IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 December 2022, accepted 19 December 2022, date of publication 29 December 2022,
date of current version 31 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3233224

APPLIED RESEARCH

DIRECT: Toward Dialogue-Based Reading Comprehension Tutoring

JIN-XIA HUANG, YOHAN LEE, AND OH-WOOG KWON
Language Intelligent Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Republic of Korea

Corresponding author: Jin-Xia Huang (hgh@etri.re.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through the Korea Government (MSIT) (Development of Semi-Supervised Learning Language Intelligence Technology and Korean Tutoring Service for Foreigners) under Grant 2019-0-00004.

ABSTRACT A major challenge in education is to provide students with a personalized learning experience. This study aims to address this by developing a dialogue-based intelligent tutoring system (ITS) that imitates human expert tutors. The ITS asks questions, assesses student answers, provides hints, and even chats to encourage student engagement. We constructed the Dialogue-based Reading Comprehension Tutoring (DIRECT) dataset to simulate real-world pedagogical scenarios with the assessment labels and key sentences to support tutoring. The DIRECT dataset is based on RACE, which is a large-scale English reading comprehension dataset. In addition, we propose a neural pipeline approach to model the tutoring tasks and conduct a comprehensive analysis on the results, including a human evaluation. The results show that our model performs well in generating questions, assessing answers, and chatting, showing high potential although some challenges remain. The proposed model provides a good basis for further development of dialogue-based ITSs.

INDEX TERMS Computer aided instruction, dialogue-based tutoring, educational technology, intelligent tutoring, natural language processing.

I. INTRODUCTION

There is an increasing interest in education individualized for the background and achievement level of each student. In fact, it has been shown that a lack of individualized instruction results in an educational gap [1], [2], [3]. As a solution to this, intelligent tutoring systems (ITSs), which aim to scale up individualized education by imitating human expert tutors, are actively explored. Many ITSs have been successfully deployed to improve students' achievements and learning efficiency in a broad range of educational domains such as language learning and scientific reasoning [4], [5], [6].

In particular, dialogue-based tutoring is one of the most promising tutoring methods because it provides a learning environment similar to natural student–tutor interactions [7], [8]. In this setting, the tutor communicates with the student to test the student's understanding of the tutoring materials and gives appropriate instructions. For example, the tutor could ask questions or provide hints to guide the student to find the answer on their own. Moreover, they could engage in small talk, which has been shown to be an efficient tutoring technique for encouraging student engagement [9]. To address the diversity of tutoring strategies and difficulty of understanding and responding educationally to students' utterances, data-driven approaches with natural language processing techniques have been introduced for ITSs [10], [11]. However, there are few public datasets handling the diverse tutoring strategies in the discourse, despite their importance to the development and evaluation of data-driven models.

In this paper, we present the large-scale Dialogue-based Reading Comprehension Tutoring (DIRECT) dataset, in which the tutor generates questions, assesses the student's understanding of a given passage, and provides appropriate feedback. To simulate natural one-on-one tutoring scenarios, we construct multi-turn dialogues for each passage with three types of tutoring tasks, question generation, feedback, and passage-related chatting, and integrate them into one

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

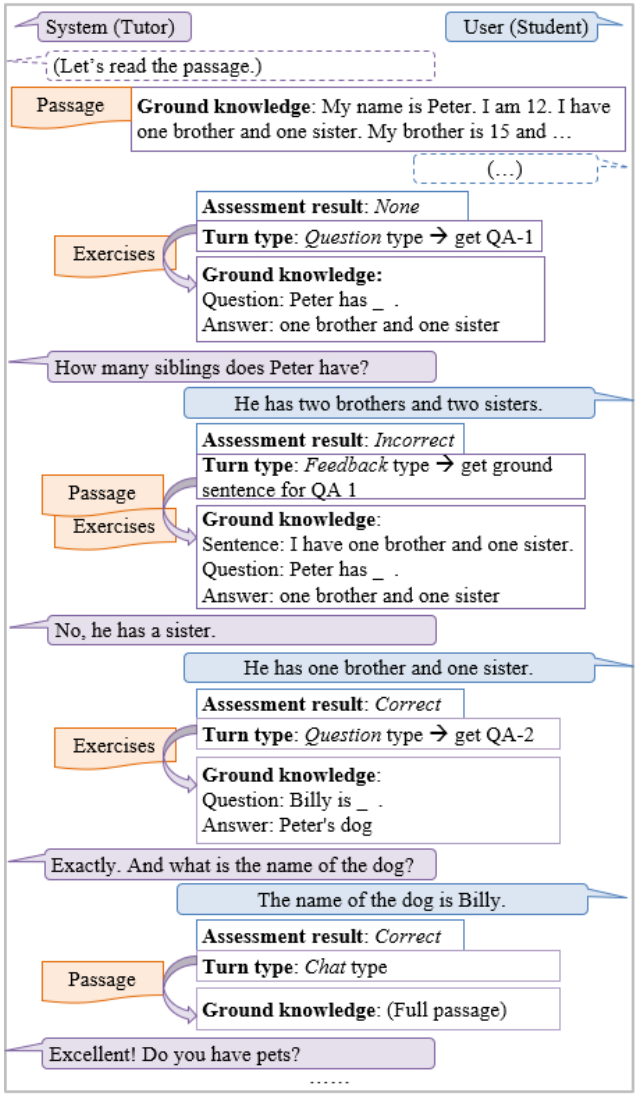

FIGURE 1. Example of the multi-turn interaction between a user (student) and a reading comprehension tutoring system.

dialogue, as shown in Fig. 1. To support the tutor's response, we provide the correctness of the student's answer and the key sentences.

To construct the data efficiently, we use the passages and high-quality exercises in RACE-M [12], which is a reading comprehension dataset collected from English exams for middle school Chinese students. Each exercise consists of an average of four questions, which we expand into multi-turn dialogue.

We treat the workflow of the tutoring process as a task-oriented dialogue (TOD) by formulating it using a pipeline schema, which is a mainstream method in the TOD field [13], [14]. The TOD system reads user utterances, tracks the dialogue state, searches a database, decides on system actions, and generates responses over the turns of the dialogue sessions [14]. Similarly, our tutoring model reads student utterances, assesses their answers (i.e., dialogue state tracking in the TOD), determines the tutoring turn type (i.e., makes system action decisions), selects the ground knowledge (i.e., database searching), and generates the responses.

