Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

Development of an Unsupervised Learning-Based Automated Evaluation System of Descriptive Assessment

Minsu Ha[1], Jinseon Choi[2], Daniel Rim[3], Sungchul Choi[2], Jaegul Choo[3], Chol Kyun Shin[4], Jisun Park[5*]
[1]Seoul National University
[2]Pukyong National University
[3]Korea Advanced Institute of Science & Technology
[4]Kangwon National University
[5]Ewha Womans University

비지도 학습에 기반한 서술형 문항 자동 평가 시스템 개발

하민수[1], 최진선[2], 임다니엘[3], 최성철[2], 주재걸[3], 신철균[4], 박지선[5*]
[1]서울대학교, [2]부경대학교, [3]한국과학기술원, [4]강원대학교, [5]이화여자대학교

*Corresponding Author: Jisun Park, jpark29@ewha.ac.kr(Corresponding Author's e-mail)

ABSTRACT

Current research on automatic scoring using traditional supervised learning methods cannot grade responses to questions that are newly generated or created spontaneously. Additionally, relying on pre-developed questions and their corresponding scoring models for lesson planning may limit the creativity and diversity of instruction. This study proposes a method that can quickly evaluate student responses and generate feedback without the need for pre-developed models. We introduces the SAAI system, which employs unsupervised learning techniques to instantly create scoring models based on student responses, thereby generating evaluation and feedback information. The SAAI system complements the automatic scoring of traditional supervised learning methods and supports scoring for a wide range of newly generated questions. This research elucidates the principles and significance of this system.

Key words: unsupervised learning, automatic evaluation, descriptive assessment, AI-based assessment

OPEN ACCESS

Brain, Digital, & Learning
2023, Vol. 13, No. 4, 339-351
https://doi.org/10.31216/BDL.20230021

Received: November 21, 2023
Revised: December 11, 2023
Accepted: December 14, 2023

Introduction

Constructivism learning is an educational approach where the learner synthesizes and comprehends new information provided by educators, building upon existing knowledge. This pedagogy underscores the active engagement of learners in constructing new information, as opposed to passively receiving it, by leveraging their prior experiences and knowledge (Bada & Olusegun, 2015; Fosnot, & Perry, 1996). Within constructivism learning, the learner's contextual understanding of knowledge is prioritized to enhance its relevance, facilitating a deeper comprehension, stimulating interest in the educational material, and aiding in the retention of information over extended periods. The significance of constructivism learning stems from its facilitation of a learner-centric educational paradigm. As learners engage with new information within a framework that resonates with their personal understanding and experiences, the learning process becomes more personalized, potentially elevating the significance and the intrinsic value of the educational experience (Campbell et al., 2007). Consequently, this learner-driven approach is conducive to assimilating complex notions and bolstering higher-order cognitive processes such as problem-solving and creative thinking. A critical component of constructivism learning is the assessment of students' pre-existing conceptual understanding, their learning trajectory, and the nature of the knowledge or skills acquired as a result (Fosnot & Perry, 1996). Diagnostic evaluation, which scrutinizes the learning process, constitutes the cornerstone of constructivism pedagogy.

The significance of the diagnostic function of assessment, analogous to that in constructivism learning, is supported by an important theoretical backdrop: the understanding-focused curriculum. This framework has influenced key directions in the curriculum revisions of 2015 and 2022 in Korea. An understanding-focused curriculum aims for profound comprehension and application, not just rote memorization of knowledge. Deep understanding is developed by linking individual facts and concepts, and by understanding the interactivity among them (Han & Lee, 2017). In the process of applying their knowledge to solve problems in new contexts, students enhance their related skills, experience complex thinking, and perform higher-order cognitive activities such as problem-solving and knowledge transfer (Nehm & Ha, 2011). Diagnostic evaluation, which assesses the learning process, maintains its importance within an understanding-centered curriculum. It is necessary to determine whether students are simply memorizing information or are contextualizing and understanding knowledge in a transferable manner. Such evaluations enable educators to support the students' learning processes and provide appropriate feedback to further improve their understanding. Given these considerations, the assessment of understanding levels is a key component in the successful execution of an understanding-focused curriculum.

