedback, enabling learners to identify and address weaknesses immediately. The BERT-based model developed in this study could be integrated into such a feedback system, offering learners actionable insights for improvement. Implementing personalized feedback through continuous updates of learner data represents a critical area for future expansion.

By addressing these directions, future research can significantly enhance the accuracy, transparency, and practicality of AES systems, ensuring their widespread adoption and implementation in educational settings.

Acknowledgements

This is a revision and reconstruction of the master's thesis by the Kim Su Rim (2024).

References

Alikaniotis, D., Yannakoudakis, H. & Rei, M. (2016). Automatic text scoring using neural networks. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 715–725.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. Journal of Technology, Learning, and Assessment, 4, 1-30.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M.D. Shermis & J.C. Burstein (Eds.), Handbook of Automated Essay Evaluation: Current Applications and New Directions (pp. 181-198). New York, NY: Routledge.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Burstein, J., Tetreault, J. & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), Handbook of Automated Essay Evaluation: Current Applications and New Directions (pp. 55-67). Routledge.

Cho, H., Yi, Y., Im, H., Cha, J. & Lee, C. (2021). Automatic score range classification of Korean essays using deep learning-based Korean language models -The case of KoBERT & KoGPT2-. Journal of the International Network for Korean Language and Culture, 18, 217-241.

Choi, S. (2018). A study on the patterns of source text borrowing performance in the integrated writing of the secondary students focusing on citations and paraphrasing. Research on Writing, 38, 201-245.

Choi, S. (2023). A study on the development of Korean language and essay test items for Korea college scholastic ability test. Journal of CheongRam Korean Language Education, 91, 135-178.

Clark, K., Luong, M. T., Le, Q. V. & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171-4186.

Dikli, S. (2006). An overview of automated scoring of essays. Journal of Technology, Learning, and Assessment, 5. Retrieved from http://www.jtla.org.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H. & Smith, N. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305.

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A. & Cho, Y. -I. (2024). Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. Mathematics, 12, 3416. https://doi.org/10.3390/math1221341.

Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., Lee, T. Y., Hong, H., Kim, J., Ahn, S.-Y. & Oh, A. (2024). LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. arXiv Preprint, arXiv:2310.05191. https://doi.org/10.48550/arXiv.2310.05191.

Han, S., Yu, D., On, B. & Lee, I. (2023). Empirical study on the loss functions of contrastive learning-based multi-scale BERT model for automated essay scoring. Journal of KIIT, 21, 51-63.

Horbach, A., Pehlke, J., Laarmann-Quante, R. & Ding, Y. (2024). Crosslingual content scoring in five languages using machine translation and multilingual transformer models. International Journal of Artificial Intelligence in Education, 34, 1294-1320. https://doi.org/10.1007/s40593-023-00370-1.

Hussein, M. A., Hassan, H. & Nassef, M. (2019). Automated language essay scoring systems: A literature review. PeerJ Computer Science, 5, e208.

Hussein, M. A., Hassan, H. A. & Nassef, M. (2020). A trait-based deep learning automated essay scoring system with adaptive feedback. International Journal of Advanced Computer Science and Applications, 11, 287-293. https://doi.org/10.14569/IJACSA.2020.0110516.

Isaacs, T., Zara, C., Herbert, G., Coombs, S. & Smith, C. (2013). Key Concepts in Educational Assessment. SAGE Publications.

Kim, D., Han, S., Park, S., Kim, Y., Park, S., Jo, J., Choi, S., Lee, Y., Yu, D. & On, B. (2023). Training set creation method for essay creativity score prediction models. Journal of KIIT, 21, 69–82.

Kim, D., On, B. & Jeong, D. (2023.11.). Comparison of BERT-based model performance for automated Korean essay scoring. Proceedings of KIIT Conference 2023, 435–439.

Kim, S. (2022). A Study on Automated Essay Evaluation Method for Argumentative Writing Task Using Deep Learning-Based Natural Language Processing Technique -Based on Model of Collaborative Scoring between Teacher Scorer and Machine Scorer (Unpublished Doctoral Dissertation). Korea National University of Education Graduate School, Cheongju, Republic of Korea.

Kim, S. (2024). Development of a Deep Learning Model for Automated Grading of Korean Essay (Unpublished Master's Thesis). Korea National University of Education, Cheongju, Republic of Korea.

Kim, S., Park, J. & Choi, S. (2020). Validation study of Q-matrix for argumentative writing assessment using cognitive diagnostic model. Research on Writing, 45, 181–222.

Kumar, V. & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. Frontiers in Education, 5, Article 572367. https://doi.org/10.3389/feduc.2020.572367.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259–284. https://doi.org/10.1080/01638539809545028.

Landauer, T. K., Laham, D. & Foltz, P. (2003). Automatic essay assessment. Assessment in Education: Principles, Policy & Practice, 10, 295–308. https://doi.org/10.1080/0969594032000148154.