The contributions of this manuscript are as follows:

- A large-scale dataset called DIRECT is presented that simulates a real tutoring conversation, including natural language assessment, question and feedback generation, and passage-related chat.
- We formulate a dialogue-based tutoring process with a pipeline schema, inspired by the state-of-the-art TOD model. We also conduct a comprehensive analysis on our dataset and discuss the remaining challenges.

To the best of our knowledge, our dataset is the first attempt to encompass all aspects of dialogue-based reading comprehension tutoring for a deep learning-based tutoring model. We hope that our dataset will be used as a testbed to develop more practical tutoring models. Data is partially available at: https://github.com/DIRECTDataset/DIRECT

II. RELATED WORK

A. TUTORING DATASETS

Existing public datasets for dialogue-based tutoring are scarce, and they focus on only some of the properties of tutoring. Some address vocabulary and grammar learning [15], [16] or token error prediction [17]. For reading comprehension tutoring, most studies have focused on feedback and question generation. Reference [18] released a tutoring dataset for generating natural language feedback regarding student answers. Their feedback was elaborate and challenging, but they provided direct guidance to explain which part was incorrect. In real-world tutoring, tutors respond to student errors indirectly [4]. In this respect, we believe that indirect guidance (i.e., hints) is more effective than direct guidance because it provides students the opportunity to think independently.

A pioneering large-scale question generation dataset for education is LearningQ [19]. It focuses on constructing questions that require reasoning rather than conversational tutoring. Reference [20] utilized the RACE dataset [12] for exam-like question generation while highlighting the unnaturalness of existing datasets extracted from the Web. They used only a specific type of questions, which comprise approximately 20% of the RACE questions, and an automatic tagging method to denote answer-related information in the passage without annotating key sentences, which would have led to insufficient data annotation.

Reading comprehension tutoring is closely related to question answering (QA). Most QA datasets are developed for machine reading comprehension research, with text collected from news, Wikipedia, the Web, or book corpora, and question and answers developed through crowdsourcing [21]. However, some QA datasets based on educational resources exist, including RACE [12], RACE-C [22], ARC [23], DREAM [24], and ReClor [25]. These datasets are multi-choice datasets that provide multiple options for a question, and they can be used as tutorial materials to test a learner's reading comprehension ability. They

consider challenges such as commonsense or multi-sentence reasoning, as well as addressing unanswerable questions. These datasets can be used as question generation datasets for education after excluding Cloze-type questions [20].

Unlike other tutoring datasets, we consider overall dialogue-based tutoring strategies such as questions, assessment, feedback, and chat with human-annotated ground knowledge. The DIRECT dataset, which is a dialogue-based tutoring dataset, the differences from other dialogue datasets are as follows:

First, the dialogue participants of the DIRECT dataset are students and tutors, whereas the dialogue participants for conventional TOD datasets are users and clerks, and users and wizards or only two people in conventional open domain dialogue (ODD) datasets. The main types of system utterances in DIRECT are questions, feedback, and chats, whereas the main types are request and information in TOD datasets and only chat in ODD datasets. The user's utterance type is answer, whereas the utterance type is query for TOD datasets and chat for ODD datasets. Regarding ground knowledge, in DIRECT, it is a passage of English, whereas in TOD, it is a database, and in ODD, it is an open domain text.

B. DIALOGUE MODELING

In the past few years, deep learning-based dialogue techniques have significantly grown in both the TOD and ODD fields. Recent work on ODD has explored knowledge-grounded dialogue, which is related to our work because it involves informative responses with textual ground knowledge [26], [27], [28], [29], whereas we consider a different source of ground knowledge at each turn to provide appropriate instructions. We use knowledge-grounded ODD models as a baseline. In the TOD field, modeling sequential pipeline schemas with large pretrained language models have shown state-of-the art performance [13], [14], [30]. In particular, modeling dialogue on a session level has improved system performance [13], [14]. Inspired by that success, we model the sequential tutoring conversation at session level.

Deep learning-based ITSs for generating feedback and questions have achieved promising results. Reference [11] deployed an ITS for generating personalized feedback. To provide instruction to students about concepts that were misunderstood, they analyzed the relationship between the answer and ground knowledge using natural language processing techniques such as segmentation and sematic parsing. Reference [20] proposed a sequence-to-sequence architecture for an answer-aware question generation task. Reference [31] extended their approach by adopting an iterative question–answer generation task. Unlike the above studies that only refer to passages and answers, our tutoring model refers to both question-and-answer pairs in the exercises. This is primarily because our aim is a practical tutoring system, in which reliable performance is critical. This is also similar to human reading comprehension tutoring in the real world.

III. DATASET CONSTRUCTION

The DIRECT dataset adopts the RACE dataset reading worksheets to simulate tutoring conversations. We chose RACE because it is a collection of reading comprehension exercises constructed by human experts for practical teaching [12]. We focused on the RACE-M subset for middle school examinations, and passages with one or two questions or questions with serious errors were excluded from the dialogue construction.

Four experts with a bachelor's or master's degree in English and more than two years of experience in English translation, two of them with experience teaching English, participated in the dataset construction. To effectively construct high-quality dialogue, one person was responsible for a given passage, including constructing the conversations and annotating key sentences for the associated exercises. Early in the work, the experts cross-checked the constructed data and often had discussions to reach consensus on the guidelines. Table 1 presents examples of the data provided by DIRECT and RACE for comparison.

A. TUTORING DIALOGUES

A dialogue that includes several turns between the tutor and student was constructed for each reading comprehension passage. Similar to real-world teaching, we assume that the teaching scenario in which the dialogue occurs is as follows: the tutor is provided with full information about the examination, including the passage, questions, and correct answers. The tutor leads the conversation, asks questions, and gives feedback on student responses. The feedback includes hints so that the student can determine the correct answer. The tutor also initiates passage-related chats to arouse student interest.

The student is provided with the passage, questions, and candidate answers; thus, the student will choose an answer from the candidate answers that is correct or incorrect. The student actively participates in the conversation and can seek hints when they fail to determine the correct answer.

