education sciences

MDPI

Article

Developing a Task-Based Dialogue System for English Language Learning

Kuo-Chen Li [1,*], Maiga Chang [2,3] and Kuan-Hsing Wu [1]

[1] Information Management Department, Chung Yuan Christian University, Taoyuan City 320314, Taiwan; xeaiow@gmail.com
[2] School of Computing and Information Systems, Athabasca University, Athabasca, AB T9S 3A3, Canada; maiga.chang@gmail.com
[3] Department of M-Commerce and Multimedia Applications, Asia University, Taichung City 41354, Taiwan
* Correspondence: kuochen@cycu.edu.tw

Received: 29 July 2020; Accepted: 26 October 2020; Published: 28 October 2020

Abstract: This research involved the design of a task-based dialogue system and evaluation of its learning effectiveness. Dialogue training still heavily depends on human communication with instant feedback or correction. However, it is not possible to provide a personal tutor for every English learner. With the rapid development of information technology, digitized learning and voice communication is a possible solution. The goal of this research was to develop an innovative model to refine the task-based dialogue system, including natural language understanding, disassembly intention, and dialogue state tracking. To enable the dialogue system to find the corresponding sentence accurately, the dialogue system was designed with machine learning algorithms to allow users to communicate in a task-based fashion. Past research has pointed out that computer-assisted instruction has achieved remarkable results in language reading, writing, and listening. Therefore, the direction of the discussion is to use the task-oriented dialogue system as a speaking teaching assistant. To train the speaking ability, the proposed system provides a simulation environment with goal-oriented characteristics, allowing learners to continuously improve their language fluency in terms of speaking ability by simulating conversational situational exercises. To evaluate the possibility of replacing the traditional English speaking practice with the proposed system, a small English speaking class experiment was carried out to validate the effectiveness of the proposed system. Data of 28 students with three assigned tasks were collected and analyzed. The promising results of the collected students' feedback confirm the positive perceptions toward the system regarding user interface, learning style, and the system's effectiveness.

Keywords: task-based dialogue; computer-assisted instruction; task-based language learning

1. Introduction

With the rise of Internet technology, many educational institutions are gradually turning their attention to the application of digital education. Various online learning methods enable people to make good use of their spare time to learn, greatly enhancing traditional learning efficiency. Technologies integrating various learning approaches such as pragmatic, context, or cooperated learning have shown great success in language learning [1–3]. In the context of digital language learning, speaking is considered to be one of the most important parts of learning a foreign language [4]. To address this problem, social interaction is a key factor in improving language fluency in language learning. However, the cost of creating a social language-learning environment is too high to be widely implemented [5,6]. Researchers seek opportunities to adopt computer-aided technologies to create an environment similar to speaking with native English speakers. With the

popularity of computer-assisted language learning (CALL) and the advancement of the Internet, new methods in the field of language learning are booming [7–9].

Dialogue practice has become increasingly important in computer-aided language (CAI) learning especially for foreign languages [10]. As hardware and natural language processing progress with the times, dialogue training can be accomplished through computer technologies [11–13]. Language learning places special emphasis on the training of communication skills. This study adopted task-based language learning as the fundamental curriculum design and used a task-based dialogue system to provide the interface. Dialogue practice thus can be carried out through task-based learning, and at the same time, the learning process can be conducted by the computer-aided dialogue system driven by the language learning curriculum.

Educators have promoted task-based learning for language learning [14,15]. The original concept of task-based learning was developed by Prabhu [16], who proposed a task-based leaning model to enable learners to learn a language in the process of solving a "non-linguistic task" and focus on learners in the process of performing tasks. The application of cognitive techniques such as received language messages and processing is proven to be as effective as traditional teaching methods [17]. Prabhu [16] pointed out three different kinds of activities for task-based learning:

1. Information-gap activity: Allow learners to exchange information to fill up the information gap. Learners can communicate with each other using the target language to ask questions and solve problems.
2. Opinion-gap activity: Learners express their feelings, ideas, and personal preferences to complete the task. In addition to interacting with each other, teachers can add personal tasks to the theme to stimulate the learners' potential.
3. Reasoning-gap activity: Students conclude new information through reasoning by using the existing information, for example, deriving the meaning of the dialogue topic or the implied association in the sentence from the dialogue process.

Willis outlined the teaching process of task-based language teaching as three stages: pre-task stage, task cycle stage, and language focus stage [18,19]. Stage activities can be used to construct a complete language learning process. The pre-task stage pre-approves the learner's task instructions and provide the student with clear instructions on what must be done during the task phase [19]. This helps students review the language skills associated with the task. Through the execution of task activities, the teacher can judge the students' learning status on the topic. At the task cycle stage, students use the words and grammar they learned during the task preparation phase and think about how to complete the tasks and presentations. In this process, the teacher plays the role of supervisor, giving appropriate guidance and language-related resources. In the last stage, the language focus stage, students and teachers review related issues encountered during the previous phase, such as the use of words, grammar, or sentence structure. The teacher guides the students to practice the analyzed results and improve their language comprehension.

The efficiency and crucial factors of task-based language learning have been surveyed by different aspects of studies. Research shows a significant improvement of speaking comprehension [20–22]. Rabbanifar and Mall-Amiri indicate that the reasoning-gap activity holds the key factor for speaking complexity and accuracy [21].

The present study adopted the three-stage-model shown in Figure 1 to develop the task-based dialogue system [16]. In the pre-task stage, the system needs to present the task and let students clearly understand the goals to accomplish throughout the conversation. In the task cycle, the system needs to interact with students and guide students to finish the task. For the language focus stage, the system needs to be able to evaluate the performance of the students and give the proper feedback.

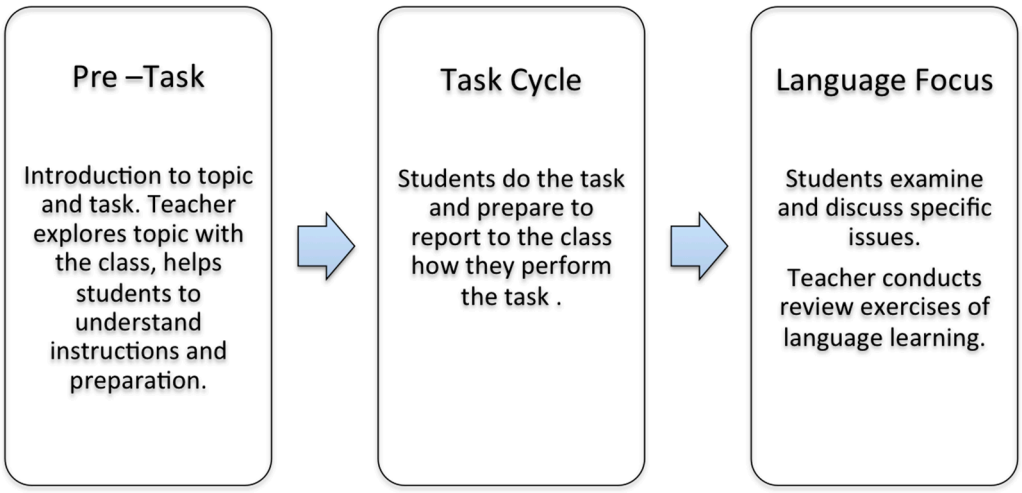

Figure 1. Task-based language learning stages.

The task-based dialogue system usually has a very clear task, such as helping users order meals or learning languages [23]. This dialogue robot contains basic modules including Dialogue Script, Dialogue Manager, Natural Language Understanding, and Natural Language Generation. As shown in Figure 2 [24], the widely used method of the task-based dialogue system is to treat the dialogue response as a pipeline. The system must first understand the information conveyed by humans and identify it as an internal system. According to the state of the conversation, the system generates the corresponding reply behavior and finally converts these actions into the expression of natural language. Although this language understanding is usually handled by statistical models, most of the established dialogue systems still use manual features or manually defined rules for identifying state and action representations, semantic detection, and problem filling [24].