In the context of both constructivism learning and the understanding-focused curriculum, the role of diagnostic assessment in appraising students' learning processes is paramount. With the recent 2022 curriculum revision, there has been an increased emphasis on diagnostic and formative assessments—assessments for learning—in classroom settings. Moreover, the assessments that need to be emphasized are descriptive assessments with robust diagnostic capabilities. Descriptive assessments inherently possess considerable advantages for effectively diagnosing the state of students' learning (Opfer et al., 2012). Such assessments can clearly reveal the extent of students' understanding of knowledge and their application of it. Rather than mere rote memorization or pattern recognition, students are required to think independently about the problems presented and to express their thoughts in written form. In this process, students employ their knowledge integratively, convey complex ideas or concepts, and provide a deep understanding of what they know and what they do not (Opfer et al., 2012). The analysis of students' descriptive responses enables educators to comprehend the structure of a student's knowledge clearly, to gain profound insights into their learning status, and to more easily monitor students' progress and personal challenges.

Moreover, within curricula that emphasize problem-solving, descriptive assessments are particularly significant due to their role in the learning process. Descriptive assessments effectively bolster students' problem-solving capabilities. While multiple-choice assessments may focus on the memorization of information or the ability to find the correct answer, descriptive assessments underscore students' ability to tackle complex problems, select necessary information, and present logical solutions. In these processes, students are compelled to use creative and critical thinking, thereby enhancing a variety of competencies (Ha, 2013). Additionally, descriptive assessments can improve communication skills, which are central to problem-solving. Through descriptive assessments, students learn how to articulate their thoughts and ideas clearly and effectively. They also practice how to provide logical evidence to support their arguments and persuade others of their perspectives. This process intensifies skills that are essential to problem-solving, such as communication, writing, effective argumentation, and the ability to understand and evaluate problems and contexts.

While the educational effects of descriptive assessments are significant, they present limitations due to the extensive time and effort required for evaluation and feedback. Feedback plays a crucial role in enhancing educational outcomes within the learning process. Immediate feedback, in particular, is highly beneficial to learning. It strengthens a student's understanding of the task they have just performed and provides an opportunity to promptly correct any mistakes. Immediate feedback helps learners clearly comprehend the connection between their actions and outcomes, thus fostering reflective thinking and ultimately being vital for learning new concepts or skills. Moreover, immediate feedback plays an important role in maintaining learners' motivation. The prompt recognition of successful outcomes boosts learners' confidence and reinforces a positive attitude towards learning. Considering these aspects, immediate feedback can be seen as an effective strategy that educators can utilize to maximize the learning effect for learners (Ha et al., 2011; Moharreri et al., 2014).

The educational efficacy of descriptive assessments, as well as the limitations of immediate feedback, are addressed by the utilization of artificial intelligence (AI) in descriptive evaluation. The application of AI has been extensively conducted in the educational field over the past decades, with the automatic evaluation of descriptive responses playing a significant role (Shin & Shim, 2021). Initially used mainly for scoring simple short-answer and multiple-choice questions, the advancement of AI has spurred research into automated evaluation technologies for more complex descriptive responses (Zhai et al., 2020). The automatic evaluation of descriptive responses utilizing AI is based on a variety of AI technologies, including machine learning, natural language processing, and deep learning. Machine learning algorithms learn the patterns of students' descriptive responses using training data, while natural language processing analyzes these responses to identify the keywords or context that serve as criteria for evaluation. To date, the technology for the automatic evaluation of descriptive responses using AI has shown considerable potential.