The following three types of turns are included in one dialogue:

- *Question* type: The tutor asks the questions in the reading comprehension examinations. Each question in the exercise has one *Question*-type turn.
- *Feedback* type: This refers to an utterance to elicit the correct answers when students fail to answer the *Question*-type tutor utterance correctly. Each question in the exercise has at most one *Feedback*-type turn.
- *Chat* type: The tutor utters passage-related chat before, after, or in the middle of the tutoring conversation. Each dialogue has at least one *Chat*-type turn.

We assume that the student initially answers about 50%–60% of the *Question*-type questions correctly and determines the correct answer after receiving feedback. To focus on dialogue-based tutoring, the following types of questions and answers are avoided:

TABLE 1. Sample tutoring dialogue in the DIRECT dataset for a passage and exercise from RACE-M.

Passage: Today is Sunday. It is sunny. Kate and her friends go to the beach. There are lots of people here now. Some are playing volleyball. Others are swimming in the sea. Look at this group of people singing and taking a sunbath on the beach. After swimming for some time, Kate feels very tired. So she has a rest at the swimming club on the beach. Where are Kate's best friends? Susan is learning to swim in the water. Gina is helping her to learn swimming. Susan is clever. I think she can swim soon.

Exercise:

Idx	Question	Candidate answers	Answer
1	Kate and Susan are___ on Sunday.	A. "on the beach"; B. "at home"; C. "in the school"; D. "in the park"	A
2	There are some people on the beach. Some are ___. Others are ___.	A. "playing basketball; singing"; B. "playing volleyball; swimming"; C. "staging; dancing"; D. "taking photos; singing"	B
3	Kate has a rest because she feels ___.	A. "happy"; B. "cold"; C. "tired"; D. "sad"	C
4	Which isn't mentioned in the article?	A. "Some people are playing volleyball."; B. "Some people are taking photos."; C. "Some people are swimming."; D. "Some people are taking a sunbath."	B

(a) A passage and its exercise from RACE-M

Dialogue:

Type	Tutor utterance	Student utterance	Label
Question	Where are Kate and Susan on Sunday?	They are at home.	*Incorrect*
Feedback	No, they go swimming.	Kate and Susan are on the beach on Sunday.	*Correct*
Chat	That's right! / Do you enjoy swimming?	Yes, my friends call me a seal because I love swimming.	*None*
Question	Good to know! / There are some people on the beach. What are they doing?	Some are playing volleyball, and others are swimming.	*Correct*
Chat	Correct. / How did you learn swimming?	My father taught me how to swim when I was a little girl.	*None*
Question	Good for you! / Why does Kate have a rest?	It is because she feels sad.	*Incorrect*
Feedback	Not really. It was after swimming for some time.	Kate has a rest because she feels tired.	*Correct*
Question	Excellent! / Is it true that some people are taking photos?	No, it is not true.	*Correct*

Key Sentences:

QA-idx	Ground Sentence
QA-1	Kate and her friends go to the beach.
QA-2	Some are playing volleyball. Others are swimming in the sea.
QA-3	After swimming for some time, Kate feels very tired. So, she has a rest at the swimming club on the beach.
QA-4	Full text

(b) A dialogue and ground sentences newly constructed for DIRECT, which is based on RACE-M

• tutor questions (such as "Do you have any questions?") that may cause the student to ask a question;

• student answers with lexical or grammatical errors because the dialogue focuses on improving student's reading skills and not their grammar;

• for *Chat*-type questions, student answers are unrelated to the question.

All dialogues were produced in English. The dialogue continues until the student finishes the given exercise. When the student answers the last question correctly, the conversation is ended without closing comments.

B. OTHER ANNOTATIONS

Key sentences are assigned to each question-and-answer pair. They can be adopted as ground knowledge to generate the tutor's questions and feedback. The key sentences are selected from the passage, and up to two sentences are allowed. If there are three or more key sentences, only the beginning and last sentences from the passage are added to the key sentences, and a "∼" is added in front of the second sentence. The following annotations are used for the key sentence annotation:

• "Full text" for questions related to the subject or summary of the given passage;

• "Background" for questions that must be answered using common sense;

• "Unknown" for questions that either cannot be answered using the information provided in the passage or for which the correct answer is "the story didn't tell us about this";

•"Other-calculation," "Other-counting," or "Other-table" are added in front of key sentences if additional calculation or counting is required to answer the question or the key information is a table.

Assessment labels "Correct" or "Incorrect" are added to the student responses for *Question* and *Feedback* types. The responses to *Chat*-type questions are assigned "None" by default.

C. DATASET STATISTICS

The statistics for the dialogues in the DIRECT dataset are summarized in Table 2.

RACE-M contains 7,139 passages with 28,293 questions [12] and there are 5,708 dialogues developed with 23,982 *Question*-type turns in DIRECT (Table 2). Hence, approximately 79.96% of the passages and 84.76% of the questions in RACE-M were adopted for constructing the DIRECT dialogues. The number of *Feedback*-type turns is approximately 43.49% of the number of

TABLE 2. Statistics of the DIRECT dialogues.

Description	Train	Dev	Test	Total
Dialogues	5,099	302	307	5,708
Question turns	21,463	1,239	1,280	23,982
Feedback turns	9,431	475	524	10,430
Chat turns	10,481	582	625	11,688
Total number of turns	41,377	2,296	2,429	46,102
Average number of turns	8.11	7.60	7.91	8.08

Question-type turns, which indicates that the percentage of students initially answering the question correctly is 56.51% and the proportion of those that answer correctly on their second try is 43.49%. A total of 69.16% of the assessment labels on the student answers were "Correct," which indicates that the correct rate of student answers was 69.16% in all *Question* and *Feedback* turns.

Table 3 presents the statistics of the key sentences. Only 58.34% of the key sentences consist of a single sentence. This indicates that the key sentence annotations in DIRECT together with the questions in the dialogue part can provide a challenging dataset for future question and feedback generation research.

TABLE 3. Key sentence statistics.