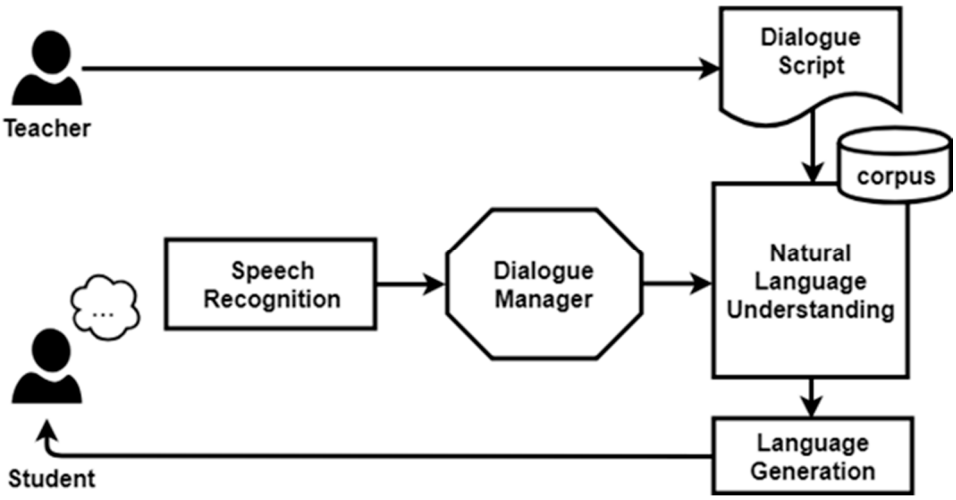

Figure 2. Task-based dialogue system.

Implementing a dialogue system for language learning has been carried out by using different algorithms years ago [25–27]. From the statistic model to pattern recognition, the applications have become more practical and widely developed with the advancement of text mining and natural language processing technologies [25]. Several advantages have been addressed using dialogue system for language learning. The language-learning dialogue system is considered fun and easy to approach for students [25,26]. In addition, the dialogue system is easily integrated with teaching methods such as grammar check and repetition [25]. Except when carrying out the task, the proposed dialogue

system needs to focus more on language learning. Functions regarding speaking comprehension need to be considered and developed in the system.

In recent years, hardware and software technologies have grown rapidly. The media attention toward artificial intelligence and machine learning continues to rise. The development of these technologies makes it possible for applications using machine learning and human–computer interaction to process large amounts of data storage and massive calculations. Many researchers have turned to applications with natural language processing [28–30]. Natural language processing is the communication channel between human and machine. It is also one of the most difficult problems in computer science, whether it is to achieve natural language understanding or natural language interaction. However, applications of natural language processing have been proposed in different fields, such as machine translation, dialogue robots, data retrieval, and abstract generation. Among those applications, the task-oriented robot shows the capability of solving the special purpose problems. For example, food-ordering robots in restaurants or customer service robots are general applications using a task-oriented dialogue robot. In education, computer-assisted teaching robots can help learners' oral fluency and build self-confidence for speaking foreign languages.

The decision-making technology of the dialogue system (chatbot) has gradually matured, an example being the Siri artificial intelligence assistant software in Apple's iOS system [31,32]. Through natural language processing technology, people can use dialogue to smoothly interact with mobile devices, such as querying weather, making phone calls, and setting up to-do items [33–35]. The use of the dialogue system is quite extensive. In terms of the fast-growing chat bots in recent years, in order to allow customers to get instant response from enterprises, many companies have invested resources into building dedicated dialogue robots to save labor costs [34,35]. The chat bot is based on the dialogue system, so it is necessary to simulate human dialogue. In addition, the dialogue has to have meaningful purpose. It still remains a challenge for today's chat bots to understand all kinds of questions and responses correctly, since human languages are ambiguous to a degree. Dialogue training still heavily depends on human communication with instant feedback or correction [32,36]. However, it is not possible to provide a personal tutor for every English learner.

Therefore, this study involved the development of a task-based dialogue system that combines task-based language teaching with a dialogue robot. The proposed task-based dialogue system contains functions to carry out the conversational task including natural language understanding, disassembly intention, and dialogue state tracking. The research objectives were as follows:

1. Development of a task-based dialogue system that is able to conduct a task-oriented conversation and evaluate students' performance after the conversation;
2. Comparison of the differences between the proposed system and the traditional methods;
3. Evaluation of the effectiveness of the proposed system.

The first step of this study was to survey the related studies on task-based learning methodology and task-based dialogue systems to establish the fundamental curriculum design and interfaces of the system. Section 2 proposes a novel framework of a task-based dialogue-learning model. Section 3 elaborates on the experiment and the results. Finally, Section 4 concludes the results and discusses limitations and future works.

2. Methodology

2.1. Proposed Task-Based Dialogue-Learning Model

This study involved the development of a dialogue system that combines task-based teaching methods to assist teachers in guiding students to complete dialogue tasks with the dialogue robot. A complete set of dialogue scripts used by teachers was constructed. Scoring criteria for the grammar, sentences, and speaking were then established similar to those used by a regular English teacher. To validate the performance of the proposed model, an experimental evaluation was designed to explore the learning style and the learning status compared to traditional teaching methods.

The dialogue system is composed of multiple modules as shown in Figure 3. In a task-based dialogue system, the dialogue is retrieved from the automatic speech recognition (ASR), and the information is recorded by the dialogue manager. The information is forwarded to the natural language understanding module to process and understand the semantics expressed by the learner in the conversation. The extracted result is converted into a semantic vector and compared with the pre-constructed dialogue script set. The statement decision module outputs the corresponding response based on the dialogue policy. Finally, the natural language generation module converts the semantic vector to corresponding dialogue. The dialogue can be delivered by a text-to-speech (TTS) module. Thus, multi-turn dialogue can be implemented so that the system can continuously correct or guide the learner back to the scope of the script collection. The system is also equipped with an exception handling function for instances where the conversation is not clear or falling off track. The task-based dialogue system not only needs to use natural language processing to understand sentences but also needs to give a reasonable response according to the current state like a real person.

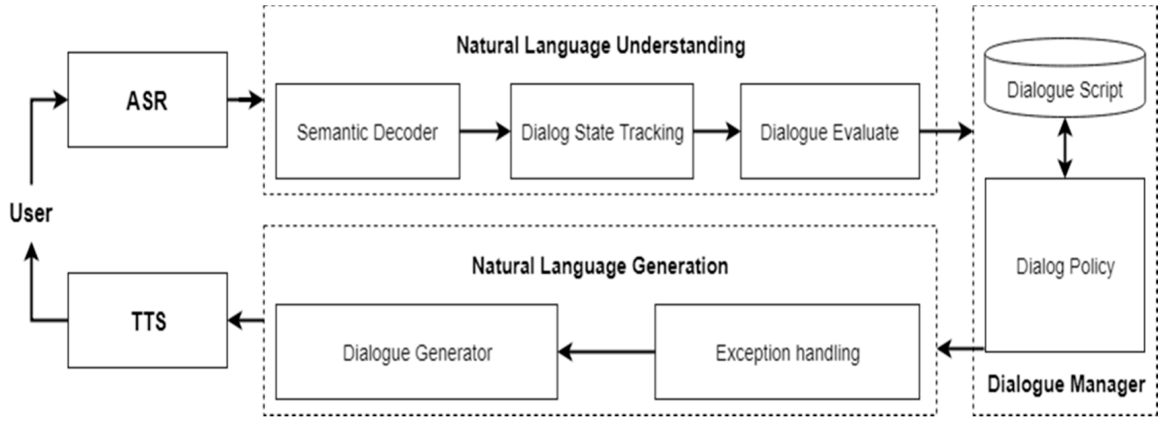

Figure 3. Proposed system model.

Before the learner uses the system, the teacher conducts course training on the topic of the conversation and designs a tree-like dialogue script set in the system in advance. Since the task always involves decision-making, series of decisions for a particular task usually can be represented by a decision tree. Based on the decision tree, the dialogue script is also in a hierarchical form. The dialogue covers various topics and specific tasks such as ordering food or buying tickets. All the scripts were designed by professional English teachers. These conversational themes include the basic elements of language such as grammar statements, language skills, and culture integration. Each dialogue topic has a sequence of tasks to complete, which can be represented by a complete dialogue structure.

Figure 4 shows an example of dialogue branches. In this example, there are three conversation rounds (N1, N2, N3) for one dialogue task, and each layer has one (inclusive) or more possible response sentences. The system determines the dialogue path based on the learner's answer. As shown in Figure 4, the dialogue presents the process of ordering food, whereas N represents non-player character and S represents student. Initially, the system starts the Q&A with the N1 layer. The system presents the learner with three possible responses in the S1 layer based on the dialogue script defined by the teacher. At this time, the learner interacts with the dialogue system to complete the task-based dialogue process by using the dialogue-topic-related information learned during the task preparation phase. This study was designed to enable learners to successfully complete conversation tasks and to guide learners to stay with the pre-defined script.