A widely utilized method in the research of artificial intelligence-based descriptive evaluation is the supervised learning approach. This study aims to highlight the issues associated with AI's automated descriptive assessment using the supervised learning method. Supervised learning, one of the core learning methodologies of AI, involves training a model using known input and output data. This method is effective when there is labeled educational data available, such as target outputs that are equivalent to scores. Utilizing this method, AI can develop a model that automatically evaluates a student's descriptive response. The development of a descriptive response evaluation model begins with securing an extensive dataset of descriptive responses, which must include results evaluated by humans (labels). Subsequently, this data is used to train a supervised learning algorithm. During this process, the algorithm learns the relationship between inputs (student responses) and outputs (scores evaluated by humans). Once this relationship is learned, the algorithm can generate predicted scores for new student responses.

The descriptive response evaluation model utilizing supervised learning has the advantage of automating and standardizing the assessment process, enhancing objectivity and efficiency. Additionally, the model can swiftly process a large volume of student responses, potentially reducing the workload for educators. However, such models require a substantial amount of high-quality training data, and the accuracy of the trained model significantly depends on the quality of this data. A notable limitation is the difficulty in adapting to spontaneously generated questions and the inability to modify items according to specific teaching situations.

Training a scoring model with the supervised learning method necessitates considerable time and effort in preparing the scored data (Ha et al., 2011). The process of data collection, cleaning, classification, and finally labeling is complex. Securing high-quality data is a critical step in supervised learning. If proper care is not taken during data preparation, the algorithm's performance can significantly diminish, adversely affecting the efficiency and effectiveness of the final scoring model. For these reasons, creating a scoring model using supervised learning can be time-consuming and costly, making it challenging to respond to the numerous spontaneously generated questions in a classroom setting. While it may be possible to create scoring models centered on core concepts that are widely used across various educational contexts over a long period, many questions in the classroom are generated spontaneously, and based on these questions, students' learning is assessed. Supervised learning methods cannot quickly adapt to such spontaneously generated questions.

The second limitation of the supervised learning method is its inability to adapt to questions tailored to specific teaching situations. Assessments are a critical element of the educational curriculum, necessary to ascertain whether students comprehend and can apply the content they have been taught. Through such assessments, teachers can gauge students' learning achievements and determine the required educational support. Given the significance of this evaluation, it is particularly crucial that the content taught by the teacher is closely linked to the assessment items. If this connection is not ensured, students might disregard the teacher's lessons and instead participate in assessments based on content learned from tutoring centers or other sources. Such behavior, where students ignore the teacher's content and focus on external materials, contradicts the fundamental purpose of school education. This situation can lead to students not focusing on the lessons and may undermine the teacher's role. Moreover, if students disregard the lesson content, they may miss important messages or points the teacher intends to convey during the class, potentially hindering educational and academic growth.

Therefore, this study emphasizes the necessity for a close connection between the teacher's lesson content and the assessment items. By developing assessment questions based on their own teaching content, teachers can encourage students to concentrate on the lessons and thoroughly understand the material. This approach assists students in making the most of the educational resources provided by the school and ensures that the teacher's lessons play a significant role in the students' learning process.

This study aims to develop a system that leverages unsupervised learning to overcome the issues associated with supervised learning. By employing unsupervised learning techniques, the system is designed to identify elements related to correct answers within students' responses and to generate evaluation and feedback materials. Additionally, the study aims to introduce the principles of the system, explain how it is used, and describe its operating procedures.

Overview of The Developed System

Unsupervised Learning-based Automated Evaluation System, SAAI (Scoring Assistant Using Artificial Intelligence)

SAAI (Scoring Assistant using Artificial Intelligence) is a technology that improvisationally scores student responses without the need for a pre-established scoring model, basing its assessment on the content of the responses themselves. As the English name implies, this technology does not predict scores using a scoring model for student responses; rather, it serves as a scoring support tool to assist teachers in rapid assessment. Utilizing the principle of mutual reinforcement, this technology identifies key terms related to the correct answer within responses, and scores are based on these terms. The fundamental principle of this technology is that words related to the correct answer tend to recur in well-constructed responses and are highly interconnected. By leveraging this feature, it is possible to estimate the relevance to the correct answer and, in turn, quantify the level of a student's response. The quantification process uses an algorithm known as the mutual reinforcement principle, introduced by Zha (2002) for the purpose of summarizing documents. Zha's method was developed to extract key terms and phrases from multiple sentences within a document set to summarize the document. This method aims to identify words that occur frequently and across multiple locations in the document. The computation used to find key terms follows the principle of mutual reinforcement: sentences containing many significant words are likely to be close to a good response, and the words contained in a response close to a good one are also likely to be significant.