Type	Train	Dev	Test	Total
Single sentence	57.98%	58.58%	64.30%	58.34%
Multiple sentences	28.26%	26.60%	22.84%	27.89%
Full text	10.49%	10.27%	10.26%	10.47%
Background	1.09%	2.28%	1.27%	1.16%
Unknown	0.06%	0.00%	0.00%	0.05%
Other	2.12%	2.28%	1.33%	2.09%

IV. METHODS

This section describes how we approach reading comprehension tutoring from the perspective of dialogue modeling and how we frame it in DIRECT. The section also presents the proposed dialogue model (Fig. 2) and its training details.

A. DIALOGUE-BASED READING COMPREHENSION TUTORING SYSTEM

The dialogue-based tutoring system aims to generate a system (tutor) utterance Y_t in a turn t, given knowledge K and user (student) utterance U_t. That is,

$$Y_t = \text{TutoringModel}(U_t, K) \tag{1}$$

Here, knowledge K refers to what the student needs to learn; in the DIRECT dataset, knowledge K includes a set of sentences p_i $(i = 1, \ldots, N)$ in passage P and a question–answer pair $e_j (j = 1, \ldots, M)$ in exercise E, where N and M are the number of sentences in the passage and the number of question–answer pairs in the exercise, respectively.

We separate the dialogue-based tutoring task into four subtasks: student response assessment (f_1), turn type selection (f_2), ground knowledge selection (f_3), and tutor utterance generation (f_4).

1) STUDENT RESPONSE ASSESSMENT

Let dialogue history $H_t = [U_0, S_0, \ldots, U_t]$; the tutoring model first needs to produce an assessment result A_t on student response U_t as

$$A_t = f_1(H_t, K), \tag{2}$$

where A_t can be one of the set $\{Correct, Incorrect, None\}$.

2) TURN TYPE SELECTION

According to the evaluation result A_t, the next utterance type T_t is determined as

$$T_t = f_2\left(A_t, T_{t-1}\right), \tag{3}$$

$$f_2 = \begin{cases} Feedback, & if\, A_t = Incorrect \text{ and } T_{t-1} \neq Feedback \\ QuestionorChat. & otherwise \end{cases} \tag{4}$$

Therefore, if the user answers the system's question incorrectly, the system will give the student a chance to try again with a *Feedback*-type utterance. Otherwise, the system proceeds to the next question in the exercise (T_t =*Question* type) or chats about the learning passage (T_t =*Chat* type).

3) GROUND KNOWLEDGE SELECTION

The next step is selecting ground knowledge K_t from knowledge K according to utterance type T_t. Assuming that at turn t, the question–answer pair is e_j, and its key sentences comprise a subset of the passage that is denoted as $S(S \subset \text{ø}P)$. The ground knowledge K_t is as follows:

$$K_t = f_3\left(T_t\right), \tag{5}$$

$$f_3 = \begin{cases} e_j, & if\ T_t = Question \\ \left[S, e_j\right], & elif\ T_t = Feedback \\ P. & else \end{cases} \tag{6}$$

Thus, if T_t is a *Question* type, a question–answer pair will be used to generate the utterance; if T_t is a *Feedback* type, K_t will be the concatenation of the question–answer pair from the previous *Question* turn, and its key sentences will be selected from the passage; if T_t is a *Chat* type, the entire passage will be considered as ground knowledge to produce topic-related free chats.

4) TUTOR UTTERANCE GENERATION

A system utterance Y_t is generated conditioned on all prior information and concatenated as a single sequence as follows:

$$Y_t = f_4(K_t, T_t, A_t, H_t). \tag{7}$$

As a result, the sequence prediction model for the dialogue-based tutoring task is similar to the pipeline model for TOD [14], [30].

B. MODEL IMPLEMENTATION

We model tutoring tasks as a sequence prediction problem. The system reads the knowledge and student response

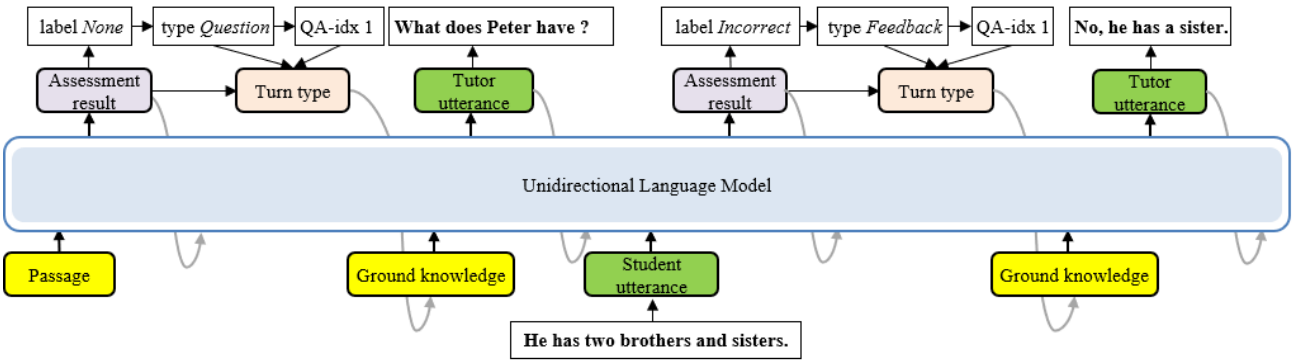

FIGURE 2. Overview of the dialogue model for reading comprehension tutoring.

and generates a sequence including the student response assessment result, turn type, and tutor utterance.

The process of the dialogue model is as follows: first, after reading the input utterance from the user, the system generates the assessment result. There is no user input for the first system utterance generation, and thus, the passage is input instead of the user utterance.

The generated assessment label is used to determine the next system turn type. Each time a *Question* type is determined, the question index number (QA-idx) is updated to store which question should be addressed in the given exercise. For a fair evaluation in the experiments, we adopted the ground-truth turn type in the case of *Question* or *Chat*.

The turn type and the QA-idx are used to retrieve the ground knowledge using (6). If the turn type is *Feedback*, the question–answer pair of the QA-idx will be adopted as the query to retrieve the key sentence from the given passage using BM25 [32]. In our experiments, the top-1 ranked key sentence is used as the ground knowledge along with the question-answer pair to generate the next system utterance.

