Journal of The Korea Convergence Society Vol. 13. No. 3, pp. 11-22, 2022
e-ISSN 2713-6353
DOI: https://doi.org/10.15207/JKCS.2022.13.03.011

Error Analysis of the Latest Conversational Agent-Based Commercial Education Platform

Seungjun Lee1 Jaehyung Seo1 Changjun Park1 Heuiseok Lim2* 1Computer Science Department, Korea University,
Integrated Master's and Doctoral Program, 2Professor, Computer Science Department, Korea University

Error Analysis of Recent Conversational Agent-based Commercialization Education Platform

Seungjun Lee [1], Changjun Park [1], Jaehyung Seo [1], Heuiseok Lim [2*]
1Master & Ph.D. Combined Student, Department of Computer Science and Engineering, Korea University 2*Professor, Department of Computer Science and Engineering, Korea University

Abstract: Research and development utilizing various artificial intelligence technologies are currently underway in the education field. Among AI-based education tools, conversational agents, in particular, enable more effective language learning by integrating diverse AI technologies, such as speech recognition and translation, without being constrained by time or space. This paper analyzes trends in commercialized educational platforms with large user bases and those utilizing conversational agents for English language learning. The analysis revealed that the conversational agents on currently commercialized educational platforms have various limitations and issues. To analyze these specific issues and limitations, we conducted comparative experiments with state-of-the-art, pre-trained, large-scale conversational models. The Sensibleness and Specificity Average (SSA) human evaluation method was used to assess the similarity of the conversational agent's responses to human responses. Based on the results of the experiments, we propose the need for conversational models trained with large parameters, training data, and information retrieval capabilities to enhance effective learning.

Keywords: conversation, agent, deep learning, educational platform, chatbot, language fusion

Abstract Recently, research and development using various Artificial Intelligence (AI) technologies are being conducted in the field of education. Among the AI in Education (AIEd), conversational agents are not limited by time and space, and can learn more effectively by combining them with various AI technologies such as voice recognition and translation. This paper conducted a trend analysis on platforms that have a large number of users and used conversational agents for English learning among commercialized application. Currently commercialized educational platforms using conversational agent through trend analysis has several limitations and problems. To analyze specific problems and limitations, a comparative experiment was conducted with the latest pre-trained large-capacity dialogue model. Sensibleness and Specificity Average (SSA) human evaluation was conducted to evaluate conversational human-likeness. Based on the experiment, this paper propose the need for trained with large-capacity parameters dialogue models, educational data, and information retrieval functions for effective English conversation learning.

Key Words : Conversation, Agent, Deep Learning, Education Platform, Language Convergence

*This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)

Received December 3, 2021 Revised December 29, 2021
Accepted March 20, 2022 Published March 28, 2022

1. Introduction

Recent advancements in artificial intelligence (AI) technology have led to the convergence and development of various industries and entertainment sectors. Among these, the education industry is one area where AI integration is crucial [1]. In particular, in the post-COVID-19 era, online learning has become increasingly popular, and learners are being given more time for self-directed learning than ever before. Consequently, AI-integrated education platforms have naturally attracted more attention than ever before.

Artificial Intelligence in Education (AIEd) refers to education utilizing artificial intelligence. AIEd has recently been developing in conjunction with various AI fields, including natural language processing and speech recognition. AIEd can be broadly categorized into three categories: Intelligent Tutoring Systems (ITS), which are suitable for handling relatively structured problems; Dialogic Tutoring Systems (DBTS), which are suitable for handling semi-structured problems; and Inquiry Learning Systems (ELE), which are suitable for handling unstructured problems [2].

An Intelligent Tutoring System (ITS) provides an effective step-by-step learning process for fields with well-defined knowledge structures. For effective tutoring, ITS consists of a pedagogical model, a domain model, and a learner model. The pedagogical model is used as a model for effective teaching by professors or teachers, and a representative example is a learning management system such as an LMS (Learning Management System). An example of a domain model used as a topic model centered on learning content is Santa TOEIC1) using reinforcement learning.

There is KnowRe2) that uses AI to apply a drill down system . A representative example of a learner model is the Active Learning Forum of Minerva Schools, which was created to maximize interaction with learners in real time. The Dialogue-Based Tutoring System (DBTS) is an advanced form of ITS that

utilizes a chatbot called a conversational agent or Dialogue System to engage students in a learning conversation. DBTS The conversational agent used is based on Natural Language Understanding (NLU) and Natural Language Generation (NLG) technologies, both of which are natural language processing technologies. Learners input text or speech, and the conversational agent recognizes this, determines the learner's level of understanding, provides feedback to correct misconceptions, or outputs an appropriate response to the task [2].

