Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

Development and Application of Web-based Machine Learning Program for Automated Assessment Model Generation

Minseok Choi[1] · Jaegul Choo[1] · Minsu Ha[2]*
[1]KimJaechul Graduate School of AI at KAIST
[2]Kangwon National Univesity

자동 평가 모델 생성을 위한 웹기반 기계학습 프로그램의 개발 및 적용

최민석[1] · 주재걸[1] · 하민수[2*]
[1]KAIST 김재철 AI대학원 · [2]강원대학교

*Corresponding Author: msha@kangwon.ac.kr

ABSTRACT

In order to enable customized learning using artificial intelligence, a scoring model that can evaluate students' responses is needed. Human-scored data and programming skills training artificial intelligence are required to generate scoring models. When teachers develop scoring models, it is very useful to secure many scoring models that can be used in schools. However, the biggest challenge for teachers to create scoring models is programming skills training artificial intelligence. Thus, in this study, we developed a web-based automatic assessment model generation program that can rapidly generate assessment models using graded descriptive responses using supervised learning, without programming. Web program development was created by dividing an analyzer, classifier, web development, and server for extracting features. By developing an assessment model with actual scored data using the developed program, it was determined that an assessment model is reliable. Using the developed program, teachers can create their assessments using scored data without prior programming experience. In addition, the developed program can be used to strengthen teachers' artificial intelligence capabilities.

Key words: Constructed response assessment, artificial intelligence, machine learning, scoring model

Brain, Digital, & Learning
2022, Vol. 12, No. 4, 567-578.
https://doi.org/10.31216/BDL.20220034

Received: October 25, 2022
Revised: November 15, 2022
Accepted: November 21, 2022

Introduction

The importance of constructed response assessment has been emphasized in the educational field, including science education. This is because of the significance of constructed response assessment in identifying the various competencies required by future society (Ha et al., 2019). For instance, learning strategies for high scores on a multiple choice assessment are distinct from those for high scores on a constructed response assessment. For multiple choice assessment, students practice the function of distinguishing between correct and incorrect answers, whereas in constructed response assessment, students extract knowledge and explain it in sentences; therefore, the ability to use knowledge in constructed response assessment is similar to the functions required in our daily lives, including when problem-solving skills and communication abilities are performed (Ha et al., 2019). As emphasis is placed on communication skills such as writing, constructed response assessment is recognized as an essential skill for enhancing communication competency. Accordingly, the ratio of constructed response assessment items has expanded even at schools. In particular, the expansion of constructed response assessment is primarily motivated by the fact that limited thinking and memorization learning, which are primarily employed by students in multiple choice assessment, prevent them from developing diverse competencies, such as creativity (Kim & Seong, 2009).

Another reason the constructed response assessment is emphasized is that it excels in diagnostic functions for learning. Assessment is acknowledged as an essential component of learning (Ha et al., 2019; Wiliam, 2011). Diagnosis is an essential step in the learning process, and it is necessary to assess the actual level of development of students and organize classes accordingly. It is well-known that a constructed response assessment is beneficial based on a diagnostic perspective. In the multiple choice assessment, the student selects the response presented by the question developer that most closely matches his or her thoughts. The choice of the student may be the student's idea, but this cannot be inferred. It was chosen as the most plausible of the multiple responses provided, despite it not being in total agreement with the individual's thoughts. However, this idea can be attributed to the student in the constructed response assessment because the student directly writes the response (Opfer et al., 2012).