Unless otherwise stated, our model is the unidirectional language model DistilGPT-2 [33], which is a distilled version of GPT2 [34], fine-tuned using the DIRECT and RACE datasets. The model was trained at dialogue-session level with the following parameters: the maximum sequence length was 1,024; the AdamW optimizer was adopted with greedy decoding, a temperature of 0.7, and batch size set to 2. The model had the best performance on the validation set after 65 epochs, and this model was selected for the evaluation on the test set.

V. EXPERIMENTS

A. SETTINGS

The model generated system utterances based on ground knowledge and all the previous sequences including the evaluation result and turn type, using all previous turns as context. Two settings were used for prediction: *DIRECT.allT*, in which all conditions for prediction were the ground truth, including ground knowledge, evaluation results, turn types, and previous turns, and *DIRECT.e2e*, which is an end-to-end setting in which all conditions except current user utterances were the automatically predicted results.

Similarity the evaluation metrics unigram F1, BLEU [35], [36], METEOR [37], and ROUGE-L [38], were adopted to evaluate the system utterances. N-gram diversity (n=4) DIST-4 [39] was also evaluated. The accuracies of the user utterance assessment, turn type prediction, and key sentence selection were also evaluated.

B. BASELINES

As the baseline model, we adopted the Lost-in-Conversation (LIC) model [40], which was the winner in ConvAI2 on the Persona dataset [26]. In our experiments, the baseline model achieved a better performance than one of the state-of-the-art models (DualGAN [41]) on the Wizard of Wikipedia dataset [27]. The LIC model was trained with two settings: i) all ground-truth conditions (*LIC.allT* in Table 4) and ii) with passages and dialogues but without other ground knowledge such as questions, answers, and ground-truth key sentences (*LIC.Net* in Table 4).

TABLE 4. Results with ground-truth conditions.

Model	Pretrained Model	F1	BLEU	METEOR	ROUGE-L
LIC.Net	GPT	0.2340	0.0551	0.1854	0.2233
LIC.allT	GPT	0.4655	0.2792	0.4282	0.4482
DIRECT.allT	GPT	0.4705	0.2864	0.4415	0.4543
DIRECT.allT	DGPT2	0.4918	0.3007	0.4582	0.4748

For direct comparison with LIC, which was fine-tuned on GPT [42], we fine-tuned *DIRECT.allT* on both GPT and DistilGPT-2. The parameters were the same as those in the fine-tuning of DistilGPT-2, except the maximum length was changed to 512 to fit GPT.

Table 4 presents the comparative evaluation results using the ground-truth conditions. *LIC.allT* performs much better than *LIC.Net*, which indicates that all types of ground knowledge in the DIRECT dataset, including the questions, answers, and key sentences, are necessary to produce reliable tutor utterances. The DIRECT model shows similar or slightly better results than those of the baseline LIC model.

C. EXPERIMENTS ON THE DIALOGUE MODEL

A dialogue system for practical tutoring purposes must be able to participate in a sequential conversation in which every utterance it generates at each turn is conditioned on previous user utterances and system responses.

The *assessment accuracy* of student responses was 92.05%. If only *Question* and *Feedback* types were considered, the assessment accuracy was 90.70%. The turn type prediction accuracy was 94.94% in the end-to-end model prediction.

The evaluation of *tutor utterance generation* is presented in Table 5. The performance of *Question*-type questions is quite high, but the *Feedback* type is still far from satisfactory.

TABLE 5. Evaluation results on the DIRECT.E2E model.

Turn type	F1	BLEU	METEOR	ROUGE-L	DIST-4
All	0.4621	0.2759	0.4240	0.4441	0.6525
Question	0.5978	0.3795	0.5753	0.5703	0.7248
Feedback	0.2298	0.0416	0.1711	0.2243	0.6425
Chat	0.3791	0.1413	0.3263	0.3698	0.5126

The key sentence selection accuracy was 35.65% according to the top-1 exact matching, which is only applicable to the ''single sentence'' type (listed in Table 3). Fuzzy matching accuracy was 57.38%, and the ''full text'' and ''background'' types (listed in Table 3) were counted as accurate by default. The performance of the key sentence selection was quite low, which affects the performance of the *Feedback*-type generation.

We evaluated the DIST-4 according to utterance type. As Table 5 shows, the diversity of *Chat* utterances is slightly lower than that of the other types. This may be because the *Chat* type asks for personal experience.

D. HUMAN EVALUATION

Human evaluation was conducted for the purpose of confirming the meaning of the results in the quantitative evaluation. To maintain the quality and consistency of the evaluation, two raters evaluated the system utterances on the entire test set. Early in the work, the raters discussed their cases frequently and wrote a detailed manual with examples to clarify the criteria. After evaluating approximately 10% of the data, the independent evaluation began. A second independent evaluation was required for the cases where the gap between their scores was 1.5 or more.

The system utterances under ground-truth conditions and with end-to-end settings were evaluated. The evaluation criteria were whether the utterances were sensible and specific [43], but with the following modifications: 2 points was given if it makes sense and is specific; 1 point was given if it makes sense but lacks specificity; 0.5 points was given if it makes sense but is partially incorrect or contains errors in spelling or grammar; and 0 points was given if it was incorrect or does not make sense. Cohen's Kappa between the two raters was 0.97.

The results of the human evaluation were consistent with the results of the quantitative evaluation (Table 6). The system shows good performance for both *Question*- and *Chat*-type utterances but poor performance for *Feedback*-type utterances.

TABLE 6. Human evaluation results.

Model	Score	All	*Question*	*Feedback*	*Chat*
DIRECT.allT	2	65.44%	78.87%	25.19%	71.68%
	1	9.67%	8.55%	18.13%	4.88%
	0.5	3.71%	5.70%	1.43%	1.52%
	0	21.18%	6.88%	55.25%	21.92%
DIRECT.e2e	2	60.44%	74.57%	18.03%	67.04%
	1	8.56%	7.27%	16.79%	4.32%
	0.5	3.48%	4.84%	1.62%	2.24%
	0	27.52%	13.32%	63.55%	26.40%

E. ANALYSIS AND DISCUSSION

The purpose of feedback is to provide hints rather than direct answers to encourage students to read and understand on their own. We analyzed 50 human-constructed feedback utterances in DIRECT and found that they can be divided into reasoning hints, key sentence hints, and general types (see the Appendix). Approximately 32% of the feedback utterances were reasoning hints that required background knowledge or reasoning. The key sentence hints, that is, trying to help students find key sentences, but avoiding exact answers, comprised 50% of the utterances. Both types are rather challenging to generate as feedback. Approximately 18% of the utterances were the general type, that is, they required students to read more sentences or asked them to reread carefully. In the automatically generated feedback, 70% of the utterances provided correct or incorrect answers based on key sentences, 8% were of the general type, and 22% were of the reasoning type, that is, they used world knowledge or informed the student why the answer was wrong.