An Exploratory Learning Environment (ELE) is a system that provides automated feedback to correct learners' incorrect learning outcomes, rather than following a set, step-by-step learning sequence. Representative AIEd technologies utilizing ELE include Grammar Error Correction (GEC) and Automated Essay Scoring (AES). In AI-enabled education (AIEd), interactive tutoring systems possess the following characteristics. First, while traditional

learning typically takes place in classrooms and is subject to spatial and temporal constraints, AIEd-integrated conversational agents allow for free learning without these constraints. Second, AIEd can be implemented by combining natural conversational interfaces with various AI technologies. Combining STT (Speech-to-text) and TTS (Text-to-speech) technologies with conversational agents can provide a conversational experience similar to a real person, while combining GEC can provide effective grammar correction feedback to learners [3,4]. Taking advantage of the characteristics of these conversational tutoring platforms, educational platforms that combine conversational agents and language learning have recently been released. However, considering current artificial intelligence technology, utilizing conversational agents as language learning educational platforms for

AIEd has several limitations. Most goal-oriented conversational agents require manual data entry into a knowledge base and are only applicable to specific languages or limited domains. Furthermore, open-domain conversational systems, known as chit-chat models or non-goal-oriented conversational systems, lack consistent personas, lack long-term memory and thus cannot consider previous conversations, provide non-specific answers such as "I don't know," and have difficulty responding properly when discussing other topics [5,6]. Due to these limitations, conversational agents have not been applied as language learning educational platforms to date.

1) https://www.riiid.co/

2) https://www.knowre.com/

There are difficulties in doing so [7]. This paper analyzed an English education platform utilizing a conversational agent among artificial intelligence-based education (AIEd). Existing studies on educational platforms have not included comparative experiments with large-scale conversational agent models. This paper conducted a human evaluation of an English education platform and a large-scale conversational model to suggest a plan for a future educational platform. Based on this, we analyzed the limitations and problems of the current English education platform using a conversational agent, and based on the improvement plan, we propose a utilization plan for the future development of an English education platform using a conversational agent.

2. Related research

2.1 Conversational Agents Conversational agents effectively enable human-computer interaction through dialogue. They utilize technology that allows computers to naturally understand, process, and respond to voice or text input from virtual assistants or applications. Furthermore, various research is underway on conversational agents focused on personalization. Human-computer conversation systems have largely evolved into two types.

There are two types of dialogue systems: open-domain dialogue systems and task-oriented dialogue systems. Unlike task-oriented dialogue systems that must perform a specific purpose, open-domain dialogue systems must converse with users (chit-chat) naturally, like a human speaking, like a chatbot. [8] explained that existing chit-chat models have problems such as inconsistent responses, unnatural responses, and non-specific responses. In addition, [6] revealed that most chatbot responses are meaningless, predictable, and lack personas.

Accordingly, recent open-domain conversation systems (Chit-chat) have been developed in combination with personas. A representative example is the Persona conversation dataset (PERSONA-CHAT) [8]. A conversation reflecting personas is a system in which personal information such as the age, occupation, family, and hobbies of two speakers are determined. This is a conversation that generates utterances that match the personal characteristics assigned to the item. By equipping computers with persona information, they can have more consistent and natural conversations. In [9], Blenderbot, an open-domain conversation system that combines various conversation skills such as empathy and knowledge, was developed. While previous conversation models focused on increasing the number of parameters, Blenderbot improved model performance by utilizing not only large-scale parameters but also a dataset that can learn human conversation skills well.

2.2 Research on interactive agent-based education

Over the past few decades, there has been significant interest and attention in various technological aids, including AIEd, in the field of language learning. Language learning involves lessons and exercises that enhance vocabulary, grammar, and the four language skills of listening, reading, speaking, and writing. Using educational conversational agents that incorporate these features allows learners to not only grow through language use but also receive immediate feedback on their achievements. [10] reported that chatbots can motivate students to perform better academically. [3] investigated the role of chatbots in language learning and identified six characteristics that can enhance student language learning: first, students feel more comfortable interacting with bots than with people; second, bots allow for unlimited conversation; third, bots enable students to practice listening and reading; fourth, bots are engaging in themselves; and fifth, bots enable students to learn vocabulary and various language expressions. Finally, using a bot can provide effective feedback on learners' grammatical and spelling errors. [11] In the study, ROBOSEM (Educational Service Robot) with TTS function was developed to improve learners' intonation and speaking speed, and the study was conducted to see whether learners' pronunciation improved depending on whether TTS was used. Most users After using ROBOSEM's TTS function, it was found to be effective in improving speaking skills and pronunciation.

[12] conducted an English grammar test before and after using the conversational agent to determine the effect of chatting. The results of the pre- and post-evaluation analysis showed that the improvement in English grammar skills due to chatting with the conversational agent was statistically significant.