For instance, in a response to a question about the contraction and dilation of the iris, words such as "iris," "contraction," "pupil," and "expansion" would be key terms of a good response. In this response, "iris" and "pupil," being technical scientific terms with no synonyms, are likely to appear commonly in correct sentences. Words like "contraction" and "expansion" would score highly because they appear in the same sentences as "iris" and "pupil." Even if "contract" and "expand" are replaced with "narrow" and "enlarge," they would still score because they are contained in important sentences that include key terms like "iris" and "pupil." Thus, the scores of words and sentences are interconnected, allowing for conditions of mutual reinforcement. This process is significant because it identifies key terms in student responses, enabling the recognition of words used meaningfully by students in their responses, beyond those expected in the model answers by teachers. This is akin to the context where various linguistic features, unforeseen during the generation of scoring models in supervised learning processes, are employed.

To effectively use SAAI, there needs to be a certain ratio of correct to incorrect responses within the student response corpus. SAAI struggles to perform well if a question is so difficult that there are hardly any correct responses, or if all responses are correct. It also has difficulty with very open-ended questions. Therefore, if SAAI's performance is low, it may be due to a large disparity in the ratio of correct and incorrect responses, or because the question's responses are very open-ended. However, this aspect can be utilized to further extend the application of SAAI. In formative assessments, the lack of correct responses among students' answers could indicate that the question is not well-constructed. Likewise, a question that all students answer correctly is also not considered good, as it may be too easy. An appropriately challenging question, which would be considered good, might be one where about 40% of students are able to construct strong responses.

SAAI analyzes student responses to generate three types of information. First, it produces a list of significant words used in the students' responses. Teachers can review this list of key words to approximately assess the accuracy of SAAI's analysis. The scores of these key words are used to generate sentence scores, which are continuous rational numbers with many decimal places, rather than integers like 1, 2, or 3 points. While it is possible to automatically convert continuous numbers into several levels of scores based on specific criteria, this process is designed to be determined by the teacher. Teachers can use the student response scores generated by SAAI to categorize them into top, middle, and lower scores. Finally, SAAI collects information that could form exemplary answers based on the important words and responses with high SAAI scores. Utilizing such information from SAAI allows for quantitative feedback about the level of responses, qualitative feedback by providing words that could be reference points for thought among those not used by the students, and the creation of feedback using well-constructed exemplary answers suitable for the students' levels.

Method for System Development

System Development Environment

Using Google's Jamboard, an Information Architecture (IA) was created, which served as the foundation for designing User Interfaces (UI) and wireframes using Figma. For the web service implementation, the front-end development employed the Vue.js(v.2.6.10) framework. Vue.js, a JavaScript-based framework, offers features such as declarative UI, virtual DOM, reactive and component-based architecture. Each component was designed in a single-file format to ensure modularity and ease of maintenance. Backend development was carried out using the Express.js (v.4.16.1) framework based on the Node.js (v.16.14.0) environment. MySQL (v.8.0.29) was used for database construction.

Data exchange between the front-end, back-end, and database utilizes a JSON-based RESTful API. The artificial intelligence scoring and individual report generation functionality were implemented in Python 3. Upon receiving data, the AI scoring system cleanses it using natural language processing libraries such as NLTK, spaCy, and KoNLPy, before passing it to the scoring module. The scoring module then converts the scoring results into an Excel file for user delivery. The individual report generation functionality uses python-docx to create and return Word file format reports based on the user-entered scoring results Excel file. The generated report files are stored in AWS S3 using the Boto3 library, and the link to the stored file is provided to the user. Communication between each function and web page is facilitated through a FastAPI-based web API. The deployment of the developed system utilizes AWS EC2 servers. Inside the EC2 instance, front-end, back-end, database, AI scoring, and individual report generation functionalities were containerized using Docker (v.20.10.11). Docker-compose(v.1.29.2) is used to run the containers simultaneously and manage the system effectively. All code storage and version control were managed using Git and GitHub.