Question-type system utterances were generated using exercise questions and answers as ground knowledge. We considered whether the system learns the dialogue or outputs the original question in the grounded knowledge unchanged. In particular, this could happen because there are many complete sentence-type questions in the RACE dataset. To determine this, we compared the generated utterances with the human-developed target utterances and their corresponding questions in the RACE exercise.

TABLE 7. Evaluation of question-type generation with human-constructed and original questions.

	F1	BLEU	METEOR	ROUGE-L
Pred-Q vs. Human-Q	0.5978	0.3795	0.5753	0.5703
Pred-Q vs. Orig-Q	0.5652	0.1798	0.6486	0.5240
Orig-Q vs. Human-Q	0.5152	0.1596	0.6004	0.4717

As shown in Table 7, the generated utterances are closer to the human-constructed tutor utterances (the ''Pred-Q vs.

Human-Q'' row in the table, and the *Question* row in Table 5) but less similar to the corresponding questions in the exercises (''Pred-Q vs. Orig-Q'' in Table 7). This indicates that the system was well-trained with target utterances, rather than outputting the ground-truth questions. Furthermore, the target utterances had a relatively low similarity to the questions of the exercises (''Orig-Q vs. Human-Q''), which indicates that the human experts tried to construct target tutor utterances using expressions that were different to those in the given questions.

We finally consider whether it is necessary to generate all of the questions related to a passage in one dialogue. This question is reasonable if only question generation is considered; in this case, the dialogue context for previous questions does not contribute to subsequent questions. However, our goal is to propose a tutoring model for four sub-tasks, including student response assessment, turn-type selection, knowledge selection, and tutor utterance (question or feedback) generation. The prediction results of these four subtasks depends on the previous ones, just as it does in TOD tasks. The pipeline schema proposed in our paper is necessary if we want to obtain a solution using a unified model.

However, a dialogue model that utilizes ground knowledge from different dialogue contexts in different turn types is worth investigating. In such a model, ground knowledge from at least one previous turn is used to generate tutor *Feedback*-type utterances, and those from as many previous turns as possible are used to generate *Chat*-type utterances. This type of model must be trained on turn-level sequences, which performed much worse than the dialogue-session level training adopted in [14]. We aim to investigate such a system in our future work.

VI. CONCLUSION

We proposed a large-scale dialogue dataset called DIRECT for reading comprehension tutoring. Tutors ask questions and give feedback, while student answers can be correct or incorrect. Similar to a tutor in a real one-on-one tutoring situation, this dataset also includes a chat about a given passage. Other annotations, including key sentences and student answer assessment labels, are also provided so that the DIRECT dataset can be used for other tutoring tasks, including question and feedback generation as well as student answer assessment.

We also formulated a dialogue-based reading comprehension tutoring process with a pipeline schema and implemented a neural network model. It sequentially generated assessments of the student responses, tutor questions, and feedback. A series of experiments indicate that the performance is quite promising as a first attempt at a dialogue-based tutoring model. However, there is still much room for improvement, particularly with respect to the key sentence selection and feedback generation. Considering that most reading materials do not have exercises for reference, further research on a series of tutoring tasks involving only passages and answers is required.

TABLE 8. Reasoning hint feedback example 1.

	Contents
Question	The most important reason for the popularity of kitesurfing is that ___.
Answer	Its equipment progress makes it easier and safer.
Key	Now it is becoming easier and safer because of the safer kite design.
Student	Because all people can learn and take part in it.
Feedback	It's not right. A hint is that it's useful for people.
Auto-Ques	Good. What is the most important reason for the popularity of kitesurfing?
Auto-Stud	Because all people can learn and take part in it.
Auto-Key	With the development of its equipment progress, kitesurfing is becoming even safer.
Auto-Feed	Not really. The answer is related to the changes in kitesurfing. (score 2)

The dialogue-based tutoring model proposed in this paper can be used in a variety of educational contexts, including tutoring related to reading comprehension such as language and history in primary and secondary schools, as well as professional training programs.

VII. LIMITATIONS

Compared to other ODD or TOD datasets, our dataset is not particularly large (5,708 dialogue with 46,102 turns; see Table 2). For example, the Wizard-of-Wikipedia dataset, which is a document-grounded ODD dataset, contains 22,311 dialogues with 201,999 turns [27], and the MultiWOZ dataset, which is widely used for TOD research, contains 8,438 dialogues with 113,556 turns [44].

However, our dataset cannot be considered small compared to datasets used for tutoring purposes or TOD datasets released prior to MultiWOZ. For example, the question generation dataset EQG-RACE for reading comprehension [20] consists of 20,486 questions (ours includes 23,982 questions, 10,430 feedback utterances, and 11,688 chats; see Table 2), and the SAF dataset [18], which is a short answer feedback dataset, contains 4,519 submissions of German and English questions.

Most datasets used for tutoring purposes are small due to the lack of publicly available tutoring materials. We hope that our dataset will stimulate more research in the field of tutoring.

APPENDIX: EXAMPLES OF FEEDBACK

There are three types of feedback in the DIRECT dataset according to our analysis: hints that need reasoning, hints to help student find key sentences, and general-type hints.

In the following examples, ''Question,'' ''Answer,'' ''Key sentence'' (''Key''), ''Student,'' and ''Feedback'' (shown in black) are ground-truth data provided by the RACE and DIRECT datasets. ''Auto-key sentence'' (''Auto-Key''), ''Auto-Question'' (''Auto-Ques''), and ''Auto-Feedback'' (''Auto-Feed'') (shown in gray) were retrieved or generated by *DIRECT.e2e*. ''Auto-Student'' (''Auto-Stud'') is the input to the model, which is the same as the ground-truth ''Student.'' The first short sentence in an ''Auto-Ques'' is

TABLE 9. Reasoning hint feedback example 2.