Another aspect of the important discussion of assessment is the feedback provided to students after the assessment. The assessment results indicate that providing feedback to promote student learning is an essential learning process. Shute (2008) emphasized that feedback can assist students in developing a variety of learning strategies by increasing their motivation to learn, making them aware of their current and desired skill levels, and enhancing their understanding of their abilities. There are several ways to provide immediate and slow feedback. In some instances, delayed feedback is effective; however, fast feedback has a greater educational impact than delayed feedback (Lee & Sohn, 2018). Immediate feedback is more difficult than delayed feedback because student responses must be analyzed. Therefore, we require an automated scoring system powered by artificial intelligence (AI) that provides quick feedback. An online-based automated scoring system that assesses students' understanding of evolutionary ideas and outputs the findings was created by Moharreri et al. (2014). Recently, an artificial intelligence (AI)-based formation assessment tool for Korean has been developed (Ha et al., 2019). This study also briefly introduces the WA^{3}I program, which automatically evaluates more than 100 items. Although possibility of AI in creating an effective analysis of assessment results and immediate feedback has been highlighted, AI assessment systems are still viewed with skepticism. For instance, there is a concern regarding the number of questions that must be created, as well as a concern regarding the scoring model.

AI-based automated scoring of constructed response assessment were developed by collecting student responses and employing a generated assessment model from an AI training. The data of students' responses were scored according to the scoring criteria. In order to score a constructed response assessment using AI, a scoring model must be trained in advance, and both the classification of the response and the training process of the model are necessary. It is challenging for AI researchers to collect student responses and teacher scoring data, but this information can be generated daily in the educational field such as school, where constructed response assessment is being used. Consequently, the number of AI assessment models can be significantly increased by utilizing teacher-scored data. However, it is difficult for teachers to acquire the computer programming skills required to develop and utilize assessment models and perform assigned tasks. It is possible to present a cooperative model between teachers and programmers, but close collaboration between them is required because teacher confirmation is necessary to confirm the accuracy of the assessment model. Therefore, the most efficient solution is to develop a tool that enables teachers to easily train and analyze assessment models without prior programming experience. Consequently, the purpose of this study is to develop a web-based automatic assessment model generation program that trains AI using constructed response data that has been evaluated and to provide field and educational implications by demonstrating practical use cases. The following are the specific research issues:

(1) Development of a web-based application for the generation of automatic assessment models that train AI using scored data.

(2) Application of a web-based application for the generation of automatic assessment models.

Literature Reviews

Natural Language Processing and Artificial Intelligence

AI has been developed to effectively perform repetitive tasks. Therefore, AI can be used as an important tool for analyzing. The development of an AI assessment model utilizes natural language processing (NLP) technology to process language. NLP, which involves the process of analyzing a language and extracting its meaning, is an essential component of language analysis using AI and is commonly used to develop constructed response assessment models. An example of linguistic task is morpheme analysis, which involves identifying and analyzing the smallest units of meaning (morphemes) within sentences in natural languages. Another example is developing a model for categorizing the proper nouns (names of people, places, and organization) in sentences and in the case of homonyms and polymorphs, classification according to context is possible. As part of the NLP technology, the analysis of dependencies, such as perception, subjectivity, and objectivity, of dominant words in sentences has also been performed. Korean is generally considered to be challenging language for NLP, although there have been significant advancements in NLP technology for Korean in recent years. According to Kim (2019), word order is unimportant in Korean grammar, and it is difficult to analyze sentences or morphemes because a single word can generate numerous derivatives. Because the spacing rules have changed significantly and related data are insufficient or inconsistent, the distinction between plain text and interrogative sentences can only be determined by punctuation marks, and subject omissions make NLP challenging. Accordingly, a substantial amount of data is required for the development of Korean NLP technology.

AI must learn the assessment rules to scored responses. There are numerous ways to learn AI, and supervised learning has become increasingly popular in recent years. Supervised learning refers to learning in which a computer searches for the rules of human judgment by receiving human-judged data and then applies the rules to other data to mimic human judgment. If the points in descriptive responses have been assigned, users can enter the data into a computer and generate an assessment model using NLP and suitable classifiers. If the assessment model is evaluated and its accuracy is found to be high, the computer can evaluate the next task instead of a human.

Web-based Automated Assessment Using AI (WA^{3}I)

Web-based automated assessment using AI (WA^{3}I) is a formative assessment program that uses responses scored by supervised learning to create an assessment model and immediately evaluate input student responses. Currently, this service is available at www.wai.best. The prototype version of the WA^{3}I was announced in 2019 (Ha et al., 2019), and the WA^{3}I presented in this study is the final version that has been upgraded and includes multiple features.