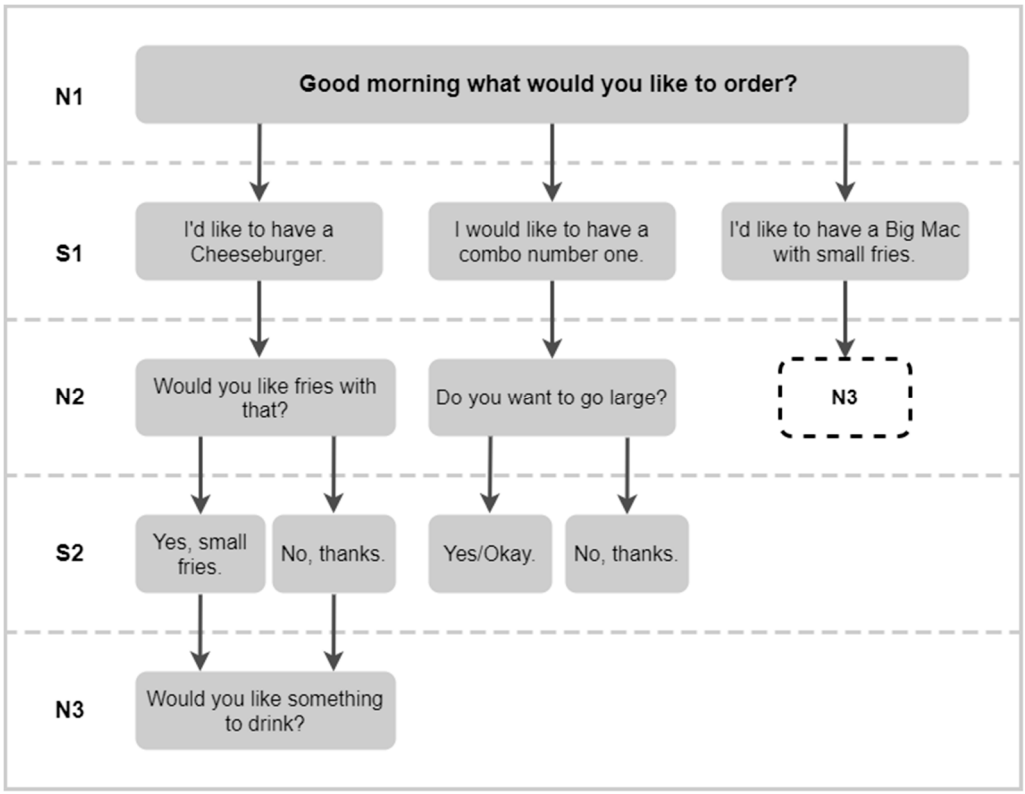

Figure 4. Example of dialogue branches.

Note that the N3 block in the N2 layer is designed to be the continuation of the third answer selected by the learner at the S1 layer; that is, the conversation jumps to the content of the N3 layer so that the task-based dialogue can be completed. The system can flexibly convert or jump back to a conversation. When the learner is led away from the topic, the system can moderately guide the conversation back to the topic. The learner can repeatedly practice and successfully complete the dialogue task and improve their English speaking ability.

In order to train the dialogue robot for natural language understanding, the Wikipedia Corpus was used in this study [37]. Figure 5 shows the word2vector model, which represents the semantic meanings of the sentences and words based on the given Wikipedia Corpus data. A total of 14,000 articles were inputted and used to train the model. Table 1 shows the similarity test for two sentences based on the trained model. Cosine similarity is commonly used to measure the similarity of sentences or texts. Let $s1$ and $s2$ be two vectors to compare the similarity; the cosine similarity can be measured by Formula (1). For the sentences with similarity score 0.8 or above, the trained model is able to obtain the correct semantic meaning.

$$Sim(s1, s2) = \frac{s1 \cdot s2}{||s1||||s2||} \quad (1)$$

The proposed model adopts a sequence-to-sequence recurrent neural network as the dialogue generator. The sequence-to-sequence model consists of two recursive neural networks, the encoder and decoder, to simulate human thinking. When the machine receives a natural language sentence, it gives a corresponding reply according to what it understands. The encoder is responsible for digesting the input sequence and converting it into vector information containing the contents of the original sequence. The decoder generates text based on the converted vector so that it can process input and output sequences of variable length, such as inputting a question and outputting a reply. The designed dialogue robot thus can interact with the leaners based on the pre-defined scripts.

```
import logging
import os
import sys
import multiprocessing

from gensim.models import Word2Vec
from gensim.models.word2vec import LineSentence

if __name__ == '__main__':
    program = os.path.basename(sys.argv[0])
    logger = logging.getLogger(program)

    logging.basicConfig(format='%(asctime)s: %(levelname)s: %(message)s')
    logging.root.setLevel(level=logging.INFO)
    logger.info("running %s" % ' '.join(sys.argv))

    # check and process input arguments
    if len(sys.argv) < 4:
        print(globals()['__doc__'] % locals())
        sys.exit(1)
    inp, outp1, outp2 = sys.argv[1:4]

    model = Word2Vec(LineSentence(inp), size=400, window=5, min_count=5,
                     workers=multiprocessing.cpu_count())
```

Figure 5. Word2vector model for Wikipedia Corpus.

Table 1. Similarity test for the trained model.

Target Sentence	Input Sentence	Similarity
I'd like to have a cheeseburger.	Hello can I have a hamburger please.	0.915
One ticket to Taipei, please.	I want to buy a ticket.	0.831
Yes, small fries.	Yes, I would like some fries.	0.840
Hi, I'd like to make a reservation.	How are you today.	0.760

The dialogue scoring system designed in this study records the sentences expressed by the learners. The system analyzes and evaluates the content of each statement and analyzes the learner's speaking comprehension. The results are presented to the teacher as a language enhancement phase in the three-stage approach proposed by Willis (1996) to enhance the learner's language ability. When the learner selects the topic of the conversation, the dialogue scoring system acts as a teacher to evaluate the learner's conversation. The system grades learners' conversations in similar scoring mechanisms to professional English teachers such as timing, grammar, and correct responses.

2.2. Experiment Procedure and System

The experiment was conducted in a college-level English speaking class. Twenty-eight beginner-level students participated in this experiment with three different tasks after three weeks of traditional English teaching classes. Table 2 shows the given three tasks for the experiment. The teacher first designed three dialogue tasks based on the textbook and labeled them with the proper level to reflect the dialogue difficulty. The details of the three dialogue tasks are listed in Appendix A. According to three-stage task-based language learning, the teacher first explained the task and let students understand what needed to be done during the conversation. In the task cycle, students then entered the system and talked to the dialogue system to complete the task. During the process, the system interacted with the student and recorded the behaviors of the students including pause time, answer time, number of errors, the number of repetitions, and the number of hints (reminders). In the language focus stage, the system evaluated students' performance and gave the feedback to students.

Table 2. Designed tasks.

Theme	Task	Task Description	Level
Accommodation	Hotel Reservation	Reserve a room for two adults, two nights from 8th August.	2
Transportation	Buying Train Tickets	Buy a train ticket to Taipei with a 500 bill.	1
Restaurant	Ordering food in a Restaurant	Order a burger and a coke.	1

Figures 6 and 7 show the student interfaces of the proposed system. Student can receive the task cards given by the teacher and current progress of each task as shown in Figure 6. Once the student completes the task, the system gives the score right away and allows students to trace back the records of the corresponding task as shown in Figure 7. The student is able to replay or redo the task to improve the score of the task.

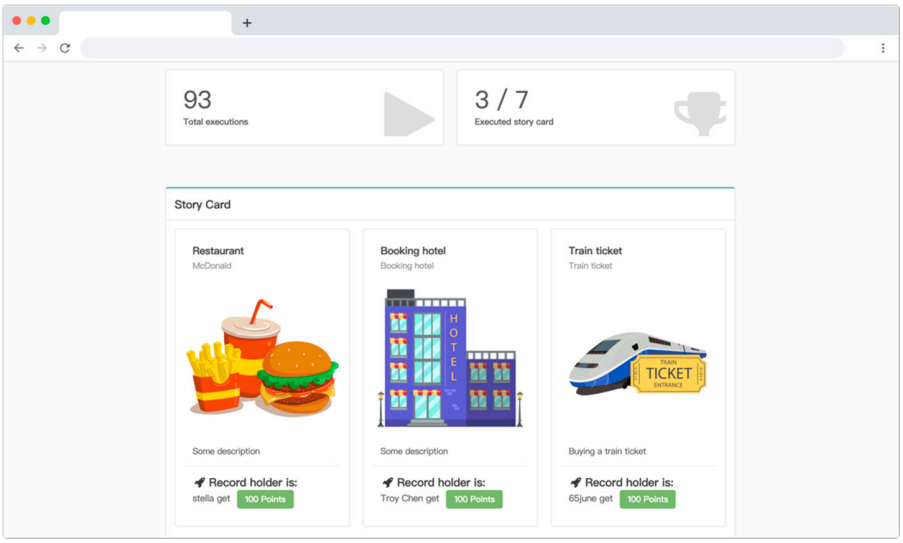

Figure 6. System demonstration: task cards.