	Contents
Question	The dog, Pluto, is ___.
Answer	A fool that often does something wrong.
Key	This dog does some foolish and wrong things wherever he goes.
Student	He is a clever dog that can do everything well.
Feedback	No, you can get hints from Pluto's behavior.
Auto-Ques	Now you've got it. and what is the meaning of the dog, Pluto, according to the passage?
Auto-Stud	He is a clever dog that can do everything well.
Auto-Key	This dog does some foolish and wrong things wherever he goes.
Auto-Feed	It's not. He is not a good friend. (score 0)

TABLE 10. Key sentence hint feedback example.

	Contents
Question	People in Beijing work ___.
Answer	Hard
Key	People in Beijing are in high spirits and hard-working.
Student	They work slowly.
Feedback	Think it over. The writer says they are in high spirits and they work in this way.
Auto-Ques	Exactly! And how do people in Beijing work?
Auto-Stud	They work slowly.
Auto-Key	People in Beijing are in high spirits and hard-working.
Auto-Feed	Not really. They are not hard-working. (score 0)

TABLE 11. General hint feedback example.

	Contents
Question	How many times has the writer been to Beijing?
Answer	Twice
Key	It's the second time for me to come to Beijing.
Student	He has been to Beijing once.
Feedback	No, please read the passage thoroughly.
Auto-Ques	Good to know! How many times has the writer been to Beijing?
Auto-Stud	He has been to Beijing once.
Auto-Key	It's the second time for me to come to Beijing.
Auto-Feed	No, he has been to Beijing twice. (score 1)

a short response to the previous student's answer, and the next new "Question" is then asked. The score of the "Auto-Feedback" ("Auto-Feed") is provided by a human rater.

We first consider reasoning hints. In the example in Table 8, to provide ground-truth "Feedback," the model should know that "the equipment becomes easier and safer" means "it's useful for people."

In the example in Table 9, the model should know that "doing something wrong" is a "behavior" to provide ground-truth "Feedback."

We also provide examples for key sentence hints. In the example in Table 10, "Feedback" helps student find the key sentence ("Key") by referring to the phrase "in high spirits."

By contrast, the general-type "Feedback" does not give specific hints. The example is given in Table 11.

REFERENCES

[1] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educ. Researcher*, vol. 13, no. 6, pp. 4–16, Jun. 1984, doi: 10.3102/0013189X013006004.

[2] C. A. T. Kegel and A. G. Bus, "Online tutoring as a pivotal quality of web-based early literacy programs," *J. Educ. Psychol.*, vol. 104, no. 1, pp. 182–192, Feb. 2012, doi: 10.1037/a0025849.

[3] C. G. Brinton, R. Rill, S. Ha, M. Chiang, R. Smith, and W. Ju, "Individualization for education at scale: MIIC design and preliminary evaluation," *IEEE Trans. Learn. Technol.*, vol. 8, no. 1, pp. 136–148, Jan. 2015, doi: 10.1109/TLT.2014.2370635.

[4] B. D. Nye, A. C. Graesser, and X. Hu, "AutoTutor and family: A review of 17 years of natural language tutoring," *Int. J. Artif. Intell. Educ.*, vol. 24, no. 4, pp. 427–469, Sep. 2014, doi: 10.1007/s40593-014-0029-5.

[5] V. Rus, N. Niraula, and R. Banjade, "DeepTutor: An effective, online intelligent tutoring system that promotes deep learning," in *Proc. 29th AAAI Conf Artif. Intell.*, Mar. 2015, vol. 29, no. 1, pp. 4294–4295, doi: 10.1609/aaai.v29i1.9269.

[6] I. Roll and R. Wylie, "Evolution and revolution in artificial intelligence in education," *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 582–599, Jun. 2016, doi: 10.1007/s40593-016-0110-3.

[7] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue," *IEEE Trans. Educ.*, vol. 48, no. 4, pp. 612–618, Nov. 2005, doi: 10.1109/TE.2005.856149.

[8] M. Ventura, M. Chang, P. Foltz, N. Mukhi, J. Yarbro, A. P. Salverda, J. Behrens, J. Ahn, T. Ma, T. I. Dhamecha, S. Marvaniya, P. Watson, C. D'helon, R. Tejwani, and S. Afzal, "Preliminary evaluations of a dialogue-based digital tutor," in *Proc. Int. Conf. Artif. Intell. Educ.*, London, U.K., Jun. 2018, pp. 480–483, doi: 10.1007/978-3-319-93846-2_90.

[9] J. Luk, "The dynamics of classroom small talk," *Issues Appl. Linguistics*, vol. 14, no. 2, pp. 115–132, 2004, doi: 10.5070/L4142005072.

[10] E. Kochmar, D. D. Vu, R. Belfer, V. Gupta, I. V. Serban, and J. Pineau, "Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems," *Int. J. Artif. Intell. Educ.*, vol. 32, no. 2, pp. 323–349, Jul. 2021, doi: 10.1007/s40593-021-00267-x.

[11] M. Grenander, R. Belfer, E. Kochmar, I. V. Servan, F. St-Hilaire, and J. C. K. Cheung, "Deep discourse analysis for generating personalized feedback in intelligent tutoring systems," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 17, pp. 15534–15544, doi: 10.1609/aaai.v35i17.17829.

[12] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding comprehension dataset from examinations," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 785–794, doi: 10.18653/v1/D17-1082.

[13] Y. Lee, "Improving end-to-end task-oriented dialog system with a simple auxiliary task," in *Proc. Findings Assoc. Comput. Linguistics, (EMNLP)*, 2021, pp. 1296–1303, doi: 10.18653/v1/2021.findings-emnlp.112.

[14] Y. Yang, Y. Li, and X. Quan, "UBAR: Towards fully end-to-end task-oriented dialog systems with GPT-2," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 16, pp. 14230–14238, doi: 10.1609/aaai.v35i16.17674.