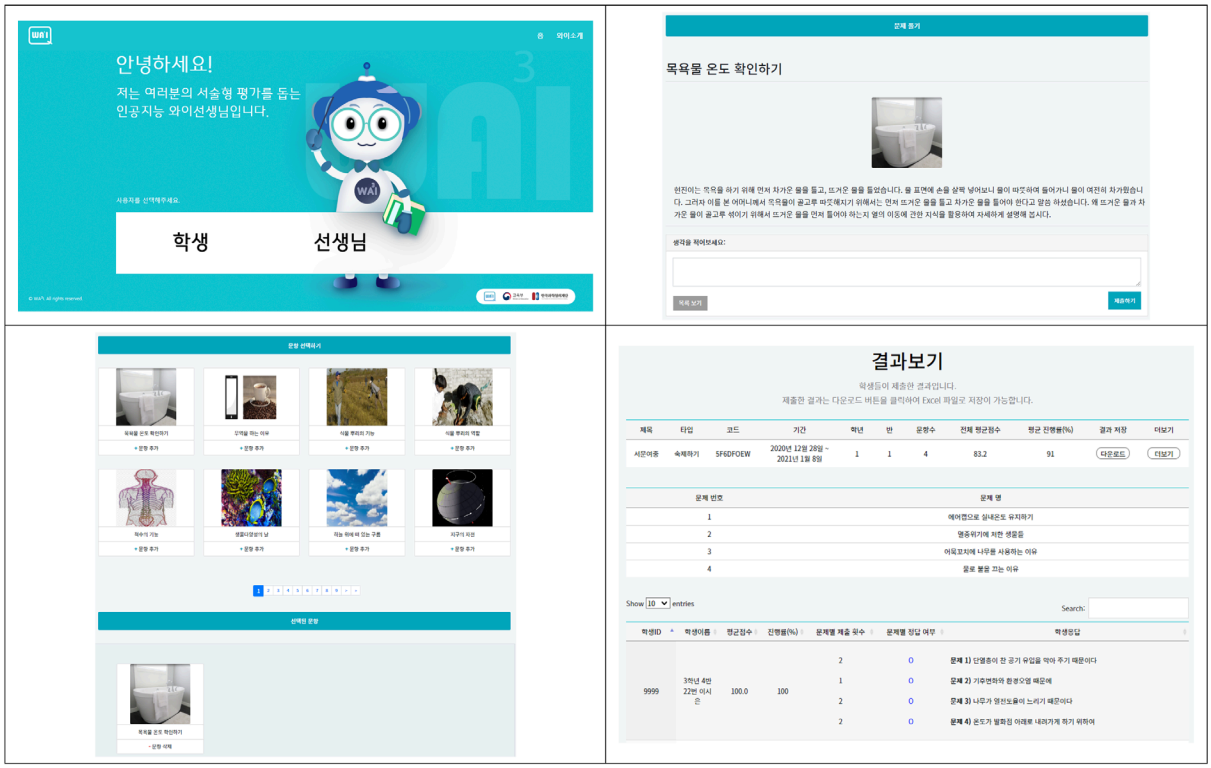


Fig. 1. Key elements of the WA^{3}I system

WA^{3}I (pronounced as "Why") is an abbreviation for web-based automated assessment using artificial intelligence, and the majority of questions begin with "why?". The users of WA^{3}I are students and teachers; the students are able to use AI without membership, meanwhile access for teachers required a registrations with a confirmed personal email with school domain address. Instead of registering as a member, the student accesses the teacher's questionnaire by entering the teacher-generated question code, and the system links the teacher's information with that of the student. The WA^{3}I was created as an assessment model that collected responses from approximately 5,000 elementary, middle, and high

school students. This system scored them based on a scoring criteria developed by experts, and trained the scoring results using a machine learning technique.