Lee, J., Park, J. & Shon, J. (2022). A BERT-based automatic scoring model of Korean language learners' essay. Journal of Information Processing Systems, 18, 282-291.

Lee, K. (2021). Do it! Natural Language Processing using BERT and GPT. Easyspublishing.

Lee, S. (2022). A Study on the Development of Korean Language Learning and Scoring Tools applying Artificial Intelligence. Grammar Education, 44, 29-54.

Lee, Y. & Park, K. (2022). Exploring a way to build an AI automated essay scoring model with insufficient data. Journal of Education & Culture (JOEC), 28, 25–42.

Lee, Y., Choi, Y. & Lee, S. (2023). The development study of an automated scoring program for Korean essays, PASTA-I. Journal of Educational Evaluation, 36, 711–730.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5, 115–133.

Nguyen, H., & Dery, L. (2016). Neural networks for automated essay grading. CS224d Stanford reports, 1-11.

Nitko, A. J. & Brookhart, S. M. (2007). Educational Assessment of Students (5th ed.). Pearson.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 43–54). Lawrence Erlbaum Associates Publishers.

Park, H. (2019). Do it! Introduction to deep learning by honestly coding. EasysPublishing.

Park, H., Kim, S., Kim, K., Lee, M., Kim, K. & Kim, J. (2019). Substantializing Methods of Restricted and Extended Response Essay Assessment through Enforcing the Instruction-Assessment Alignment. Korea Institute of Curriculum and Evaluation Research Report RRE 2019-6.

Park, H., Kim, S., Kim, K., Lee, M., Kim, K. & Kim, J. (2019). Substantializing Methods of Restricted and Extended Response Essay Assessment through Enforcing the Instruction-Assessment Alignment. Korea Institute of Curriculum and Evaluation Research Report RRE 2019-6.

Park, J. & Choi, S. (2023). A study on the development of automated Korean essay scoring model using random forest algorithm. Brain, Digital, & Learning, 13, 131–146.

Park, J., Choi, S. & Kim, T. (2019). Development of empirically-derived and descriptor-based diagnostic (EDD) checklist for the cognitive diagnostic assessment in argumentative writing. Research on Writing, 41, 131–160.

Park, J., Lee, S., Song, M., Lee, M., Lee, M. & Choi, S. (2022). A Study on the Development of Automated Scoring Method for Computer-Based Essay and Short Answer Question Type Assessment (I). Korea Institute of Curriculum and Evaluation Research Report RRE 2022-6.

Park, J., Song, M., Lee, S., Park, S., Choi, S. & Lee, M. (2023). A Study on the Development of Automated Scoring Method for Computer-Based Essay and Short Answer Question Type Assessment (II). Korea Institute of Curriculum and Evaluation Research Report RRE 2023-7.

Park, K. & Lee, Y. (2022). Deep Learning Algorithm Exploration for Automated Korean essay Scoring. Journal of Educational Evaluation, 35(3), 465-488. 10.31158/JEEV.2022.35.3.465

Pearson Languages. (2023). The role of AI in English assessment. Retrieved October 8, 2024, from https://www.pearson.com/languages/community/blogs/2023/05/the-role-of-ai-in-english-assessment.html.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E. & Kukich, K. (2000). Comparing the validity of automated and human essay scoring. ETS Research Report Series, 2000(2), i-23.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. Artificial Intelligence Review, 55. https://doi.org/10.1007/s10462-021-10068-2.

Ravichandiran, S. (2021). Getting Started with Google BERT: Build and Train State-of-the-Art Natural Language Processing Models Using BERT. Packt Publishing.

Song, M., Noh, E. & Sung, K. (2016). Analysis on the Accuracy of Automated Scoring for Korean Large-scale Assessments. The Journal of Curriculum and Evaluation, 19(1), 255-274.

Sun, C., Qiu, X., Xu, Y. & Huang, X. (2019). How to fine-tune bert for text classification?. In Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18 (pp. 194-206). Springer International Publishing.

Taghipour, K. & Ng, H. T. (2016). A neural approach to automated essay scoring. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1882-1891.

Yu, D., Kim, Y., Han, S. & On, B. (2023). CLES-BERT: Contrastive learning-based BERT model for automated essay scoring. Journal of KIIT, 21, 31–43.

Zhang, H. & Litman, D. (2018). Co-attention based neural network for source-dependent essay scoring. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 399–409). Association for Computational Linguistics. New Orleans, LA. https://doi.org/10.18653/v1/W18-0547.

Authors Information

Kim, Surim: Korea National University of Education, Graduate Student, First Author, https://orcid.org/0000-0002-1647-1638

Choi, Sookki: Korea National University of Education, Professor, Corresponding Author, https://orcid.org/0000-0002-7846-7204