Components and Sequence of System Design

The design of SAAI system consists of eight components in sequence, reflecting the needs of the user, the teacher. Fig. 1 shows the componets and sequence of the SAAI system and the gray area indicates the section where the teacher can increase the accuracy of the results by reviewing them.

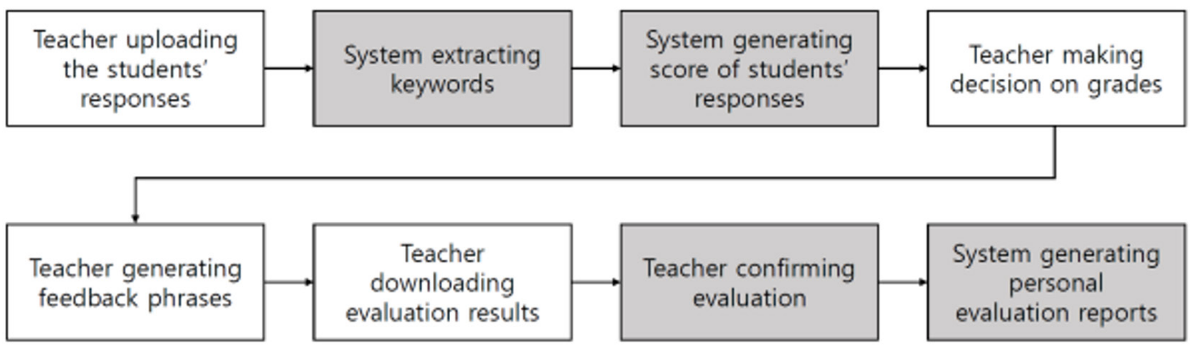


Fig. 1. The components and sequence of the SAAI system

Initially, student responses are collected using survey tools and then saved in an Excel file. The student responses analyzed by the SAAI system are formative assessments conducted within the classroom and evaluation items submitted as assignments, which are low-stakes assessments that can be collected as text that can be stored on a computer. This system is intended for evaluative purposes to aid learning, and it is expected that it will not be used for high-stakes evaluations.

When the response file is uploaded, as previously described, the system automatically extracts important keywords. The important keywords are designed to be sorted based on their scores. The teacher checks whether the keywords have been appropriately extracted, which helps to gauge whether the SAAI system can grade the bundle of responses. If the keywords deemed important by the teacher are not extracted, it signifies that the SAAI system has limitations in evaluating that particular bundle of responses. In such cases, the teacher has no choice but to score manually. This scenario typically arises when certain patterns are not apparent in student responses, often when the correct or good answers are very scarce in the bundle of responses, or when the answers to the questions are highly diverse and do not converge to a particular response. The teacher can manually select keywords to be used for score generation. Once the keywords are appropriately selected, response scores are generated. After scoring each student's response, they are sorted from highest to lowest score. By examining the responses from highest to lowest score, the teacher can make a preliminary assessment of the SAAI system's accuracy. If the system has high scores for responses that are, in reality, not good according to the teacher's judgment, or if there are many good responses with low scores, the SAAI system is limited in grading that bundle of responses. If the SAAI scores generally arrange the students' scores appropriately, it can be assumed to have a degree of accuracy. Then, the score ranges for each response are determined, and quantitative feedback for each score range is generated. For example, student responses can be divided into three levels, with the highest grade receiving 3 points, the middle 2 points, and the lowest 1 point. However, more than such numerical scores, phrase-based feedback is encouraged. For instance, feedback for a 3-point response could be, "You have written the anwers requested by the problem conscientiously and well." This kind of phrased feedback is more approachable for students and is likely to have a higher educational effect. Once the SAAI system's evaluation results are finalized, they are downloaded as an Excel file. From the point of downloading the Excel file, the teacher manages the students' data on the computer. Upon opening the Excel file, the teacher can review the student responses and feedback phrases again. If well-written responses received low scores, the feedback for such responses can be adjusted. Once the evaluations and feedback phrases to be reported to the student are finalized, individual reports are generated.