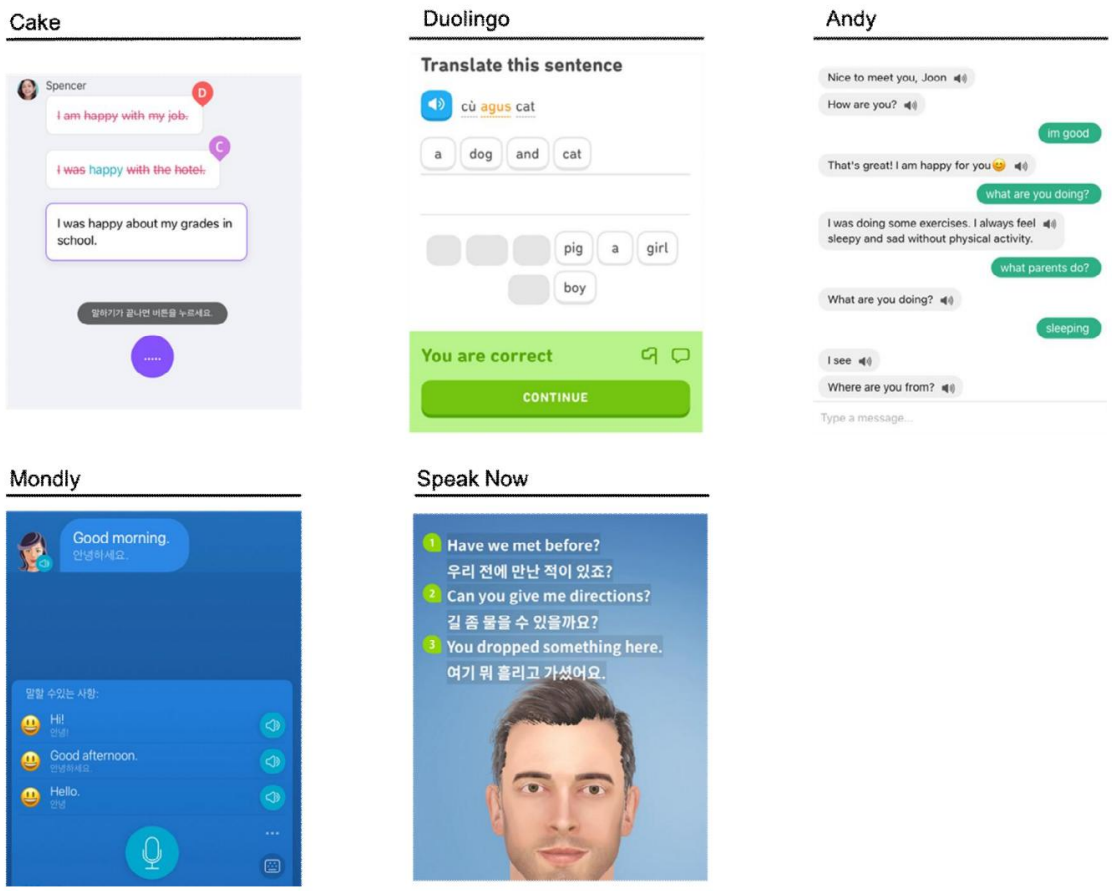


Fig. 1. The design and layout of AIEd platforms using conversational agent

3. Trends in Conversational Agent-Based Educational Platforms

Most existing research on chatbots for English education has focused on AI speakers such as Amazon Alexa or Google Assistant. This can be seen as a move to leverage the information-providing advantages of these general-purpose chatbots while considering their limitations in engaging in conversations tailored to specific situations [13]. While general-purpose chatbots are useful for simple chatter and information retrieval [14], they have limitations when used in English education activities [13]. Therefore, an analysis of chatbots developed for English education purposes is

necessary. This paper conducted a trend analysis by selecting apps among commercially available educational platforms with a large user base and those designed with conversational agents for English learning.

3.1 Cake

Cake is an English learning platform based on video content. Learners learn various expressions used by native speakers in American TV dramas, programs, and movies through videos. There are differences between English learned academically and English expressions actually used by native speakers. In this respect, Cake English education platform allows learners to learn English conversational expressions actually used by native speakers through video content such as TV programs and movies. In addition, STT and

You can use TTS to evaluate whether your pronunciation is similar to that of a native speaker or practice by listening to actual pronunciation.

As shown in Fig. 1, it has the form of a chatbot, but instead of actually having a conversation, it only has the function of evaluating pronunciation by pressing a button given in the app, or displaying word and expression search results.

3.2 Duolingo

Duolingo is one of the most popular language learning chatbot platforms. It supports learning over 30 languages. While chatbot functionality was available for some languages before 2020, this feature has since been removed. Duolingo's most notable features include gamification and translation-based learning. It uses gamification strategies to motivate learners to continue learning [16].

The gamification strategy is as follows: Learning assessment provides a customized level of learning (unit) for the learner, and as the level of each unit increases, more difficult learning can be done. Alternatively, if an incorrect answer is given, the points are lost and the test must be restarted or a specific lesson must be completed. STT and It provides a wide range of English learning, including pronunciation learning through TTS, grammar learning through GEC, and situation-based learning such as daily life, restaurants, and travel, based on speaking, reading, writing, and listening learning through artificial intelligence.

While in-class activities primarily focus on pronunciation, vocabulary, listening, and grammar, most of the learning process is based on translation practice. While Duolingo's translation-based approach may seem overly simplistic compared to other AIEs, its structured learning design and curriculum, which includes pronunciation, vocabulary, and GEC, can have a positive impact on students' acquisition of accurate pronunciation, vocabulary, and basic grammar structures.

3.3 Andy :

Andy is a virtual tutor app that helps learners learn English. This chatbot, based on a conversational agent, teaches new vocabulary, grammar, and provides language learning games. Specifically, learners can learn the language by using English in real-life conversations. For example, they can learn English conversation by exchanging greetings or discussing specific topics. The chatbot can also engage in discussions with learners. Learners can discuss various topics with Andy, such as movies, travel, art, and humor. After several discussions and conversations with the chatbot, they can also learn the language through games.