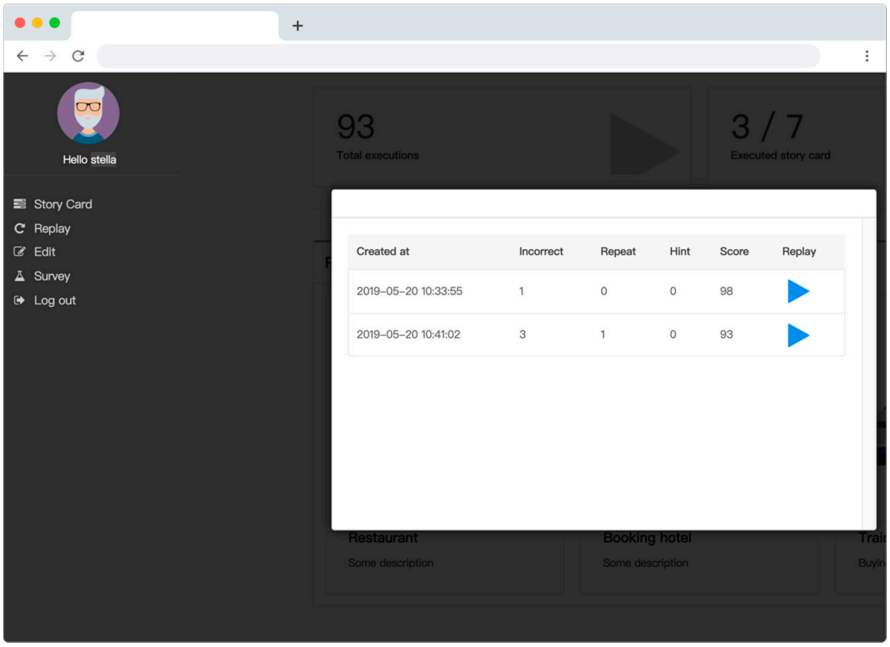

Figure 7. System demonstration: system records.

Figure 8 shows the teacher interface for editing the conversational tree. The tree structure conveys the possible paths for the assigned task. Based on the tree, the system can determine and guide the conversation accordingly based on students' dialogue. When students conduct the task, the system monitors students and helps them to complete the task by providing hints or repeating the question.

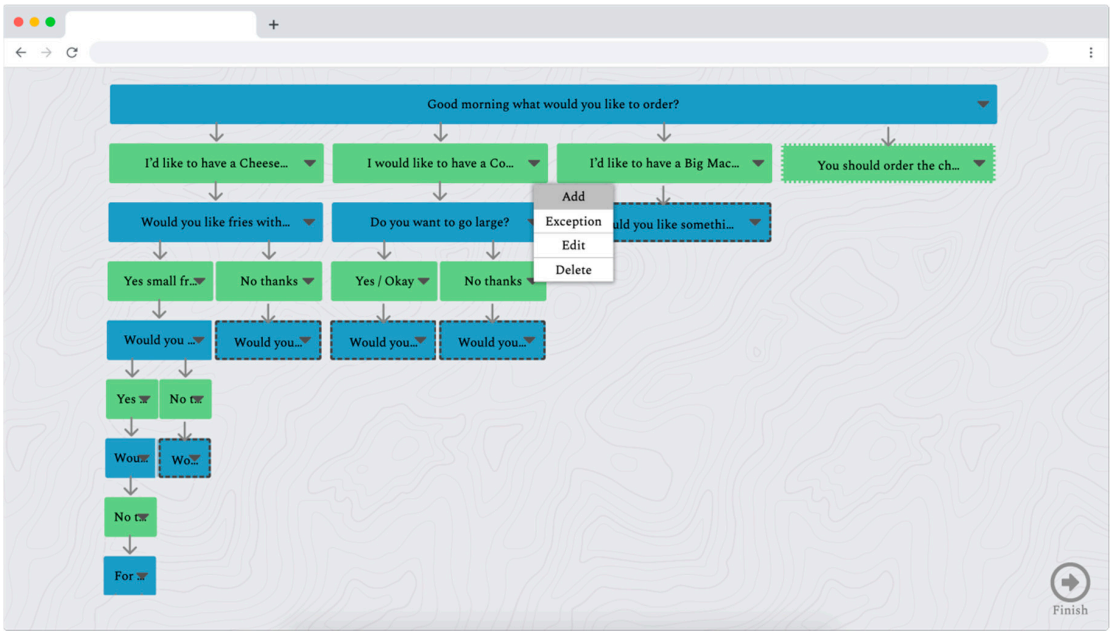

Figure 8. Task demonstration: conversational tree editing interfaces.

Figure 9 shows the teacher interface for the task-based dialogue system. The teacher can create and manage the NPCs to interact with students. The system dashboard allows the teacher to monitor the progress of students.

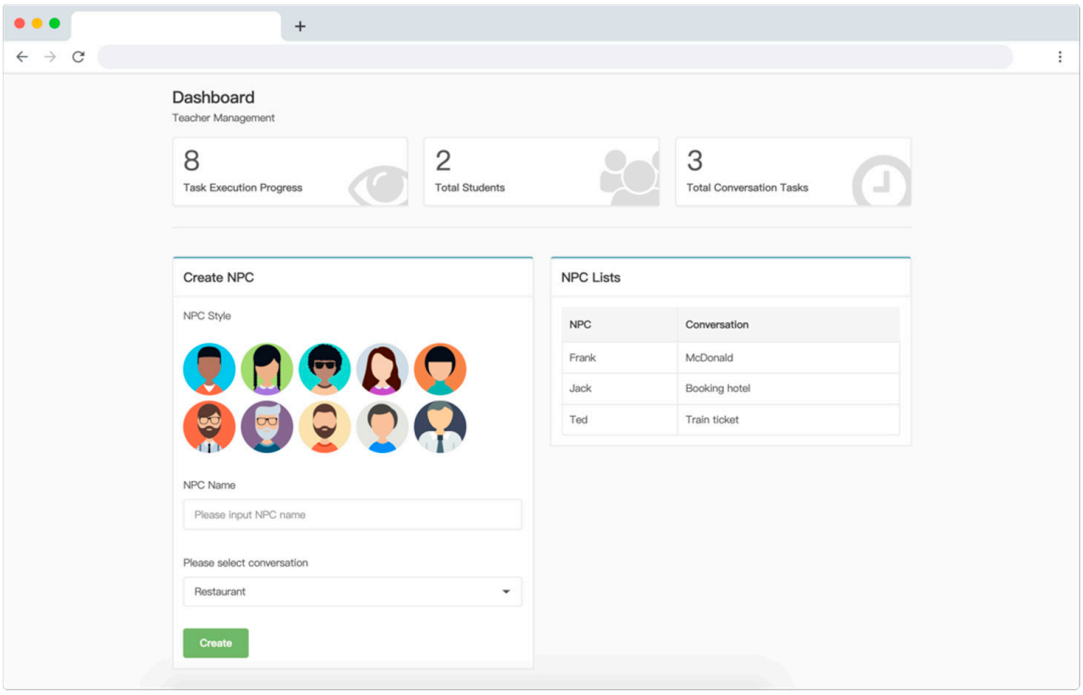

Figure 9. Teacher editing interfaces.

Figure 10 shows the learning status of students including the detail scores, time, and number of completions for each task. The score is given by the auto-scoring module in the system. Different scoring mechanisms can be selected, though the default scoring method is a rule-based point-deduction scoring method. The score is deducted by the number of wrong answers, number of repetitions, and number of hints used throughout the task. Through this interface, the teacher can replay and manually evaluate the conversations.