The essence of the SAAI system's design lies in continually enabling teachers to judge the information provided by the SAAI system according to their own scoring criteria. Evaluation items are created by teachers, and the understanding of the evaluation criteria is most accurately judged by the teachers who created the items. The second key element is the individual report provided to each student. Individual reports offer both quantitative and qualitative feedback. Students can reflect on their written responses and learn from the quantitative feedback provided. They can also consider how to improve their answers by reviewing the keywords that were not included in their responses. This process fosters reflective thinking in students and offers them the opportunity to think about how they can provide better responses, thereby promoting learning. While the SAAI system could be used to generate a comprehensive score for a student, its intended direction is to provide rapid feedback to each individual student, fostering reflective thinking and guiding them towards a deeper understanding and learning.

Result and Discussion

The Layout and Description of a Web-Based System Interface

The SAAI system is designed for easy access and use by teachers. Developed as a web-based platform, it can be accessed in any environment with an internet connection. As previously described, when responses are uploaded, the system extracts key keywords and generates scores for student responses. Teachers can monitor this process to ensure the SAAI system is evaluating properly. Teachers can adjust their level of intervention accordingly. If teachers have time to check the SAAI evaluation results meticulously, a more rigorous assessment can be conducted; otherwise, a generally acceptable, albeit less stringent, level of assessment can be achieved. Once the SAAI evaluation results are finalized, individualized feedback can be created. A detailed description of the overall screen layout and functionality is provided in Table 1.

Table 1. Participants Distribution

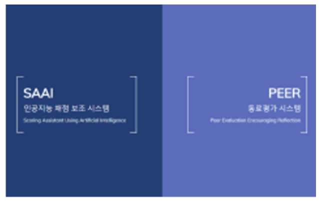
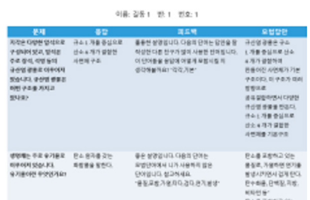
Component of system	Web screen of SAAI system	Description
-		The SAAIPEER site is designed to enable the use of both the artificial intelligence scoring assistance system SAAI and the peer review system PEER together. To evaluate students' responses and generate feedback information using artificial intelligence, one enters through SAAI. http://saaipeer.com/
Teacher uploading s tudents' responses		Before entering SAAI response analysis, a user must register. Registration can be done with a name, email, and password. After logging in, clicking 'Start Response Analysis' begins the analysis process.
		The data to be graded should be entered into Excel in a specified format. If there are multiple questions, they should be entered on separate sheets. Upon data upload, the total number of sheets is displayed, and the questions on each sheet, as well as the students' responses to them, are visible. After naming the scoring sheet, click next.
System extracting keywords		The SAAI system uses the principle of mutual reinforcement to determine the importance of keywords within responses, and sorts words accordingly. In this process, teachers check to see if words they consider to be part of the ideal answer are ranked highly. If the top-ranked words are highly relevant to the ideal answer, it is likely that the evaluation will be accurate. The default setting for analysis is set to evaluate based on the top 10 words, while the user-defined analysis allows the selection of keywords to be used in the evaluation, and adjustment of scores for semantically similar keywords.
		The next step is to select important missing keywords from each student response for provision and conclude the evaluation. The selected keywords not found in a student's response are provided in the feedback.
System ge nerating score of students' responses		When the analysis is performed, scores are generated for each student's response. By clicking on the scores and sorting from high to low, it's possible to review the responses starting with the highest scores to estimate the accuracy of the evaluation. This process is repeated for the number of questions.
Teacher making decision on grades		Based on SAAI's scores, one defines the ranges for feedback phrases. After setting the score ranges with reference to a graph of SAAI's scores, appropriate feedback phrases for each range are generated. The feedback ranges can have up to six levels, but more refined feedback ranges based on SAAI scores can be defined in Excel in the future.
Teacher generating feedback phrases		Once the evaluation of student responses is complete, the last sheet of the Excel file is used to enter each student's information by ID. Feedback tailored to each student is automatically generated based on the entered information. As individual reports contain students' names, they are generated and immediately deleted without being saved.