Language learning includes vocabulary and grammar learning. Vocabulary learning involves a chatbot explaining words and their meanings to the learner, and providing feedback on how well the learner knows the given words. After the definitions and explanations are complete, the learner Learning progresses by evaluating the extent to which a word has been learned using four scales (Very Bad, Explain, Easy, Ok). For grammar learning, the learner selects the grammar they wish to learn, and the chatbot provides an explanation of the grammar, followed by a grammar test. The grammar test is multiple-choice, with the learner selecting one correct answer from multiple options. After completing the training, the results of the word and grammar learning are stored in a vocabulary or database for review or use for learning statistics. However, the Andy chatbot has the following limitations. It cannot store previous conversations with the learner and only responds to the most recent conversation input. Furthermore, if it does not understand the user input, it responds with responses such as "I see" or "Got it" and then questions within a set topic. In other words, it does not generate diverse responses to user input, but only uses a set of expressions to conduct the conversation. This prevents in-depth conversations with the user, which can lead to inefficient language learning. Andy is a purpose-based chatbot that only has a nominal non-purposeful conversational format, but is unable to generate diverse sentences.

3.4 Mondly

Mondly is one of the world's most well-known language learning apps. This app utilizes a chatbot to assist with language learning, making it more effective than traditional English language learning. Like other apps, Mondly offers a learning method that utilizes repetitive vocabulary learning and flashcards. The Mondly chatbot supports 33 languages, allows users to chat with native speakers after learning words and sentences, and supports speech recognition to help with pronunciation. Furthermore, practical language learning is possible through language learning based on everyday scenarios, such as ordering food at a restaurant. For vocabulary learning, the app provides correct answers as feedback when users input incorrect words.

Mondly has the following limitations. As shown in Fig. 1, it is not a free-flowing conversational platform, but rather a platform that only allows for responses to input. Because input is limited, it is difficult for learners to engage in conversations on topics of their choice and to learn diverse expressions.

3.5 Speak Now is an

educational app for learning English conversation using AI avatars. Learners can select an avatar with their preferred style as a tutor. Speak Now offers conversational lessons based on various scenarios, such as expressing joy or talking to the opposite sex. Unlike other English learning platforms, it also offers a free conversation feature. Including this learning method, the conversation input supports voice recognition, giving users the feeling of actually speaking with a native speaker. Like other educational platforms, it also offers a repetitive vocabulary learning feature.

SpeakNow's conversation system includes free conversation and contextual conversation features. Free conversation primarily involves asking for greetings and engaging in conversations on specific topics. For "greetings," if the user inputs a question that deviates from the greeting, the chatbot repeats the greeting until the desired greeting is received. For conversations on specific topics, the chatbot must input a set pattern, such as "Let's talk about ~" or "I would like to speak ~." Otherwise, the chatbot is forced to continue speaking in that pattern. Furthermore, because the available topics are limited, inputting a topic not stored in the knowledge base forces the chatbot to switch to another topic or to speak on a different topic. Furthermore, the overall conversation suffers from long-term memory dependency, preventing it from reflecting previous conversations. Contextual conversation requires selection from the learner's options, allowing only a limited dialogue, as

shown in Figure 1. Consequently, learners cannot respond to desired responses and in-depth conversations are impossible. In other words, both free conversation and contextual conversation features are merely conversational agents in form, preventing effective conversation practice.

4. Experiments and Experimental Results

This paper conducted a human evaluation on the educational platform that enables free conversation (Chit-chat) analyzed above and the latest open-domain conversation model released as open source.

4.1 Dataset

Model responses based on conversation input as data for experiments

For comparison, we used the MTB benchmark dataset proposed in [17]. The MTB dataset consists of 1,477 conversational contexts, each containing 1 to 3 conversational turns. It also includes conversations asking personas, such as “Do you like cats?”, to evaluate the consistency of personas. For the experiments in this section, 30 conversations were randomly selected from the 1,477 conversational contexts and used as the experimental dataset.

4.2. Experimental model

This paper selected Andy, an educational platform capable of conducting conversations, including a conversational agent, and SpeakNow, and Blenderbot 2.0, a large-scale pre-learning conversational model, as experimental subjects for comparison models for the experiment.

Cake and Mondly appear to be conversational agents, but they don't allow learners to input desired dialogue. For this reason, Andy and Speak Now were selected as the subjects of this chapter's experiments. Andy is primarily a chatbot, allowing learners to ask questions about grammar and vocabulary through chat input, or engage in free-flowing conversations. Speak Now offers a separate free-conversation feature for learners to practice conversation.

Blenderbot 2.0, an open-domain conversational system that improves upon Blenderbot 1.0, was released in July 2021 [18,19]. Blenderbot 1.0 was the first chatbot to incorporate various conversational skills, such as empathy and knowledge. Blenderbot 1.0 not only includes large-scale parameters but also utilizes a dataset that is well-suited for learning human conversational skills. Major issues with existing models, such as Blenderbot 1.0 and

GPT-3 [20], included making statements that contradicted previous conversations or generating conversations that did not reflect the latest information. Blenderbot 2.0 performs consistent conversations across multiple sessions and can even search online for up-to-date information for conversations. This represents a significant difference from existing conversational models. Additionally, it is characterized by using the Wizard of the Internet (WizInt) dataset [18] and the Multi-Session Chat (MSC) dataset [19] through crowdsourcing.