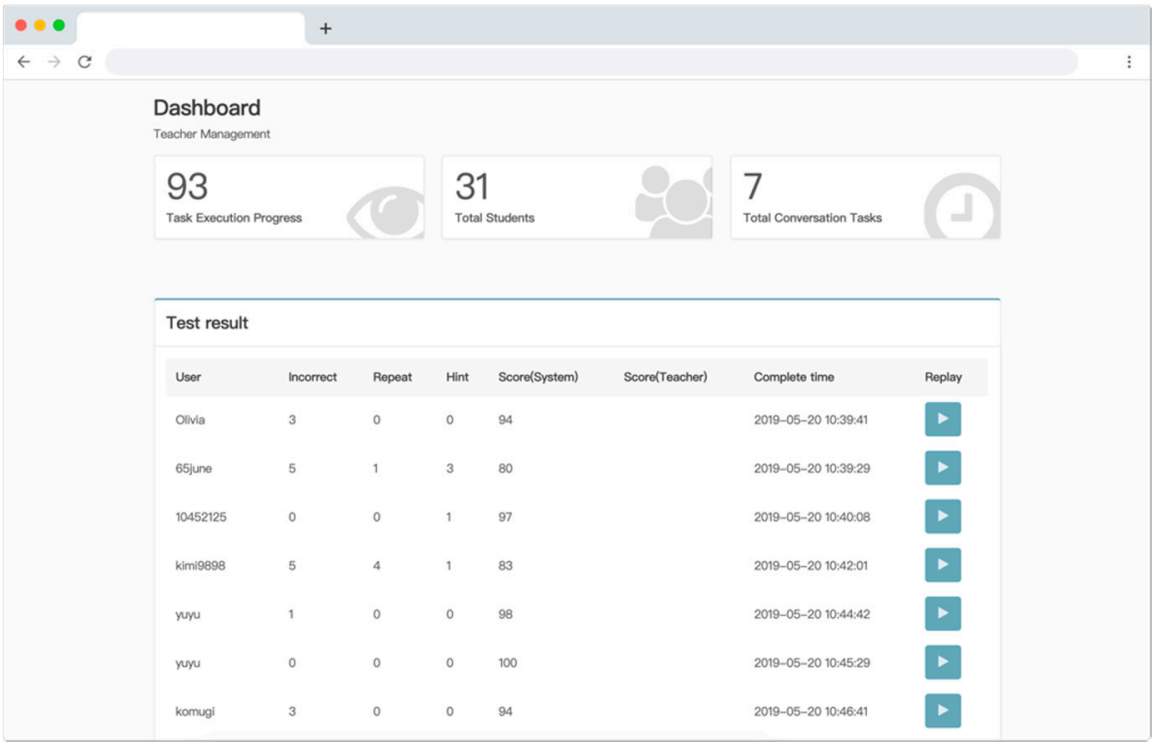

Figure 10. Learning status.

3. Results

To address the research objectives, each conversation was recorded and evaluated by an English teacher. The score was compared to the different scoring mechanisms provided by the system to check the accuracy of the scoring system. In addition, a questionnaire was distributed after the experiment to evaluate the efficiency of the dialogue system.

The study collected a total of 636 records and 51 complete task data. Data for each complete task were evaluated using five scoring criteria by the task-based dialogue system and the same teacher who taught this class. The "correct" score refers to the score given by the teacher. The teacher was able to evaluate all the data recorded by the system while students were performing the tasks. The task dialogue can be re-produced based on the system records so that the teacher can evaluate the students' performance and give a score similar to that of face-to-face scoring criteria. Different criteria from the system were combined and tested to obtain accurate prediction. The criteria included pause time, answer time, number of errors, the number of repetitions, and the number of hints (reminders). Based on the suggestion from the teacher, the teacher judged students' performance based on the pause time after the question was asked. The number of incorrect responses also reflects the comprehension of the given dialogue. The number of repetitions and number of hints are also possible criteria suggested by the English teacher. The system recorded those criteria and used them to train a model to predict the "correct" score given by the teacher. Three different methods were tested in this experiment. The first method was a rule-based evaluation method. The rating rule was based on point-deduction rules given by the teachers. The number of errors, the number of repetitions, and the

number of hints were considered for this method. Points were deducted whenever the rule is triggered. The deducted point for each rule was also suggested by the teacher. The second and the third methods used machine-learning algorithms to predict the scores. A multilayer feed-forward neural network was used to train and predict the score with different criteria as the input data and the final score as the output. The second method used neural network prediction taking the same criteria from the first method as input data, namely number of errors, the number of repetitions, and the number of hints. The third method was also a neural network approach considering all the five criteria recorded by the system, namely pause time, answer time, the number of errors, the number of repetitions, and the number of hints (reminders). The prediction models of the neural network methods were trained using the corresponding criteria and expected scores given by the teacher. The system uses the M5P algorithm to predict the nonlinear model. M5P is a machine-learning algorithm published by Yong Wang and Ian H. Witten in 1997 [38]. In practice, most of the prediction targets (classes) to be predicted by many machine learning research problems are continuous values, but only a few machine learning methods can handle continuous numerical predictions. M5P is one of the machine-learning algorithms that is able to predict the continuous value. Training involves 10-fold cross-validation. The 10-fold cross-validation is used to test the accuracy of the algorithm. The validation divides the data set into 10 parts, and takes turns using nine parts as training data and one part as test data for testing. Each test obtains the corresponding correct rate (or error rate). The average of the correct rate (or error rate) of the results of 10 repetitions is used as an estimate of the accuracy of the algorithm. Generally, it is necessary to perform multiple 10-fold cross-validation (for example, 10 times 10-fold cross-validation) and then find the average value as the algorithm accuracy rate. Based on 10-fold cross validation, 90% of the data were used as training data, and the remaining 10% were used as testing data.

Figure 11 shows the predicted results of the different system methods. The X-axis shows the completed 51 tasks. The Y-axis shows the corresponding scores for each task given by three different automatic grading methods and manually by the classroom teacher. The detailed scores can be found in Appendix D. As shown in the figure, three different methods all gave an evaluation close to the teacher's evaluation. Table 3 shows the error estimation among three different methods, namely system rating with point-deduction rules, the machine-learning prediction model with three features, and the machine-learning prediction model with five features. For the predicted score p_i and the correct score t_i given by the teacher, root mean squared error and mean absolute error were measured based on Formulas (2) and (3). Machine learning prediction using five criteria shows the closest evaluation to the expected scores. This shows that the pause time and the answer time are crucial factors while the teacher is rating the students' conversations.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - t_i)^2}{n}} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{n}|p_i - t_i|}{n} \tag{3}$$

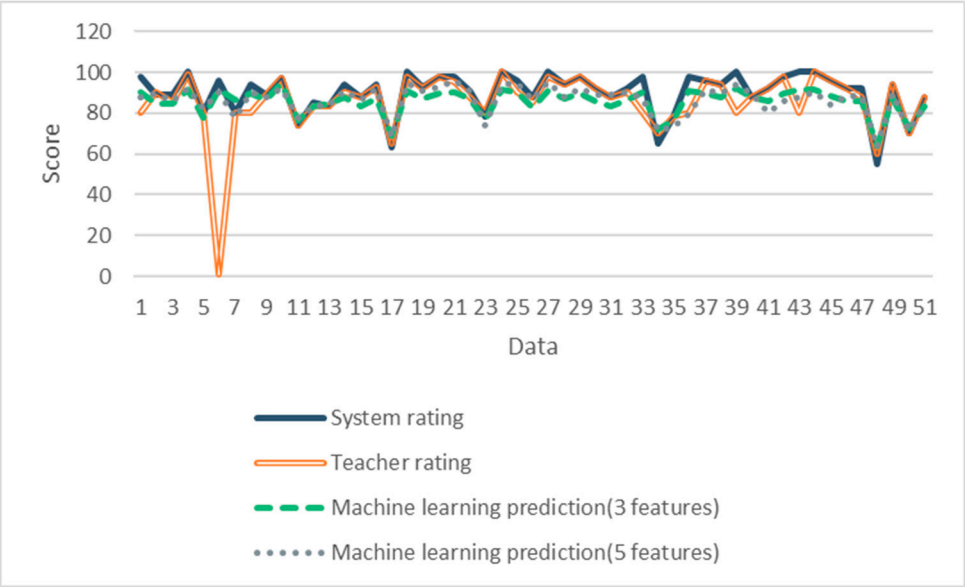

Figure 11. Predicted results of system methods.

Table 3. Error estimation.

Methods	Root Mean Squared Error	Mean Absolute Error
System rating	14.81	5.02
ML prediction (three features)	14.08	6.98
ML prediction (five features)	13.59	5.82

Right after the chatting experiment, participants were requested to fill out the online survey with 12 statements. The 12 statements were designed based on a five-point Likert scale measuring three aspects: (1) participants' perception of the user interface, (2) participants' perception of the chatting process compared to traditional instruction, and (3) participants' perception of the overall effectiveness of the system. Table 4 shows the results of the survey. The averaged score (AVG) was calculated based on three aspects. Each aspect was evaluated by four sections of the questionnaire, as shown in Appendix B. One point was scored when the strongly disagree option was given. Five points were given when the strongly agree option was given. The result of the questionnaire is shown in Appendix C. The average score of four sections was calculated to represent the perspective of the participants of the corresponding aspect.

Table 4. Survey results with three aspects.