[15] K. Stasaski, K. Kao, and M. A. Hearst, "CIMA: A large open access dialogue dataset for tutoring," in *Proc. 15th Workshop Innov. Use NLP Building Educ. Appl.*, Seattle, WA, USA, 2020, pp. 52–64, doi: 10.18653/v1/2020.bea-1.5.

[16] A. Caines, H. Yannakoudakis, H. Edmondson, H. Allen, P. Pérez-Paredes, B. Byrne, and P. Buttery, "The teacher-student chatroom corpus," in *Proc. 9th Workshop Natural Lang. Process. Comput. Assist. Lang. Learn. (NLPCALL)*, Nov. 2020, pp. 10–20, doi: 10.3384/ecp2017510.

[17] B. Settles, C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani, "Second language acquisition modeling," in *Proc. NAACL-HLT Workshop Innov. Use NLP Building Educ. Appl. (BEA)*, pp. 54–56, Jun. 2018.

[18] A. Filighera, S. Parihar, T. Steuer, T. Meuser, and S. Ochs, "Your answer is incorrect…Would you like to know why? Introducing a bilingual short answer feedback dataset," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland, May 2022, pp. 8577–8591, doi: 10.18653/v1/2022.acl-long.587.

[19] G. Chen, J. Yang, C. Hauff, and G.-J. Houben, "LearningQ: A large-scale dataset for educational question generation," in *Proc. Int. AAAI Conf. Web Social Media*, Jun. 2018, vol. 12, no. 1, pp. 481–490.

[20] X. Jia, W. Zhou, X. Sun, and Y. Wu, "EQG-RACE: Examination-type question generation," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 14, pp. 13143–13151, doi: 10.1609/aaai.v35i14.17553.
[21] D. Dzendzik, C. Vogel, and J. Foster, "English machine reading comprehension datasets: A survey," 2021, *arXiv:2101.10421*.
[22] Y. Liang, J. Li, and J. Yin, "A new multi-choice reading comprehension dataset for curriculum learning," in *Proc. 11th Asian Conf. Mach. Learn.*, Nagoya, Japan, Oct. 2019, pp. 742–757.
[23] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? Try ARC, the AI2 reasoning challenge," 2018, *arXiv:1803.05457*.
[24] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie, "DREAM: A challenge data set and models for dialogue-based reading comprehension," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 217–231, Nov. 2019.
[25] W. Yu, Z. Jiang, Y. Dong, and J. Feng, "ReClor: A reading comprehension dataset requiring logical reasoning," in *Proc. 8th Int. Conf. Learn. Represent.*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–26.
[26] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2204–2213, doi: 10.18653/v1/P18-1205.
[27] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," 2018, *arXiv:1811.01241*.
[28] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-grounded dialogue generation with pre-trained language models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 3377–3390, doi: 10.18653/v1/2020.emnlp-main.272.
[29] B. Kim, J. Ahn, and G. Kim, "Sequential latent knowledge selection for knowledge-grounded dialogue," 2020, *arXiv:2002.07510*.
[30] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher, "A simple language model for task-oriented dialogue," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 33, 2020, pp. 20179–20191.
[31] F. Qu, X. Jia, and Y. Wu, "Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2583–2593, doi: 10.18653/v1/2021.emnlp-main.202.
[32] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.
[33] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, Feb. 2019.
[35] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318, doi: 10.3115/1073083.1073135.
[36] B. Chen and C. Cherry, "A systematic comparison of smoothing techniques for sentence-level BLEU," in *Proc. 9th Workshop Stat. Mach. Transl.*, Baltimore, MD, USA, 2014, pp. 362–367, doi: 10.3115/v1/W14-3346.
[37] A. Lavie and M. J. Denkowski, "The meteor metric for automatic evaluation of machine translation," *Mach. Transl.*, vol. 23, nos. 2–3, pp. 105–115, Nov. 2009, doi: 10.1007/s10590-009-9059-4.
[38] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop ACL Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81.
[39] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, San Diego, CA, USA, Mar. 2016, pp. 110–119, doi: 10.18653/v1/n16-1014.
[40] S. Golovanov, A. Tselousov, R. Kurbanov, and S. I. Nikolenko, "Lost in conversation: A conversational agent base on the transformer and transfer learning," in *Proc. NIPS*, Nov. 2019, pp. 295–315, doi: 10.1007/978-3-030-29135-8_12.
[41] S. Kim, O.-W. Kwon, and H. Kim, "Knowledge-grounded chatbot based on dual Wasserstein generative adversarial networks with effective attention mechanisms," *Appl. Sci.*, vol. 10, no. 9, p. 3335, May 2020, doi: 10.3390/app10093335.
[42] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, Tech. Rep., Jun. 2018.
[43] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," 2020, *arXiv:2001.09977*.
[44] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gasic, "MultiWOZ—A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 5016–5026.

JIN-XIA HUANG received the B.S. degree in physics from Jilin University, China, in 1991, the M.S. degree in computer science from KAIST, Republic of Korea, in 2001, and the Ph.D. degree in computer science from Jeonbuk National University, Republic of Korea, in 2018.

From 1994 to 1997, she worked as an Engineer at the Yanbian University of Science and Technology (YUST), Yanji, China. From 2001 to 2003, she worked as a Researcher at Microsoft Research Asia, Beijing, China. Since 2008, she has been working with the Language Intelligent Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, where she is currently a Principal Researcher. Her research interests include natural language processing, dialogue systems, and intelligent tutoring.

YOHAN LEE received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, Republic of Korea, in 2015 and 2017, respectively. Since 2017, he has been working with the Language Intelligence Research Section, Artificial Intelligence Laboratory, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea. His research interests include natural language processing, machine translation, and machine learning.

OH-WOOG KWON received the B.S. degree in computer engineering from Kyungpook National University, Republic of Korea, in 1992, the M.S. degree in computer science from KAIST, Republic of South Korea, in 1995, and the Ph.D. degree in computer engineering from the Pohang University of Science and Technology (POSTECH), Republic of Korea, in 2001.

Since 2004, he has been working with the Language Intelligent Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea, where he is currently a Principal Researcher. His major research interests include natural language processing, dialogue systems, and text mining.