The WizInt dataset was used to generate answers to users' Internet search inputs, and the MSC dataset was used to generate new information from multi-session conversations.

This data set was used to store acquired knowledge in long-term memory and generate responses based on it. Using this dataset, it is possible to output conversations that reflect Internet information search results without compromising the given persona, and the concept of sessions enables the generation of naturally continuous conversations even over long intervals.

4.3 Experimental design

Most existing studies on educational chatbots [13-15,21] have not directly compared them with the latest large-scale pre-trained conversational models. This is because evaluating conversational agents is inherently difficult [22], and building evaluation datasets and organizing evaluators for evaluation are challenging. Furthermore, the conversational agents used in commercialized educational platforms are not open source, making direct evaluation difficult. This paper conducted a comparative

experiment using human evaluation of the latest conversational models and the conversational models used in educational platforms. Through this comparative experiment, we analyze the limitations and issues of the conversational models used in educational platforms and propose improvements based on these findings. The latest conversational model, Blenderbot 2.0, and the educational platform conversational models, Andy and SpeakNow, were selected as the subjects of the experiment. Furthermore, for human evaluation, we conducted static and conversational evaluations utilizing the Sensibleness and Specificity Average (SSA) evaluation metric proposed in [17].

4.4 Evaluation Criteria

The SSA evaluation metric is a human evaluation methodology used in Meena[17], a conversational agent announced by Google Brain in January 2020. The SSA evaluation metric is an index to evaluate whether a chatbot can converse like a human. It can evaluate whether an open-domain chatbot can converse like a human without relying on the evaluation methods of existing knowledge bases or rule-based systems. Since effective conversational learning is possible if an educational chatbot provides learners with the feeling of conversing with a real person, this paper
SSA evaluation indicators were used.

The SSA evaluation metric is whether the chatbot's response makes sense (Sensibleness) and is specific to the user's input, i.e., the context. (Specificity) is assessed. Specifically, Sensibleness is It evaluates whether the response makes sense, is logical, and is consistent. However, Sensibleness can consistently give high scores to boring and non-specific answers such as “I don’t know.” To improve this problem, Specificity evaluates whether the answer is specific enough to fit the context when Sensibleness is evaluated as 1. Therefore, both Sensibleness and Specificity scales have values of 0 or 1, but if Sensibleness is evaluated as 0, Specificity is set to 0.

SSA evaluation methods are categorized into static evaluation and interactive evaluation. Static evaluation is a methodology that evaluates responses to a common conversation using SSA evaluation metrics to compare multiple conversation models. This paper used the Mini-Turing Benchmark (MTB) [17] dataset, which was created for static evaluation, as an evaluation dataset and conducted experiments with three evaluators. While static evaluation can be suitable for comparing multiple models using the same dataset, it can be biased depending on how the static evaluation dataset is structured. Therefore, interactive evaluation should also be conducted. Interactive evaluation is a methodology in which evaluators conduct SSA evaluations while interacting with actual conversation models, rather than using a predefined dialogue input. This paper requested at least six conversations with each chatbot system (three user utterances and three chatbot system sessions), and ultimately conducted experiments on ten conversation sessions.

4.5 Experimental Results

4.5.1 Quantitative Analysis

In this paper, we conducted an experiment by randomly selecting 30 MTB benchmark datasets proposed in the SSA[17] evaluation index methodology for static evaluation.

The experimental results of the static evaluation are shown in Fig. 2. Blenderbot shows high scores in all evaluation criteria. Blenderbot was followed by Andy and SpeakNow in that order. In the case of SpeakNow, we can see that the Specificity score for the MTB dataset is close to 0. The experimental results of the interactive evaluation are shown in Fig. 3. Similar to the static evaluation, Blenderbot showed high scores in most evaluations. Only SpeakNow showed a slightly higher score in the Sensibleness evaluation. SSA evaluation index

BlenderBot, SpeakNow, and Andy were evaluated in that order. Compared to the static evaluation, both Andy and SpeakNow showed improved scores.

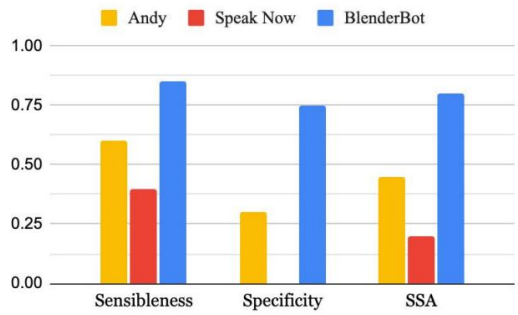


Fig. 2. Results of static human evaluation

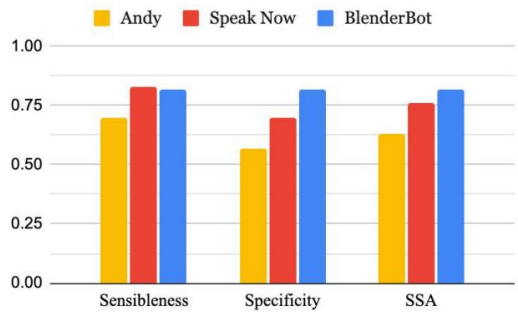


Fig. 3. Results of interactive human evaluation

4.5.2 Qualitative Analysis

This paper conducted a qualitative error analysis based on experimental results to analyze the problems and limitations of the English education platform. The results of the qualitative error analysis are presented in Table 1.
It is as shown in Fig. 4.