Survey Topics	AVG
participants' perception of the user interface (Q1–Q4)	3.125
participants' perception of the chatting process compared to traditional instruction (Q5–Q8)	3.545
participants' perception of the overall effectiveness of the system (Q9–Q12)	3.511

Based on the results shown in Table 4, even though participants showed less agreement on the user interface (<3.5), they agreed that using the system to practice English conversation is better than traditional conversation practice (>3.5), and the system (including composing dialogue and practicing dialogue) is effective in general (>3.5).

The results for the first section of the questionnaire (The user interface is simple and easy to use) indicate that most participants consider the platform to be clearly designed and easy to use. However, many students were not satisfied (2.72) with the recognition accuracy rate of the speech-to-text software (Q4: The speech-to-text recognition is accurate). Based on an unofficial interview with the instructor, many students became frustrated when the machine replied that their answers cannot be recognized (because of the pronunciation, accent, or not using the pre-designed words or phrases). Once the instructor reminded the students to use only the words or phrases that were taught or focused, all the students successfully completed the three tasks. During the process, however, some students still experienced the issue that their speech could not be recognized smoothly. For example, a couple of students kept saying "Two nights", but the system showed "tonight" in the chat room. Therefore, the speech-to-text function will be modified accordingly in order to increase the accuracy rate.

Figure 12 shows the overall results of the online survey. The blue line "UI" represents participants' perception of the user interface. The red line "V.S traditional" represents participants' perception of the chatting process compared to traditional instruction. The green line "Effectiveness" represents participants' perception of the overall effectiveness of the system. The X-axis indicates the corresponding section of the questionnaire. The Y-axis shows the average score for each section.

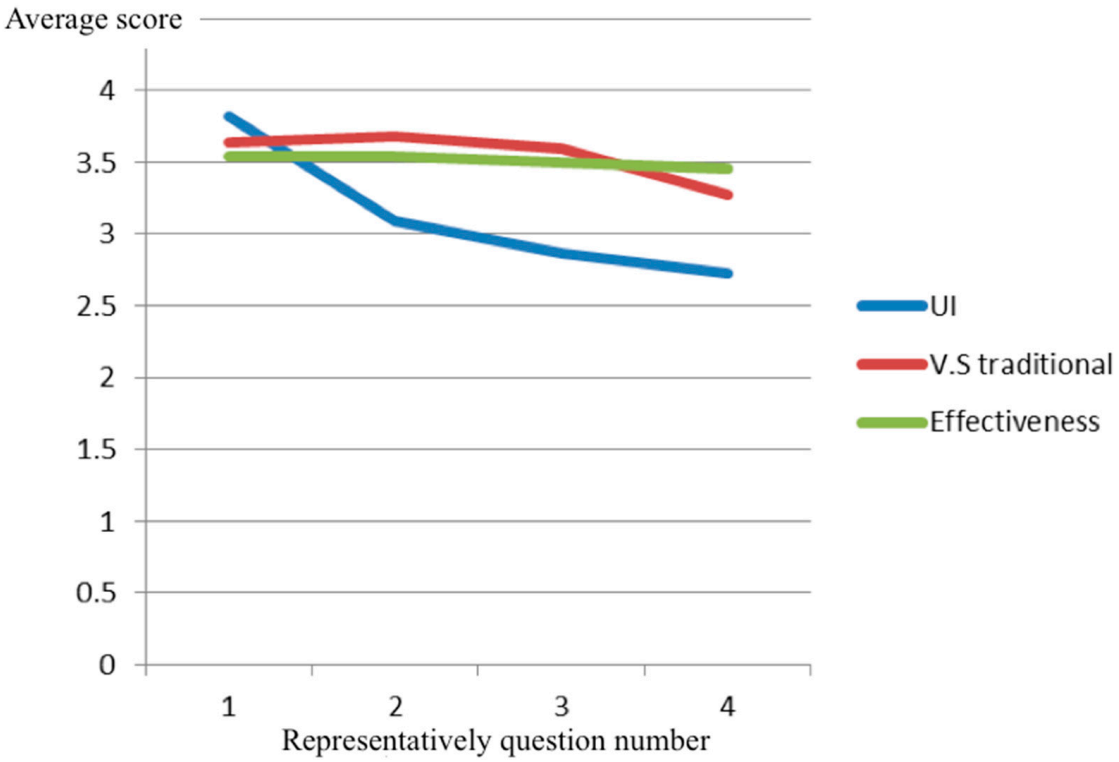

Figure 12. Overall survey results.

As shown in Figure 12, the participants responded positively regarding the overall effectiveness of the system and the chatting process compared to traditional instruction. Regarding the user interface, since the students encountered unexpected problems with the speech-to-text recognition software, and they still tended to reply with simple phrases instead of complete sentences, students did not respond with high satisfaction. All in all, however, students still expressed above average satisfaction with the conversation process and the system. They believed that computer-assisted learning environment did improve their learning motivation. Basically, they considered that the overall system design is effective for English language learners to practice speaking, and they will continue to use the system.

4. Conclusions

This study analyzed a task-oriented "English conversation" learning system. The system simulates professional English teachers to establish a grammar and sentence scoring mechanism. A task-based dialogue framework was proposed, and a preliminary system was developed to test the effectiveness of the proposed framework. The system was used in a college-level English speaking class to test the perceptions toward the system regarding user interface, learning style, and the system effectiveness. This research collected data to evaluate the possibility of replacing the traditional English speaking practice with the proposed system. During the process of performing tasks, the proposed system records the details of the learner's learning data. In addition to the grammar and vocabulary, it also includes the pause time of the dialogue and the number of repeated answers. The proposed task-based dialogue robot simulates the real life conversation. Based on the task-based language learning, students can learn the language by executing the conversational task assigned by the system. This study uses a pre-defined dialogue tree to describe the conversational task and a large quantity of Wikipedia Corpus data to train the natural language capability for the dialogue robot. Based on the collected students' feedback, results confirm the positive perceptions toward the system regarding the learning style and the leaning outcomes. The system provides better semantic understanding and more accurate task-based conversation control.

Compared to the traditional learning method, the system in this study conducts assessment automatically and analyzes learning status. Using the proposed framework, the dialogue is recorded, accessed, and compared to the regular conversation evaluation. The score is given by the auto-scoring module in the dialogue system. Three auto-grading methods were tested in this research. The dialogue system recorded the criteria suggested by teachers and used them to train a model to predict the "correct" score given by the teacher. Coherent grading using these evaluation methods was expected. In addition, the results of the questionnaire show effective learning using the task-based dialogue system. The qualitative feedback from students also provides the evidence of ease of use, usefulness of repetitive practice, and instant response.

5. Limitations and Future Works

Several limitations were observed in this study. This study only collected 51 test data generated by 28 learners for three topics. The small quantity of data affected the results of the scoring system in the machine learning training prediction model. Furthermore, the experiment was carried out in a computer lab at a university. Due to the frequent use of microphones in the dialogue system, the interference of students' voices is often an issue in a closed classroom space. The hardware equipment could be the crucial factor in a closed space environment. Finally, in this research, a language model was introduced to the dialogue manager module, so that the module could determine the corresponding response sentence by calculating the similarity between sentences. To avoid language ambiguity, increasing the corpus from more sources of the language model could be one of the possible solutions along with a challenge task.

Since the current study focused on beginner-level learners, future research should further examine and confirm these initial findings by exploring the effectiveness of the system being applied for higher-level learners. Furthermore, this system and the research design could also provide a good starting point for discussion and future development investigating task-based conversation of languages other than English. Looking forward, these suggested further attempts could prove quite beneficial to the relevant literature.

Author Contributions: K.-C.L. designed the system model and the system framework. M.C. developed the theoretical formalism. K.-H.W. implemented the system. K.-C.L. and M.C. conceived and planned the experiments. K.-C.L. and K.-H.W. carried out the experiments and collected and analyzed the data. All authors discussed the results and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

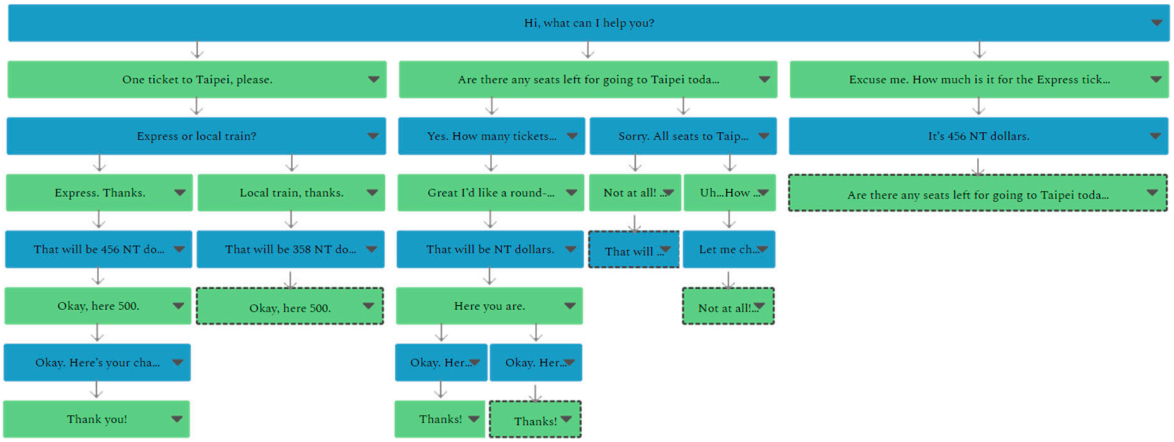

Figure A1. Task I: buying train tickets.