Table 1. Participants Distribution

Component of system	Web screen of SAAI system	Description
Teacher downloading and confirming evaluation results		In the exported Excel file, feedback phrases to be provided to students can be revised by sorting scores, word count, etc. If a score is low but the response is lengthy, it might indicate either a misunderstanding of the question or a potentially creative response, so those items are reviewed.
System generating personal evaluation reports		The feedback includes the question, missing important words, an explanation for the missing words, and an ideal answer. Although a fixed phrase is set for the explanation of the missing important words, it can be changed, and the ideal answer can either be crafted directly by the teacher or derived from responses with high SAAI scores.
		Individual reports are generated as MS-WORD files. These files can be opened and edited using various programs capable of handling MS-WORD. If there is a student who requires special feedback, that student's feedback can be reviewed and modified accordingly.

Discussion on The Educational Significance of SAAI System

The educational significance of the SAAI system lies in ensuring the agency, creativity, and diversity of teachers. As emphasized in the introduction, when teachers create questions tailored to their lessons, the existing supervised learning methods cannot support automatic evaluation. These conventional methods restrict teachers to pre-determined problems and scoring models, limiting the use of diverse questions in their classes. If we fail to develop a wide range of questions and corresponding scoring models, students in our country may only be able to solve a few problems supported by automatic scoring. This limitation hinders students' exposure to a variety of problems and restricts teachers from exploring different questions with their students. To enhance the diversity of class content and assessment items, teachers should be able to create and utilize questions of their choice. Consequently, there is a need to contemplate methods for quickly evaluating responses to such questions.

Recent studies have been actively developing automatic scoring models using supervised learning, extending beyond text to include assessments of drawings (Lee et al., 2023; Zhai et al., 2022). With sufficient budget, it is possible to develop scoring models for many questions, enabling students to experience a diverse range of problems (Zhai et al., 2021; Zhai et al., 2020). Supervised learning methods are crucial due to their accuracy and ability to generate immediate feedback. However, it is practically impossible to develop scoring models in advance for the multitude of questions continuously generated in classrooms. The SAAI system developed in this research holds significance in this regard.

Using supervised learning to develop automatic scoring models tends to result in higher accuracy, whereas using unsupervised learning may have limitations in certain cases. For example, it may not perform well when the questions are too challenging, leading to a higher rate of incorrect answers compared to correct ones. Additionally, when questions are highly open-ended, it can also face difficulties. Furthermore, it doesn't work well when all responses are correct.

Teachers or users should be aware of these limitations, and if the automated scoring system is not performing well, it can be an opportunity to examine whether the assessment items are too challenging or open-ended.

Therefore, it is necessary to harmoniously utilize both supervised and unsupervised learning approaches, rather than rigidly adhering to one or the other. Table 2 summarizes how supervised and unsupervised learning can be harmoniously utilized in response to the evaluated competencies and their importance. Various methods of automatic scoring should be proposed, developing numerous strategies to efficiently support teachers in handling descriptive questions. For instance, teachers can utilize SAAI for formative assessments, where they promptly assess students understanding of the content they taught during class through list of question items. As shown in Table 1, teachers can easily upload students' responses, check what kinds of keywords that students frequently used. Also teacher can make decision on grades based on the SAAI's scores and generate feedback phrases.

Table 2. Suggestions for the Use of Supervised and Unsupervised Learning Based on the Type of Question

Question type	Developing method
Key knowledge and competencies from a learning perspective	Developed by AI experts using extensive scoring data and rigorous supervised learning training, this model facilitates the assessment of crucial learning domains.
Numerous problems solved in everyday learning	Developed under the guidance of teachers, this model utilizes a smaller set of scoring data and machine learning software
Items that are either newly developed or created spontaneously	This is carried out using unsupervised learning methods similar to those used in the SAAI system, catering to the assessment of various spontaneous or newly created descriptive items.