Table 1 is an example of qualitative error analysis for static evaluation. BlenderBot answered appropriately for the given context, and the answer was specific and clear. In the case of Andy, the answer “Interesting” may be appropriate for the context, but it is not specific or clear. In the case of SpeakNow, the given answer is not appropriate or specific for the context. BlenderBot answered appropriately and specifically for the given context. The problems of Andy and SpeakNow, as examined through qualitative error analysis of static evaluation, are as follows. When inputting a conversation outside the knowledge base, it generates an answer that is not appropriate for the context, or is appropriate for the context but is not specific. It generates a single answer. In particular, SpeakNow was designed to accept only a limited number of input topics, so it showed a very low score on the MTB dataset containing a wide range of topics, as shown in Fig. 2. However, Blenderbot generated context-appropriate and specific answers for most conversation inputs.

Table 1. Example of generated response given the context in the MTB benchmark dataset.

Context:	Have you ever tried a mechanical keyboard?
Andy:	Interesting
Speak Now:	Thanks. Have a nice day.
BlenderBot:	I have, but I'm not very good at it. It's hard for me to press the keys correctly.

Fig. 4 is an example of qualitative error analysis for interactive evaluation. In the case of Andy, most of the answers are appropriate to the user input, but they are not specific. In Fig. 4, we can see non-specific answers such as “Nice.” or “Oh, it’s good to hear that.” For effective conversational learning, it is necessary to induce the learner’s response by including questions or providing specific answers like Blenderbot. In addition, Andy has difficulty in deeply understanding the learner’s input. Although Andy provides an English word search function, it recognized “Is” instead of “Good” in response to the question “Do you know the meaning of ‘Good’?”, showing difficulty in word search in a conversational format.

In the case of SpeakNow, there is a problem that it is vulnerable to questions that are outside the knowledge base and limits the learner's input.
Looking at Fig. 4, most of the responses are context-appropriate and specific. However, a closer look reveals that all chatbot responses include questions. This encourages the user to engage in conversations on topics the chatbot can answer. In other words, rather than generating a response for any input, it outputs a pre-stored response. Looking at SpeakNow in Fig. 4, if a question outside the knowledge base, such as “No, Do you know Leonardo DiCaprio”, is input to the question “Do you like your boss?”, it simply judges it as a negative sentence and outputs a pre-set response. By limiting the conversation in this way, it was able to receive high scores in conversational evaluations, but it showed low scores in static evaluations that included various inputs. In addition, the word search function

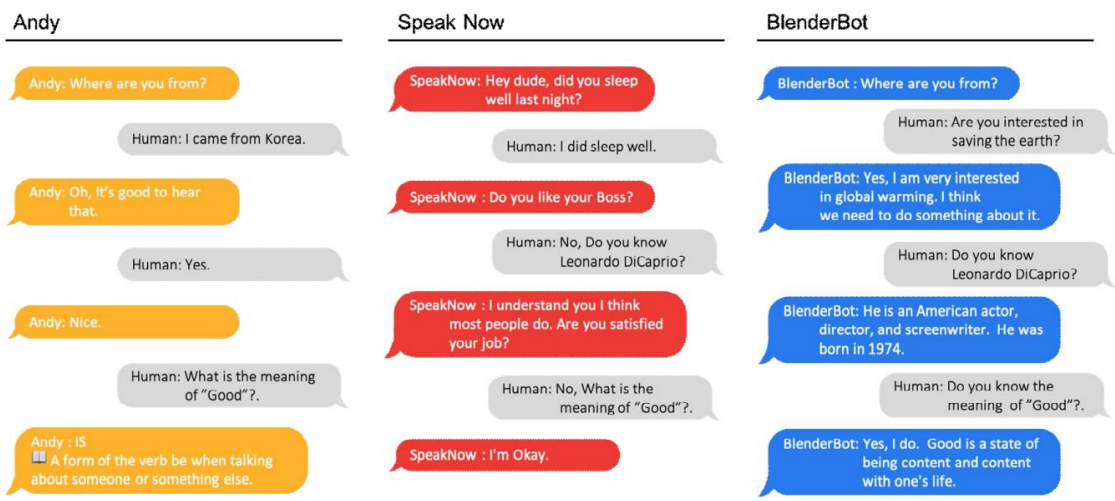


Fig. 4. Interactive examples multi-turn dialogue generation. Conversational agent are Andy, Speak Now, and Blenderbot

We can see that it does not support and cannot properly answer the question "No, what is the meaning of 'Good'?"

In contrast to Andy and SpeakNow, Blenderbot excels at recognizing learner input and provides specific responses. Furthermore, it generates answers to unknown words or information through Internet searches. As shown in Fig. 4, Blenderbot is robust to a variety of inputs, including generating answers to questions that deviate from the topic and are not included in its knowledge base, such as "Do you know Leonardo DiCaprio?" when asked "Where are you from?". Furthermore, it can perform word search. In contrast to Andy, Blenderbot provides responses such as "Yes, I do." All of Blenderbot's responses are context-appropriate and specific. This stimulates interest and motivation for English conversation learners, enabling effective learning.