This system provides features for performing assessment practices, doing assessments, homework, self-assessment, and providing utilization guides for students. If a teacher selects a question, generates a survey, and distributes it to the students, the student enters the code and completes the survey. In the teacher mode of WA^{3}I, the terms 'doing assessment' and 'homework' serve different teaching purposes. Checking students' preconceptions and concepts strengthens the diagnostic function of assessment, while homework emphasizes the learning function rather than the diagnostic function, to provide students with opportunities to strengthen their learning. Therefore, doing assessment provides students with only one opportunity to enter a response without providing feedback on the correct answer, whereas homework allows students to write responses continuously until they receive positive feedback from AI. The WA^{3}I system's user description is provided as a YouTube video (https://www.youtube.com/channel/UC8FE7ExuI3tyAzoZ5), and for the WA^{3}I program to continue to evolve, new assessment questions and models must be added. However, as mentioned in the Introduction, a large amount of data must be scored for this process. The number of questions in the WA^{3}I system could dramatically increase if teachers developed and shared scoring models for AI using data generated in the educational field. In addition, if teachers use the web-based automatic assessment model generation program developed in this study, they can create and share scoring models quickly without programming knowledge.

Materials and Methods

Web-based Automatic Assessment Model Generation Program Design and Development

For the design of the web-based automatic assessment model generation program, the existing open program that served as a reference point was first identified. The reference program was LightSide, an open-source tool developed and utilized by Carnegie Mellon University researchers. This application was created using JAVA and was installed on a computer. Based on the contents of the reference, the basic design of LightSide and the preliminary system design were confirmed. Specifically, to resolve installation errors, JAVA updates, and operating system issues revealed by LightSide, this study was conducted online to avoid installation and operating system issues. Second, it was developed with a high degree of intuitiveness to make it easy for teachers to use. The process of training the assessment model comprised primarily three phases: feature extraction, classifier selection, and predictive stages of scoring new responses. Assuming that there is a grading assistant, it is structured similarly to the process of communicating and instructing the grading rules. Without the knowledge of supervised learning or machine learning, the process of AI programming with scored data to learn the model and scoring it using the trained model is easily understood. This course was developed through continuous consultation between computer and AI automatic assessment experts.

The assessment model is designed to permit only binary classification, that is, only 0 and 1 point data are used to train student responses. There are two reasons for binary classification. The diagnostic function of binary classification, which evaluates the presence or absence of an idea, is better than multiple classification. If there is a scoring model with a partial score (e.g., 2 points for correct answers, 1 point for partially correct answers, and 0 points for incorrect answers), then

the partial score is ambiguous because the criteria for each scorer are not clear. The feedback provided to students based on ambiguous assessments can lead to confusion. If there are multiple items in the assessment and each item generates a score to obtain a combined score and provide a partial score, it is more efficient to evaluate each item separately and produce an overall score. For instance, if an assessment criterion awards 2 points when both arguments and evidence are presented, 1 point when either argument or evidence is presented, and 0 points when neither is presented, the assessment is divided in half, and the total score is calculated by adding the two scores. In this situation, two binary classification assessment models, argument and evidences, were generated, and their results were compared and analyzed.

The number of data is the second reason for using binary classification. It was difficult for teachers to generate more than 500 responses during the assessment model training process. Multiple teachers may collect more than 1000 responses through collaboration, but 1,000 responses are not a large amount of data for supervised learning. When the number of responses to be used for supervised learning is small, and there is a partial score; then, the data entered into different categories will invariably demonstrate poor performance.

Web system development was divided into an analyzer, a classifier, web development, and a server for feature extraction. Okt (open Korean text) from the KoNLPy Python Korean NLP package was used as the analyzer to extract features. The classifier used five Scikit-Learn Python machine-learning library categories: naive Bayes, logical regression, support vector machine, decision tree classifier, and random forest classifier. The front-end of the system was built with React, its back-end with Python, and its application programming interface (API) with Flask and linked to the web. A database was constructed using SQLAlchemy to encrypt the associated login information. Finally, the server utilizes the KAIST KCLOUD2 server and container images based on Ubuntu. Here, data are stored, and when the assessment model is created and the session ends, the work file is automatically deleted.

