IEICE TRANS. INF. & SYST., VOL.E97–D, NO.7 JULY 2014

PAPER

POSTECH Immersive English Study (POMY): Dialog-Based Language Learning Game

Kyusong LEE[†], **Soo-ok KWEON**[††a)], **Sungjin LEE**[†††], **Hyungjong NOH**[†], ***Nonmembers***, ***and*** **Gary Geunbae LEE**[†], ***Member***

SUMMARY This study examines the dialog-based language learning game (DB-LLG) realized in a 3D environment built with game contents. We designed the DB-LLG to communicate with users who can conduct interactive conversations with game characters in various immersive environments. From the pilot test, we found that several technologies were identified as essential in the construction of the DB-LLG such as dialog management, hint generation, and grammar error detection and feedback. We describe the technical details of our system POSTECH immersive English study (Pomy). We evaluated the performance of each technology using a simulator and field tests with users.

key words: *dialog, game, education, grammar, virtual environment*

1. Introduction

Spoken dialog systems have been developed for information-seeking tasks such as car navigation, restaurant recommendations, telephone service, and weather information. Chat-oriented dialog systems have also been developed for research and commercial purposes such as ALICE [1], and ELIZA [2]. These have been developed to handle non-task-related utterances. Recently, various other applications of dialog systems are appearing in ongoing projects for research and commercial products. For example, chat-bot systems have been developed for education and teaching [3], [4]. We have developed a spoken dialog system for second language (L2) learning.

According to Input theory [5] and Interaction theory [6] in SLA, online chatting in L2 learning settings is adequate for an educational system, as it seems to promote communication competence through lively exchanges and enhance reflective and meta-cognitive communication in L2 learning settings. To become an efficient teaching system, the system should provide educational contents combined with interesting activities through negotiated input, corrective feedback, and modified output. Through these processes learners can learn form and meaning which they are exposed to, because L2 learning is impossible without conscious awareness or attention to the input according to [7], [8] noticing hypothesis. This implies that in foreign language learning, raising learners' consciousness of meaningful comprehensible input is important. Because the process of L2 learning is fundamentally different from that of L1 learning [9] and L2 cannot be acquired without conscious attention to meaningful input. Raising consciousness or attention to meaningful input may be essential to successful L2 learning, and many researchers have investigated the importance of giving corrective feedback to facilitate the learning process in L2 learning [10]–[14].

In the study of second language learning, generating a willingness to communicate in L2 is arguably one of the central objectives of L2 pedagogy. According to [15], the implication of the willingness to communicate can be integrated into motivation of L2 learners and their use of language learning strategies. Learning strategies, which is referred to as "the conscious thought and actions that learners take in order to achieve a learning goal" [16], if well-used, can compose a significant proportion of motivated learning behavior. Also, using learning strategies can make learning quicker, easier and more effective [17]. In this respect, the function of the main components of the DB-LLG system used in the present study, such as hint generation, and grammar error detection and feedback can enhance the use of various learning strategies in order to achieve communication competence in language learning.

In this study, the methods and technology are proposed for a dialog-based language learning system using a spoken dialog system in a virtual environment. Additionally, several necessary components for the dialog-based language learning system, such as hint generation and grammatical error feedback, are presented. Hint generation, in particular, which was created by the system based on the dialogue corpus, is a unique way of keeping learners' motivation in speaking English when they have difficulty with appropriate words or expressions during the game. Through hints generated by the Tutor character in the system, learners could negotiate meaning to make input comprehensible.

CMC (Computer Mediated Communication) is conducted in a real time interaction in which users negotiate both form and meaning through modified input, output and feedback by using a keyboard. [18], [19], among others, conducted CMC research in which L2 learners interact either with each other or with native speakers of the target language and found that learners employ various kinds of

Manuscript received October 30, 2013.
Manuscript revised March 7, 2014.