5. Suggested improvement measures

Regarding the aforementioned problems and limitations, we propose improvement measures through comparison with Blenderbot, the latest large-scale conversation model, as follows.

The problems and limitations of educational platforms, including Andy and SpeakNow, are summarized as follows: First, when input is outside the knowledge base, it may be appropriate for the context, but Second, it generates vague answers. Second, it cannot generate diverse answers and only outputs fixed answers. Due to these limitations, it requires only limited input from learners, like Mondly, or continuously asks questions only for answerable questions, like SpeakNow. This limited conversational method prevents learners from having a broader conversation, which can reduce interest and motivation compared to relatively free conversation, thereby reducing learning efficiency. Therefore, this paper proposes the following improvements for educational platforms utilizing conversational agents.

5.1 Large-scale pre-training model

To achieve a deep understanding of learner input and generate appropriate and specific responses, the use of conversational models trained with large parameters is crucial. As examined in the qualitative analysis, existing commercialized educational platforms lack the ability to recognize input, which limits the topic of conversation and prevents proper utilization of chatbot functions such as word search. Blenderbot is a model trained with larger parameters than existing chatbot models. When human evaluations were conducted with existing chatbots, Blenderbot's conversations were considered the most engaging and responded that it generates more human-like conversations [23]. Compared to other educational platforms, it can provide deeper understanding and generate specific responses. Therefore, the conversational models of currently commercialized educational platforms are models trained

, which should be improved. Through concrete and more human conversations based on deep understanding, learners can become more immersed in the conversation, stimulate interest, and motivate themselves. This, in turn, leads to more conversations, increasing learning effectiveness.

5.2. Building training data

Because there has been no conversational dataset specifically designed for education, implementing chatbots for educational purposes has been challenging. Furthermore, previous research on chatbots for English education has primarily focused on AI speakers [13-15]. Because AI speakers are not designed for educational purposes, their use in education presents challenges. However, if educational-specific conversational data is built and a conversational model is developed using this data, language learning tailored to education becomes possible.

An example of an educational conversation dataset is a dataset comprised of situational conversations. Situational English conversation learning refers to learning basic English conversation and English relevant to various situations that can arise in daily life. For example, learning progresses through conversation practice in pre-determined, real-life situations, such as exchanging money at the airport or asking about symptoms at a hospital. Speak Now's scenario-based conversation feature exemplifies this approach. However, unlike Speak Now's free conversation feature, it does not allow free conversation and requires users to select from a set of given responses. This paper proposes a training dataset that enables free conversation in specific situations, rather than limited conversations like Speak Now. By constructing a multi-turn conversation dataset for various situations and training it with a conversation model trained with a large number of parameters, a conversation model capable of situational conversation learning can be constructed.

5.3. Information search function

The biggest difference between BlenderBot 2.0 and existing conversational models is that it searches the Internet for information and enables conversations about the latest information. As shown in Fig. 4, answers are generated through Internet searches for information and words outside the knowledge base. By utilizing this, learners can ask questions about words they are curious about directly in the conversation without using word searches, creating a more natural and effective learning experience. Additionally, providing learners with access to content-related lecture videos, articles, newspapers, and books through search functions can enable conversations on a wider range of topics, which can motivate and stimulate interest.

6. Conclusion

Artificial Intelligence-Enabled Education (AIEd) combines various AI technologies to deliver more effective learning than traditional learning. Among these, conversational agents have garnered significant attention due to their ability to learn freely, free from time and space constraints, and their ability to be combined with various technologies such as STT, TTS, and GEC. However, current educational platforms utilizing conversational agents are limited in form. They can only converse on a limited range of topics, and their responses are either predetermined or indefinite. These issues not only hinder effective learning for learners, but can also lead to learning decline. This paper conducted a SSA human evaluation experiment on Blenderbot 2.0, the most recently released large-scale pre-trained conversational model, and an educational platform utilizing a conversational agent. This study aims to identify these issues and limitations through comparative analysis and propose improvements. In the static evaluation, the SSA scores were Blenderbot, Andy, and SpeakNow, in that order. In the conversational evaluation, Blenderbot, SpeakNow, and Andy were ranked first. A qualitative analysis was then conducted, and the results are as follows. The conversational agent in the education platform struggled to engage in conversations outside of its knowledge base, leading to limited conversations. Furthermore, there was a problem with its answers being vague. However, Blenderbot was able to generate answers through internet searches for specific answers and information not in its knowledge base. Finally, based on the analyzed problems and limitations, we proposed improvement measures for the effective use of conversational agents in education platforms. These improvements included the need for a conversation model trained with large parameters, the construction of educational conversation data, and information retrieval capabilities. In the future, we plan to conduct follow-up research based on this paper to develop a conversation model that addresses some of the shortcomings of educational conversational agents and to build a dataset specifically for educational conversations.

REFERENCES

[1] Z. Ruttkay & C. Pelachaud. (Eds.). (2006). From brows to trust: Evaluating embodied conversational agents, 7, Springer Science & Business Media.

[2] M. Lim. (2020). A Study on the Direction of Technology Education in the Age of Artificial Intelligence. Journal of Korean Practical Arts Education, 33(4), 81-102.