Application and Assessment of a Web-based Program for Automatic Assessment Model Generation

The assessment model was created using a web-based automatic assessment model generation program (hereafter referred to as WA^{3}I-ML), and its performance was assessed by analyzing and utilizing it as actual research data. The data used to generate the automatic assessment model were the responses of 128 Korean major college students to the ACORNS question. Assessing contextual reasoning about natural selection (ACORNS) is a tool for systematically evaluating students' understanding of evolution (Nehm & Ha, 2011; Opfer et al., 2012). This is research data by Ha and Nehm (2014) that compare the evolutionary concept development patterns of Korean students with those of American students. Through using scoring data from 128 four ACORNS questions and 512 evolutionary responses, a scoring model was developed, and its accuracy was evaluated by comparing it to the study of Ha (2013), who developed an AI automatic assessment model for ACORNS. Owing to the small number of responses containing concepts, four concepts (variance, resource, difference in survival, and teleology) comprising over 20% of the responses were identified.

Results

Program for Web-based Automatic Assessment Model Generation

The web-based automatic assessment model generation program WA^{3}I-ML is a training course for model generation and a prediction process for classifying responses using the generated model, which is designed to perform the entire supervised learning procedure step-by-step. It was designed in such a way that anyone without programming experience could use it. Specifically, it was created for educational purposes by being limited to the automatic assessment of descriptive responses and not general text analysis. Classified responses for scoring model generation (scored data) initiate the processes of feature extraction, model training, model generation and storage, and the scoring and output of unclassified responses (responses to be scored) using the generated model. The generated model could be stored immediately and scored without training. Fig. 2 summarizes this procedure.

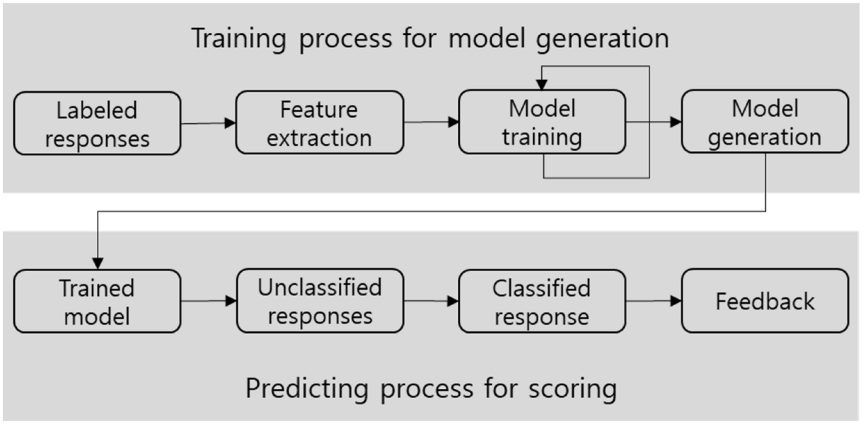


Fig. 2. WA^{3}I-ML's supervised learning process

The web-based automatic assessment model generation program WA^{3}I-ML is available at http://192.249.18.120. It can be used by entering a login credentials. WA^{3}I-ML is a web-based cloud service that, unlike LightSide, is not installed and utilized on a computer but is designed to be used directly on the web without the need for installation. Therefore, anyone with web access can use it immediately without installation. Each element of WA^{3}I-ML is described in the following table.

Examples of Web-based Automatic Assessment Model Creation Program

To generate an assessment model using the web-based automatic assessment model generation program developed in this study, and to determine how it can be utilized, an assessment model was developed using the actual scored data utilized by Ha and Nehm (2014). A total of 512 responses were divided into 384 for training of machine learning and 128 responses for testing. The students answered four questions about the evolution of snail venom, the wings of elm seed, penguins, and thorn of the rose. Two questions were about acquired traits and two questions were about degraded traits. In addition, two of the questions concerned the evolution of animals, whereas the other two concerned the evolution of plants.