Recently, LLMs (Large Language Models) such as Chat GPT have presented new challenges in automated scoring research. Large Language Models (LLMs) have recently been widely utilized in automated evaluation (Campos, & Shern, 2022; Dai et al., 2023). LLMs are recognized as powerful tools for assessment and feedback due to their ability to understand the context of a text, identify typographical errors, awkward expressions, and content elements unrelated to the answers to questions. However, limitations exist within LLMs. Firstly, while LLMs can extract general answers to questions posed by teachers from their data, they cannot ascertain the content conveyed by the teacher during class or the students' understanding of it. The knowledge formed through teacher-student interactions during class can only be identified in the students' responses. Therefore, the amalgamation of LLMs and SAAI can complete the process of evaluation and feedback. SAAI is capable of extracting key keywords from students' responses and, based on these, perform assessments and feedback grounded in the context of the lesson. LLMs can identify typographical errors and awkward expressions in student responses. Additionally, if feedback sentences are generated using generative AI, it would effectively reduce the workload of teachers.

Acknowledgements

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

Conflicts of Interest

The author(s) declared no conflicts of interest.

References

Bada, S. O., & Olusegun, S. (2015). Constructivism learning theory: A paradigm for teaching and learning. Journal of Research & Method in Education, 5, 66-70.

Campbell, R. J., Robinson, W., Neelands, J., Hewston, R., & Mazzoli, L. (2007). Personalised learning: Ambiguities in theory and practice. British journal of educational studies, 55, 135-154.

Campos, J. A., & Shern, J. (2022, May). Training language models with language feedback. In ACL Workshop on Learning with Natural Language Supervision. 2022.

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023, July). Can large language models provide feedback to students? A case study on ChatGPT. In 2023 IEEE International Conference on Advanced Learning Technologies (ICALT) (pp. 323-325). IEEE.

Fosnot, C. T., & Perry, R. S. (1996). Constructivism: A psychological theory of learning. Constructivism: Theory, perspectives, and practice, 2, 8-33.

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

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. CBE—Life Sciences Education, 10, 379-393.

Han, H. Lee, J. Y. (2017). Curriculum by Comparison with the Discipline-based Curriculum and the Understanding by Design. The Journal of Curriculum Studies, 35, 203-221.

Lee, J., Lee, G. G., & Hong, H. G. (2023). Automated assessment of student hand drawings in free-response items on the particulate nature of matter. Journal of Science Education and Technology, 1-18.

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.

Shin, D., & Shim, J. (2021). A systematic review on data mining for mathematics and science education. International Journal of Science and Mathematics Education, 19, 639–659.

Zha, H.(2002), Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering, In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 113-120. https://doi.org/10.1145/564376.564398

Zhai, X., C Haudek, K., Shi, L., H Nehm, R., & Urban‐Lurain, M. (2020). From substitution to redefinition: A framework of machine learning‐based science assessment. Journal of Research in Science Teaching, 57, 1430-1459.

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. Journal of Research in Science Teaching, 59, 1765-1794.

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

Minsu Ha : Seoul National University, Professor, First Author https://orcid.org/0000-0003-3087-3833

Jinseon Choi : Pukyong National University, Undergraduate student, Co-author https://orcid.org/0009-0005-8152-0378

Daniel Rim : Korea Advanced Institute of Science & Technology, Graduate student, Co-author https://orcid.org/0009-0008-2161-9650

Sungchul Choi : Pukyong National University, Professor, Co-author https://orcid.org/0000-0002-5836-3838

Jaegul Choo : Korea Advanced Institute of Science & Technology, Professor, Co-author https://orcid.org/0000-0003-1071-4835

Chol Kyun Shin : Kangwon National University, Professor, Co-author https://orcid.org/0009-0003-6959-605X

Jisun Park : Ewha Womans University, Professor, Corresponding Author https://orcid.org/0000-0002-7740-2046