[†]The authors are with the Department of Computer Science and Engineering, Pohang University of Science and Technology, Korea.
[††]The author is with Division of Humanities and Social Sciences, Pohang University of Science and Technology, Korea.
[†††]The author is with Language Technologies Institute, Carnegie Mellon University, USA.
a) E-mail: soook@postech.ac.kr (Corresponding author)

DOI: 10.1587/transinf.E97.D.1830

strategies to negotiate meaning and form through the use of modified input and output. It is widely accepted that CMC can enhance communication competence of language learners by providing input and feedback and bridging the gap between speaking and writing via computer technologies [20]–[23]. In the present study, we followed the basic premise of CMC with slight modification, in which Korean elementary students chatted through speaking with the system instead of typing a keyboard. The students negotieated both meaning and form using spoken-dialog technology that enables the system to speak with them instead of involving a human interlocutor.

Therefore, the primary goal of this study is to describe how the computer system can possibly function as much as a human interlocutor can do and how the system can facilitate the processes of English learning by the young learners with low level of proficiency.

Previous research in SLA has primarily focused on face-to-face interaction between native speakers and L2 learners [24], [25]. Especially, the verbal interactions between interlocutors can improve communication competence by avoiding and repairing impasses in conversation at the syntactic, lexical and phonological level of discourse structure [26]. In this respect, a native-speaker teacher would perceivably be the best resource to improve speaking ability in conversational interactions in the sense that he can provide both positive and negative evidence in timely and appropriate occasions.

However, due to the high cost and limited number of native English-speaking teachers, among other factors, many learners of English as a foreign language (EFL) have limited opportunities to practice English in a natural language learning setting (i.e., low motivation, insufficient input, interference with affective filter) Therefore, despite spending considerable time and energy learning English, many Korean students still find communicating in English difficult. These reasons have prompted considerable interest in research on English education in Korea. Robot learning and e-learning have been proposed to examine future second language acquisition [27]. Our research group is examining more interesting, motivating, economical and efficient ways to learn English. Game-based learning is currently being evaluated as an educational method because it garners high user motivation, attention, and interest [28]–[30]. Moreover, the concept of social community is gaining popularity, such that many people are spending time interacting in a virtual environment with their own characters. Recently, as an alternative to studying abroad in English speaking counties, some people studied with native speakers in virtual environments such as Second Life [31]. These developments led us to create a spoken dialog-based language learning game (DB-LLG), called POSTECH immersive English study (Pomy), in which users use the senses of sight, hearing and touch to receive a full-immersion experience. Thus, users can develop into independent EFL learners while increasing their memory and concentration abilities. The system employs spoken-dialog technology that enables computers to speak with humans. One advantage of the DB-LLG approach is that the learner becomes more actively engaged with the tutor in the game rather than with a teacher in a class. The ultimate goal is to give students the opportunity to speak English in an interactive environment comparable to on-line games. However, in pilot tests, students often did not know the proper response when interacting with Non-Player Characters (NPC). Therefore, we developed several hint generation methods to help students maintain conversations with the NPCs. In this paper, we will introduce a ranking-based dialog system with integrated hint generation that keeps students highly involved in the DB-LLG in a 3D environment built with game contents.