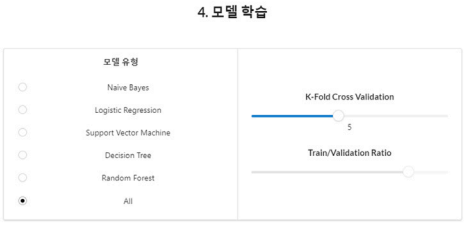
Table 2. Description of each element of the developed program

There is a description of WA^{3}I-ML as the initial screen of WA^{3}I-ML. 'Training New' will direct users to training assessment models, and 'Recall' will direct users to predicting scores immediately when the assessment model is stored. 'End Session' is a logout.

Uploading the file as the first step for using WA^{3}I-ML. The response must be entered in Excel, the student ID in column A, the response in column B, and the score in column C. The scores in column C are limited to 0 and 1.

It is the feature selection step of WA^{3}I-ML. An analysis unit, an N-gram, and a preprocessing process may be selected. Because the performance of the model may vary depending on the choice, the results should be compared in various combinations.

After extracting the features of WA^{3}I-ML, it is a feature assessment step to confirm how important each feature is in the assessment. If important words are at the top of the assessment, it can be judged that they are appropriately extracted.

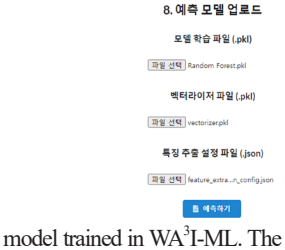


It is a classifier selection step for WA^{3}I-ML. It contains five widely used classifiers and can analyze them all at once. The k-fold cross validation is selected according to the number of data.

It is the result of cross-validation after WA^{3}I-ML learning. Select and save the model by referring to the values of accuracy, precision, recall, and F1 score.

It is a function that stores the model of WA^{3}I-ML. The model can be stored and used again for the next scoring.

It is a process of predicting data to be scored using the stored assessment model. Uploading a response that requires grading.



Uploading the model trained in WA^{3}I-ML. The assessment model is stored in three files, and each file must be uploaded.

This is the score predicted by the assessment model. There is a button at the bottom to download the assessment results.

The performance of machine learning models varies depending on the extracted features (part of speech, POS), nouns, etc.), N-gram (number of connected elements), and classifiers. To determine the model with the highest performance, it is necessary to train it in multiple combinations and then choose the model with the highest accuracy. In this study, the model with the highest training accuracy was chosen by combining POS and nouns, unigrams (one word), Bigram (two-word string), and five classifiers. A total of 128 assessment responses were scored, and their accuracy was compared to that of expert scoring. Currently, five accuracy values are used. First, accuracy is the ratio of how well the assessment model predicts the accuracy divided by the number of predictions. Precision is the ratio of what the assessment model classifies as possessing a concept to what actually possesses a concept. Recall is the proportion of how precisely the concept is predicted during the actual conceptual response. The F1 score is the average precision and recall. Cohen's kappa is the most commonly used measure of the degree of agreement between scorers. Several studies have provided criteria for kappa values, with the most prevalently employed by Landis and Koch (1977) being the range of 0.21–0.40 (either fair), 0.41–0.60 (appropriate, moderate), 0.61–0.80 (substantial), and 0.81–1.00 (almost perfect).

POS, nouns, and N-grams exhibited distinct patterns based on the four conceptual components. The decision tree classifier is generally excellent, whereas the random forest model are sometimes excellent. The optimal combination for the concept of mutation was 'noun + Bi-gram + random forest', 'noun + Bi-gram + decision tree' was optimal for the concept of resource, and 'POS + Uni-gram + random forest' and 'POS + Bi-gram + decision tree' were optimal for the concept of survival. This combination may change as the amount of training data increases. In the study by Ha (2013), for instance, a bi-gram was more useful for the concept of survival difference, but more than 7,000 training data were used.