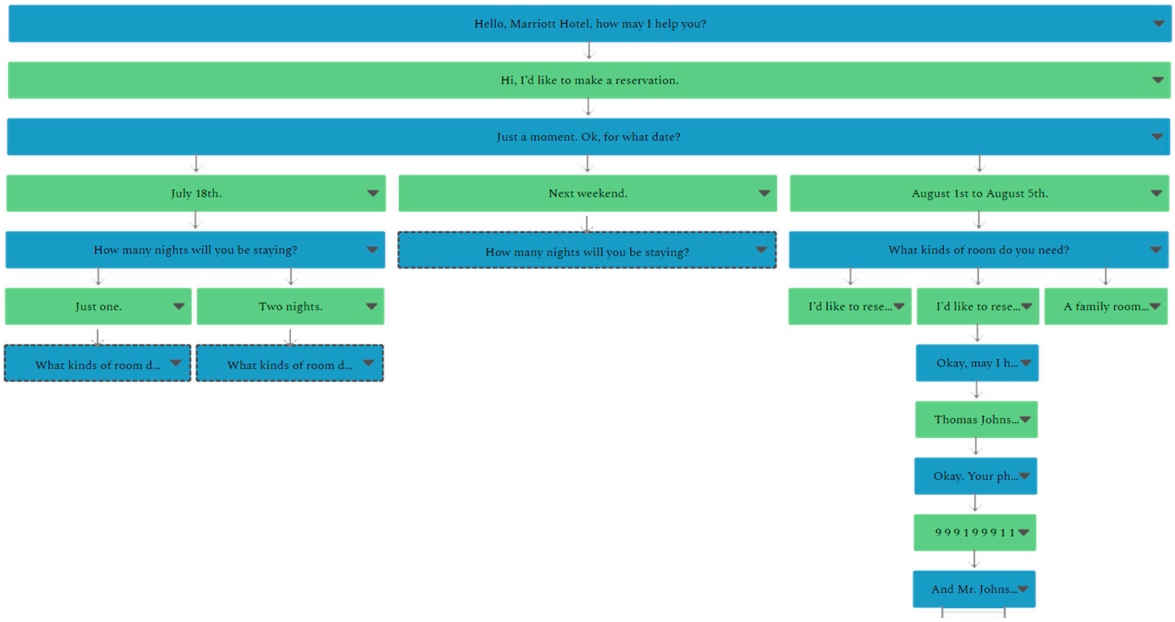

Figure A2. Task II: hotel reservation.

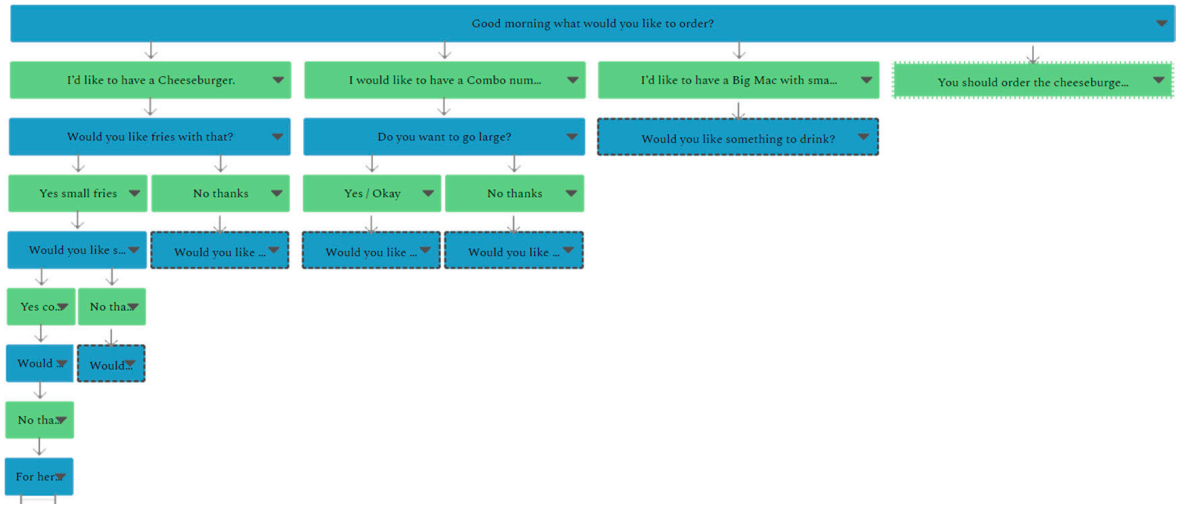

Figure A3. Task III: ordering food in a restaurant.

Appendix B

Questionnaire

Q1. The user interface is simple and easy to use.
Q2. The dialogue guidance is helpful for completing the task.
Q3. The task flow is smooth and easy to follow.
Q4. The voice recognition in the system is accurate.
Q5. Compared to the traditional classroom teaching, the content in the system is easier to access.
Q6. Compared to the traditional classroom teaching, the language-learning content in the system is easier to learn.
Q7. Compared to the traditional classroom teaching, I am better motivated to learn by using this system.
Q8. Compared to the traditional classroom teaching, I perform better by using this system.
Q9. The scoring board and social interactive functions keep me motivated.
Q10. Overall, this system is helpful in English speaking practice.
Q11. Overall, I am satisfied with the experience of using this system.
Q12. I would like to continue to use this system.

Appendix C

Table A1. Survey Results.

id	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
1	3	3	3	3	3	3	3	3	3	3	3	4
2	5	1	2	4	4	3	3	3	4	4	4	3
3	4	4	3	3	3	4	3	3	4	4	4	3
4	4	4	2	3	4	4	4	4	4	4	4	4
5	4	3	2	2	4	3	4	3	3	3	3	2
6	3	3	3	2	4	4	5	4	4	3	3	3
7	4	3	4	3	4	4	3	3	4	3	3	3
8	5	5	4	3	4	4	3	4	5	3	4	4
9	4	4	3	3	4	4	4	4	4	3	3	4
10	4	3	4	3	4	3	3	3	4	4	4	3
11	4	2	2	3	4	4	3	4	3	3	3	2
12	4	3	4	2	4	3	4	3	4	4	4	3
13	4	4	4	4	3	3	2	2	3	4	4	3
14	3	3	2	2	4	4	3	3	3	3	3	3
15	5	4	3	3	4	4	5	4	5	5	5	4
16	4	4	4	3	4	4	4	4	4	4	4	4
17	4	3	4	3	3	4	4	3	3	4	4	4
18	3	2	2	3	4	4	4	4	4	4	4	3
19	3	3	2	2	3	3	3	3	3	3	4	3
20	3	2	3	3	4	4	4	2	3	3	3	3
21	5	2	1	1	2	4	4	4	2	4	3	2
22	2	3	2	2	3	3	4	3	4	3	4	4
	AVG 3.125				AVG 3.545				AVG 3.511			

Appendix D

Table A2. Scores given by the teacher and system predictions.