2. Related Work

Conversational agents in a virtual world have been developed by Deutsches Forschungszentrum fur Kunstliche Intelligenz (DFKI). Their group investigated conversational agents capable of reasoning and inference using a knowledge-based approach of semantic web technology with ontologies. The Institute for Creative Technologies (ICT) developed the Mission Rehearsal Exercise System, which is designed to teach critical decision-making skills to small-unit leaders in the U.S. Army [32]. Furthermore, the Natural Interactive Communication for Edutainment project developed a fairy-tale game that uses spoken conversation and gestures through a spoken dialog system [33]. The aforementioned work was developed for various purposes, including chatting, military training, and gaming. Our research group focuses on education using a dialog system in virtual environments. Several systems have been developed for language teaching and learning in interactive environments. The tactical language and culture training system is one of the most successful systems developed to date. It targets members of the U.S. military who need to acquire basic communication skills in Arabic and knowledge of cultural differences in a given zone of operations, such as Iraq [34]. Spoken electronic language learning (SPELL) [35] provides opportunities for learning languages in functional situations such as going to a restaurant and expressing likes and dislikes. The key development of SPELL is the ability to recognize grammatical errors, especially those errors made by non-native speakers. Recast feedback is provided if the learner's response is semantically correct but includes grammatical errors. This system combines semantic interpretation and error checking in the speech recognition process. Thus, it uses a special speech recognition feature to identify and respond to both correct and erroneous speech. DEAL, a spoken dialog system developed at Kungliga Tekniska högskolan, Royal Institute of Technology (KTH), focuses on creating entertaining gameplay [36]. DEAL uses the trade domain, specifically a flea market, and provides hints about things the user might try to say if he or she is having difficulty remembering names of items or if the conversation has stalled for other reasons. One contribution of our current study is that hint generation was essential to enable

L2 learners to maintain continuous dialog with the system. We employ three methods which we categorized as follows: proper sentence choice questions, error identification questions, and keyword presentation. We would like to mention that the systems introduced in the present study, such as the Dialog Management (DM) system, including Hint Generation, and the Grammar Error Detection system have been developed originally by our team for the last decade. Among various systems developed by our team, we prioritized educational systems, e.g., DM and the Grammar Error Detection system. In particular, DM, which was developed for language learning by L2 learners, focuses on providing a flood of input to increase learners' competence in communication. For this purpose, the Hint Generation module was tailor-made by modifying the key functions of DM. Therefore, in this study we established a method of Hint Generation uniquely applied to DM. Moreover, to the best of our knowledge, our study is the first attempt to employ DM, Grammar Error Detection, and a Hint Generation system in one study. Finally, we evaluated speaking improvement, such as fluency and correctness through field testing. To assess fluency, we compared the number of words spoken by students pre-test and post-test. To explore correctness, we tagged the number of grammatical errors in pre-test and post-test speech. This paper introduces POMY system, accounting for its technical details and application for the field test.

3. POSTECH Immersive English Study

This spoken dialog system makes use of several technologies: automatic speech recognition (ASR), spoken language understanding (SLU), dialog management (DM), natural language generation (NLG), text-to-speech (TTS), hint generation (HG), grammar error detection and feedback (GED and F) (Fig. 1). ASR is the first step of the Dialog System to translate noisy speech input into text data. If a user said "Where is the Korean restaurant", while some background noise was present, the ASR might output "Um where Korean restaurant" from the speech. The purpose of the SLU is to determine the user's intention from the spoken utterance input. The intentions consist of 3 parts": the dialog act (DA), the main goal (MG), and the named entities (NE). The DA is a domain-independent label of an utterance at the level of illocutionary force (e.g., STATEMENT, REQUEST, or WH-QUESITON). The MG indicates the domain-specific user goal of an utterance (e.g., GUIDE−LOC, SEARCH−LOC, or SEARCH−PHONE). The provided utterance text such as "Um where Korean restaurant?" is changed to [DA=WH−QUESTION], [MG=SEARCH−LOC], and [NE=(LOC−Type=Korean restaurant)], which the computer can understand [37]. The role of the DM is to generate system responses according to the learner's intention and to generate corrective feedback accordingly. The DM ensures that proper system actions can be mapped from a user intention (i.e., the output of the SLU). The results of the DM can be used for generating various system utterance expressions for education and HG. The NLG functions by outputting natural language (e.g., "Korean Restaurant is located at 2nd street.") from the result of the DM, which is a machine-friendly code (e.g., Inform (loc-name, loc-address)). Finally, users can listen to an audio speaker by TTS. In previous dialog systems, such conversations between human and machine are similar to the example shown in Table 1. However, in a language learning dialog system, there are many cases wherein the user's sentences contain many grammatical errors. Moreover, users do not know what to say or how to say it in a given situation. Therefore, our system has different conversation patterns such as grammatical detection and feedback, so that the system can suggest hints, such as question types. It is important that a language learning dialog system provide these functions to help the learner improve their English. In grammar error detection and feedback, two step approaches, such as restate and correction, are conducted during the conversation. Local errors are defined as grammatical errors that are relatively small, such as inflection, derivation, preposition choice, article usages. These errors can be corrected by changing a few words. Global errors are those that require that the sentences be completely changed. Grammar learning is conducted during conversation in the following way (Table 2). If a grammatical error is detected, the tutor restates the part where the user made a mistake at the first