Second, in terms of assessment accuracy, all were "almost perfect, " except for the concept of survival difference (kappa > 0.80). In a study by Ha (2013), who trained using English data, 4,048 responses and 1,913 responses were required to achieve a kappa accuracy level of 0.81 for the concept of difference in survival. In this study, a kappa value of 0.618 was confirmed using 384 responses, 191 of which included the concept of a difference in survival. In addition, errors in expert scoring were identified using this procedure. For instance, experts determined that there was no survival concept in the original data for the following two responses; however, the assessment model trained with WA^{3}I-ML found such a concept. The existence of this concept was confirmed by verification. Thus, WA^{3}I-ML can be used to improve the accuracy of the assessment, as it is possible to check for human scoring errors.

> The environment in which elm trees lived would have been disadvantageous to spread seeds, and this led to the evolution of winged seeds caused by accidental mutations to the present day.
>
> Elm trees without wings could not disperse seeds far away, so they would have spread their offspring around them and lived there. Then, the seed, a genetic trait that can spread far away from the seed of an elm, acquired the ability to make wings, and elm with winged seeds in the same environment would have evolved into a dominant species by spreading offspring far away. I will explain that.

Table 3. Accuracy of natural selection assessment models generated by the developed program

Concepts	Feature	N-gram	Classifier	Training (n=384)				Testing (n=128)				
				Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Kappa
Variation	POS	Uni	Decision Tree	0.943	0.923	0.920	0.920	0.922	0.875	0.913	0.894	0.832
	POS	Bi	Decision Tree	0.945	0.927	0.918	0.922	0.922	0.846	0.957	0.898	0.835
	Noun	Uni	Random Forest	0.927	0.962	0.823	0.886	0.961	1.000	0.891	0.943	0.913
	Noun	Bi	Random Forest	0.904	0.989	0.737	0.842	0.961	1.000	0.891	0.943	0.913
Resources	POS	Uni	Decision Tree	0.909	0.904	0.814	0.855	0.891	0.794	0.794	0.794	0.720
	POS	Bi	Decision Tree	0.899	0.855	0.835	0.844	0.875	0.750	0.794	0.771	0.686
	Noun	Uni	Random Forest	0.919	0.981	0.774	0.863	0.906	0.958	0.676	0.793	0.735
	Noun	Bi	Decision Tree	0.914	0.903	0.836	0.867	0.938	0.964	0.794	0.871	0.830
Differential survival	POS	Uni	Random Forest	0.818	0.821	0.809	0.814	0.813	0.811	0.754	0.782	0.618
	POS	Bi	Naive Bayes	0.794	0.765	0.842	0.800	0.734	0.667	0.807	0.730	0.473
	Noun	Uni	Naive Bayes	0.763	0.735	0.816	0.773	0.703	0.656	0.702	0.678	0.403
	Noun	Bi	Naive Bayes	0.771	0.747	0.811	0.777	0.688	0.635	0.702	0.667	0.374
Teleology	POS	Uni	Decision Tree	0.927	0.925	0.877	0.897	0.891	0.886	0.813	0.848	0.763
	POS	Bi	Decision Tree	0.943	0.931	0.920	0.922	0.914	0.951	0.813	0.876	0.811
	Noun	Uni	Decision Tree	0.880	0.854	0.821	0.835	0.883	0.902	0.771	0.831	0.742
	Noun	Bi	Decision Tree	0.875	0.873	0.781	0.823	0.883	0.867	0.813	0.839	0.747

Conclusions and Implications

As the importance of constructed response assessment increases, the number of AI learning tools that analyze and provide immediate feedback on student responses is also increasing. However, to continue the development of AI systems, it is necessary to create an assessment model that can continuously analyze new questions and responses. If teachers can create questions, collect and analyze student responses, and train AI assessment models, the automatic AI assessment system will continue to evolve. In the educational field, new questions are continuously formulated and student responses are collected. In addition, data for scoring student responses are continuously generated. For this course, we developed a web-based program for the generation of AI model for constructed response assessment that enable rapid access and analysis without specialized programming knowledge. Additionally, it was confirmed that an assessment model was available and developed using actual data.