[3] L. Fryer & R. Carpenter. (2006). Bots as language learning tools. Language Learning & Technology, 10(3), 8-14.

[4] N. Y. Kim, Y. Cha & H. S. Kim. (2019). Future Eng lish learning: Chatbots and artificial intelligence. Multimedia-Assisted Language Learning, 22(3), 32-53.

[5] J. Li & D. Jurafsky. (2016). Neural net models for open-domain coherence. discourse arXiv preprint arXiv:1606.01545.

[6] B. Chantarotwong. (2006). The learning chatbot. Final year project.[Online]: http://courses.ischool.berkeley. edu/i256/f06/pro jects/bonniejc. pdf.

[7] D. Lee. (2018). A study for the development of an English learning chatbot system based on Artificial Intelligence. 45-68. Secondary English Education, 11(1),

[8] S. Zhang. et al. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?. arXiv preprint arXiv:1801.07243

[9] S. Roller. et al. (2020). Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637.

[10] J. Jia & M. Ruan. (2008, June). Use chatbot csiec to facilitate the individual learning in english instruction: A case study. In International conference on intelligent tutoring systems (pp. 706-708). Springer, Berlin, Heidelberg.

[11] J. In & J. Han. (2016). The prosodic changes of Korean English learners in robot assisted learning. Journal of The Korean Association of Information Education, 20(4), 323-332.

[12] N. Y. Kim. (2019). A study on the use of artificial intelligence chatbots for improving English grammar skills. Journal of Digital Convergence, 17(8), 37-46.

[13] Y. Kim. (2020). Analysis of chatbots and chatbot builders for English language learning. Multimedia-Assisted Language Learning, 23(4), 161-182.

[14] J. E. Hyun & H. J. Im. (2019). Analysis and Implications of AI Speakers as English Learning Tools. The Journal of Mirae English Language and Literature 24(1), 201-219.

[15] H. Kim & D. Shin. (2019). A study of AI chatbot as an assistant tool for school English curriculum. The Journal of and Learner-Centered Curriculum Instruction, 19(1), 89-110. DOI; 10.22251/jlcci.2019.19.1.89

[16] K. Teske. (2017). Duolingo. 393-402.Calico Journal, 34 (3),

[17] J. Jang & S. Ye. (2021). Towards Continual Knowledge Learning of Language Models. arXiv preprint arXiv:2110.03215

[18] M. Komeili, K. Shuster & J. Weston. (2021). Internet-augmented dialogue generation. arXiv preprint arXiv:2107.07566

[19] J. Xu, A. Szlam & J. Weston. (2021). Beyond goldfish memory: Long-term open-domain conversation. arXiv preprint arXiv:2107.07567

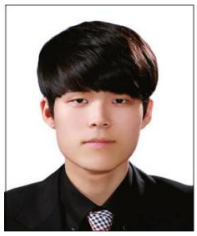
[20] L. Floridi & M. Chiriatti. (2020). GPT-3: Its nature, scope, limits, and consequences. 681-694. Minds and Machines, 30(4),

[21] T. Yoon & S. Lee. (2021). Effects of Primary ELLs' Affective Factors and Satisfaction through AI-based Speaking Activity. The Journal of the Korea Contents Association, 21(9) 34–41. DOI: 10.5392/JKCA.2021.21.09.034

[22] J. Sedoc. et al. (2019, June). ChatEval: A tool for chatbot evaluation. In Proceedings of the 2019 Conference of the North American Chapter of the for Association Computational Linguistics (Demonstrations) (pp. 60-65).

[23] S. Roller. et al. (2020). Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637

SeungJun Lee [Student Member]



February 2021: Hankuk University of Foreign Studies, Department of Industrial and Management Engineering (B.S.) July 2021 - Present: Human-inspired AI Research Institute

ÿ ÿÿÿÿ : Natural Language Processing, AI in Educa tion, Text-Mining ÿ E-Mail : dzzy6505@gmail.com

Chanjun Park [Student Member]

February 2019: Busan University of Foreign Studies, Language Processing Creative Convergence Major (Bachelor of Engineering) June 2018 ~ July 2019: SYS TRAN Research Engineer ÿ September 2019 ~ Present: Integrated Master's and Doctoral Program in Computer Science, Korea University

ÿ ÿÿÿÿ : Data-Centric AI, Machine Translation, Grammar Error Correction, Deep Learning

ÿ E-Mail : bcj1210@naver.com

Jaehyung Seo [Student Member] ÿ August 2020: Korea University, Department of English

Language and Literature and Department of Business Administration (B.A., B.A.) ÿ September 2020 ~ Present: Korea University, Department of Computer Science, Integrated Master's and Doctoral Program ÿ Areas of Interest: Graph Encoder, Computer Reasoning ÿ Email: seojae777@korea.ac.kr

Heuiseok Lim [Life Member]

ÿ 1992: Korea University, Department of Computer Science (Bachelor of Science) ÿ 1994: Korea University, Department of Computer Science (Master of Science) ÿ 1997: Korea University, Department of Computer Science (Doctor of Science)

ÿ 2008 ~ Present: Professor, Department of Computer Science, Korea University ÿ Areas of Interest: Natural Language Processing, Machine Learning, Artificial Intelligence ÿ E-Mail: limhseok@korea.ac.kr