System Rating	Teacher Rating	Machine Learning Prediction (Three Features)	Machine Learning Prediction (Five Features)
98	80	90.35	87.74
89	90	84.70	89.53
89	85	84.70	85.96
100	99	91.76	92.34
80	80	77.65	82.01
96	1	91.49	90.07
80	80	86.70	77.77
94	80	89.89	89.91
89	89	86.70	87.40
97	97	94.68	91.83
74	74	76.49	77.36

Table A2. *Cont.*

System Rating	Teacher Rating	Machine Learning Prediction (Three Features)	Machine Learning Prediction (Five Features)
85	83	83.39	83.38
83	83	84.77	83.91
94	90	87.53	91.11
88	88	83.39	86.25
94	93	86.98	92.01
63	65	69.15	68.21
100	98	90.81	94.86
93	92	86.98	90.72
98	98	89.53	92.93
98	95	89.92	95.78
91	88	87.19	90.86
78	80	76.29	73.75
100	100	91.28	93.91
96	90	89.92	94.49
88	85	82.91	86.55
100	98	90.88	95.36
94	94	86.89	87.82
98	98	89.55	92.15
92	92	85.56	89.11
88	88	83.42	88.79
92	90	86.26	90.06
98	80	90.54	87.27
65	70	72.02	71.44
78	78	76.30	73.46
98	80	90.79	79.43
96	96	89.33	91.61
94	94	87.88	89.74
100	80	92.24	94.21
88	88	87.88	89.75
92	92	85.81	80.17
98	98	89.88	85.90
100	80	91.24	88.02
100	100	91.24	90.28
96	96	88.53	83.70
92	92	85.91	88.12
92	88	85.91	88.33
55	60	63.72	63.18
94	94	87.21	89.65
70	70	72.85	71.37
88	88	83.30	84.08

References

1. Sykes, J.M. *Technologies for Teaching and Learning Intercultural Competence and Interlanguage Pragmatics*; John Wiley & Sons: Hoboken, NJ, USA, 2017; pp. 118–133.
2. Scholz, K. Encouraging Free Play: Extramural Digital Game-Based Language Learning as a Complex Adaptive System. *CALICO J.* **2016**, *34*, 39–57. [CrossRef]
3. Braga, J.D.C.F. Fractal groups: Emergent dynamics in on-line learning communities. *Revista Brasileira de Linguística Aplicada* **2013**, *13*, 603–623. [CrossRef]
4. Abdallah, M.M.S.; Mansour, M.M. Virtual Task-Based Situated Language-Learning with Second Life: Developing EFL Pragmatic Writing and Technological Self-Efficacy. *SSRN Electron. J.* **2015**, *2*, 150. [CrossRef]
5. Culbertson, G.R.; Andersen, E.L.; White, W.M.; Zhang, D.; Jung, M.F.; Culbertson, G. Crystallize: An Immersive, Collaborative Game for Second Language Learning. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing—CSCW '16, San Francisco, CA, USA, 27 February–2 March 2016; Association for Computing Machinery (ACM): New York, NY, USA, 2016; pp. 635–646.
6. Kennedy, J.; Baxter, P.; Belpaeme, T. The Robot Who Tried Too Hard. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction—HRI '15, Cancun, Mexico, 15–17 November 2016; Association for Computing Machinery (ACM): New York, NY, USA, 2015; pp. 67–74.
7. Morton, H.; Jack, M. Speech interactive computer-assisted language learning: A cross-cultural evaluation. *Comput. Assist. Lang. Learn.* **2010**, *23*, 295–319. [CrossRef]
8. Gamper, J.; Knapp, J. A Review of Intelligent CALL Systems. *Comput. Assist. Lang. Learn.* **2002**, *15*, 329–342. [CrossRef]
9. Hsieh, Y.C. A case study of the dynamics of scaffolding among ESL learners and online resources in collaborative learning. *Comput. Assist. Lang. Learn.* **2016**, *30*, 115–132. [CrossRef]
10. Ehsani, F.; Knodt, E. Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New Call Paradigm. *Lang. Learn. Technol.* **1998**, *2*, 45–60.
11. Tür, G.; Jeong, M.; Wang, Y.-Y.; Hakkani-Tür, D.; Heck, L.P. Exploiting the semantic web for unsupervised natural language semantic parsing. In Proceedings of the 13th Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
12. Chen, Y.-N.; Wang, W.Y.; Rudnicky, A.I. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*; IEEE: New York, NY, USA, 2013; pp. 120–125.
13. Hoang, G.T.L.; Kunnan, A. Automated Essay Evaluation for English Language Learners:A Case Study ofMY Access. *Lang. Assess. Q.* **2016**, *13*, 359–376. [CrossRef]
14. González-Lloret, M. *A Practical Guide to Integrating Technology into Task-Based Language Teaching*; Georgetown University Press: Washington, DC, USA, 2016.
15. Baralt, M.; Gurzynski-Weiss, L.; Kim, Y. Engagement with the language. In *Language Learning & Language Teaching*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2016; pp. 209–239.
16. Allen, J.P.B.; Prabhu, N.S. Second Language Pedagogy. *TESOL Q.* **1988**, *22*, 498. [CrossRef]
17. Leaver, B.; Willis, J. *Task-Based Instruction in Foreign Language Education: Practices and Programmes*; Georgetown University Press: Washington, DC, USA, 2005.
18. Yuan, F.; Willis, J. A Framework for Task-Based Learning. *TESOL Q.* **1999**, *33*, 157. [CrossRef]
19. Willis, D.; Willis, J. *Doing Task-Based Teaching*; Oxford University Press: Oxford, UK, 2007; pp. 56–63.
20. Sarıcoban, A.; Karakurt, L. The Use of Task-Based Activities to Improve Listening and Speaking Skills in EFL Context. *Sino-US Engl. Teach.* **2016**, *13*, 445–459. [CrossRef]
21. Rabbanifar, A.; Mall-Amiri, B. The comparative effect of opinion gap and reasoning gap tasks on complexity, fluency, and accuracy of EFL learners' speaking. *Int. J. Lang. Learn. Appl. Linguist. World* **2017**, *16*, 55–77.
22. Fallahi, S.; Aziz Malayeri, F.; Bayat, A. The effect of information-gap vs. opinion-gap tasks on Iranian EFL learners' reading comprehension. *Int. J. Educ. Investig.* **2015**, *2*, 170–181.
23. Wen, T.-H.; VanDyke, D.; Mrkšić, N.; Gasic, M.; Barahona, L.M.R.; Su, P.-H.; Ultes, S.; Young, S. A Network-based End-to-End Trainable Task-oriented Dialogue System. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Volume 1. Long Papers.

24. Chen, H.; Liu, X.; Yin, D.; Tang, J. A Survey on Dialogue Systems. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 25–35. [CrossRef]
25. Fryer, L.; Carpenter, R. Emerging Technologies: Bots as Language Learning Tools. *Lang. Learn. Technol.* **2006**, *10*, 8–14.
26. Shawar, B.A.; Atwell, E. Chatbots: Are they really useful? *LDV-Forum Band* **2007**, *22*, 29–49.
27. Brandtzaeg, P.B.; Følstad, A. Chatbots: Changing user needs and motivations. *Interactions* **2018**, *25*, 38–43 [CrossRef]
28. Azadani, M.N.; Ghadiri, N.; Davoodijam, E. Graph-based biomedical text summarization: An itemset mining and sentence clustering approach. *J. Biomed. Inform.* **2018**, *84*, 42–58. [CrossRef]
29. Zhou, M.; Duan, N.; Liu, S.; Shum, H.-Y. Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering* **2020**, *6*, 275–290. [CrossRef]
30. Marshall, I.J.; Wallace, B.C. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Syst. Rev.* **2019**, *8*, 1–10. [CrossRef]
31. Feine, J.; Gnewuch, U.; Morana, S.; Maedche, A. A Taxonomy of Social Cues for Conversational Agents. *Int. J. Hum.-Comput. Stud.* **2019**, *132*, 138–161. [CrossRef]
32. Go, E.; Sundar, S.S. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Comput. Hum. Behav.* **2019**, *97*, 304–316. [CrossRef]
33. Ivanov, O.; Snihovyi, O.; Kobets, V. Implementation of Robo-Advisors Tools for Different Risk Attitude Investment Decisions. In Proceedings of the 14th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops (ICTERI 2018), Kyiv, Ukraine, 14–17 May 2018; pp. 195–206.
34. Jung, D.; Dorner, V.; Glaser, F.; Morana, S. Robo-Advisory: Digitalization and Automation of Financial Advisory. *Bus. Inf. Syst. Eng.* **2018**, *60*, 81–86. [CrossRef]
35. Jung, D.; Dorner, V.; Weinhardt, C.; Pusmaz, H. Designing a robo-advisor for risk-averse, low-budget consumers. *Electron Mark.* **2017**, *28*, 367–380. [CrossRef]
36. Shang, L.; Lu, Z.; Li, H. Neural Responding Machine for Short-Text Conversation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2015; Volume 1, pp. 1577–1586, Long Papers.
37. DeNoyer, L.; Gallinari, P. The Wikipedia XML corpus. *ACM SIGIR Forum* **2006**, *40*, 64–69. [CrossRef]
38. Wang, Y.; Witten, I.H. Induction of model trees for predicting continuous classes. In Proceedings of the Poster Papers of the European Conference on Machine Learning, Prague, Czech Republic, 23–25 April 1997; pp. 128–137.