Recent findings indicate that the application of artificial intelligence in the assessment of science education will rise dramatically. In addition, a significant number of studies are developing automated scoring models utilizing artificial intelligence. Using machine learning, Zhai et al. (2022) created a scoring model that assesses arguments. Maestrales et al. (2021) created a scoring models of chemistry, and physics with great scoring accuracy using supervised learning. Zhai et al. (2020) anticipated that the utilization of machine learning in science education would expand substantially. Changes in the digital-based educational environment are anticipated to considerably increase the usage of automatic scoring, despite the possibility of a number of issues. Zhai et al. (2021) confirmed the effect of various factors on scoring in an automatic scoring study using machine learning. An expert who can check scoring need to continually monitors the scoring outcomes of machine learning to mitigate this effect. In other words, teachers must regularly review the scoring

results produced by artificial intelligence. These recent changes in education enhance the value of the program built for this study. An environment in which teachers can study automatic scoring without knowledge of artificial intelligence programming will accelerate automatic scoring research using artificial intelligence more quickly.

Using the program developed in this study, it is possible to mass-produce an automatic AI assessment model. When one teacher created ten assessment models and 1,000 teachers participated, 10,000 scoring models were produced. This process may provide opportunities for teachers to improve their assessment, AI, and digital skills. Sharing the assessment model created with other educators may reduce their workloads. The assessment model can be rapidly developed by gathering responses from multiple groups to the same question. In addition, constructed response assessments are anticipated to spread and transform schooling.

Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2021S1A5A2A03061991)

References

Ha, M. (2013). Assessing Scientific Practices Using Machine Learning Methods: Development of Automated Computer Scoring Models for Written Evolutionary Explanations (Unpublished Doctoral Dissertation). The Ohio State University, Ohio, United States.

Ha, M., & Nehm, R. H. (2014). Darwin's difficulties and students' struggles with trait loss: Cognitive-historical parallelisms in evolutionary explanation. Science & Education, 23, 1051-1074.

Ha, M., Lee, G. G., Shin, S., Shin, S., Shin, S., Shin, S., ... Park, J. (2019). Assessment as a learning tool and utilization of artificial intelligence: WA3I project case. School Science Journal, 13, 271-282.

Kim, K. H. (2019). Kim's Deep Learning Camp for Natural Language Processing. Seoul: Hanbit media.

Kim, W. D., & Seong, J. E. (2009). Fundamental limitations and challenges of creative human resources development. STEPI Insight, 32, 1–20.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159-174.

Lee，B., & Sohn, W. S. (2018). Meta-analysis of feedback effects: Differences by feedback, learning tasks and learner characteristics. Journal of Educational Evaluation, 31, 501-529.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. Journal of Science Education and Technology, 30, 239-254.

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: An online formative assessment tool for automatically evaluating written evolutionary explanations. Evolution: Education and Outreach, 7, 1-14.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. Journal of Research in Science Teaching, 48, 237-256.

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: Knowing what students know about evolution. Journal of Research in Science Teaching, 49, 744-777.

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78, 153-189.
Wiliam, D. (2011). Embedded Formative Assessment. Bloomington, IN: Soluton Tree Press.
Zhai, X., Yin, Y., Pellegrino, J., Haudek, K., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. Studies in Science Education, 56, 111-151.
Zhai, X., Haudek, K. C., & Ma, W. (2022). Assessing argumentation using machine learning and cognitive diagnostic modeling. Research in Science Education, 1-20.
Zhai, X., Shi, L., & Nehm, R. H. (2021). A meta-analysis of machine learning-based science assessments: Factors impacting machine-human score agreements. Journal of Science Education and Technology, 30, 361-379.

Authors Information

Choi, Minseok: KimJaechul Graduate School of AI at KAIST, Graduate Student, First Author

Choo, Jaegul: KimJaechul Graduate School of AI at KAIST, Professor, Co-author

Ha, Minsu: Kangwon National Univesity, Professor, Corresponding Author