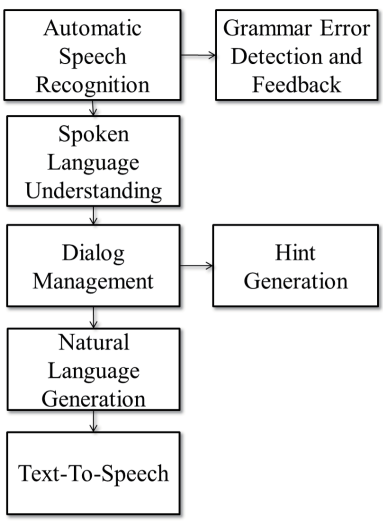


Fig. 1 Outline of pomy dialog system.

Table 1 Conversation about path-finding.

Speaker	Utterance
User	Excuse me, can you tell me where a market is?
System	You want to go to Happy Market?
User	Yes. Can you let me know how get there?
System	Just turn left at the bank, and walk along the street for three blocks. It's next to the police station.
User	How far is it?
Sytem	It is about one mile.
User	Thank you.
System	You're welcome.

Table 2 Conversation with grammar error detection and feedback.

Speaker	Utterance	Feedback
User	Excuse me, can you tell me where a market is?	
System	You want to go to Happy Market?	
User	Yes. Can you let me know how get there?	
Tutor	How get there	**DETECTION**
User	Yes. Can you let me know how get there?	
Tutor	How to get there	**CORRECTION**
User	Yes. Can you let me know how **to** get there?	
System	Just turn left at the bank, and walk along the street for three blocks. It's next to the police station.	
User	How far is it?	
System	It is about one mile.	
User	Thank you.	
System	You're welcome.	

Table 3 Conversation with hint generation.

Speaker	Utterance
User	Excuse me, can you tell me where a market is?
System	You want to go to Happy Market?
User	Uh-huh. Help me.
Tuter	**SPEAK MOST PROPER SENTENCE IN THIS SITUATION** 1. Yes I do. Can you let me know how to get there? 2. Happy market is on your right side. 3. You can miss it. 4. I want to buy apple.
User	How far is it?
System	It is about one mile.
User	Thank you.
System	You're welcome.

time. This first step just highlights the errors to the student, providing the student with an opportunity to correct the errors on their own. In the second step, if the user fails to correct the errors, the tutor corrects the user's errors. Hints are presented if the student needs them. Initially, the student can receive a hint by uttering a statement such as, "help me" or "I need help". Otherwise, when global feedback is needed for a learner, hint is provided using a question format. The data from the pilot test show that when providing the answer to the student directly, student easily become uninterested and disengaged. Moreover, students tend to rely too much on tutor without thinking on their own during conversation. Therefore, we developed the hint generation, which forces the student to talk with the system. In addition, if they do not follow the conversation, it is impossible to choose the best hint from among the other hints. An example conversation, including hints, is shown in Table 3. The system HG methods (e.g., proper sentence choice question) are discussed in subsequent sections. In this paper, we focus on the DM, the GED and F, and the HG which are the main contributions of our educational dialog system.

3.1 Scenario

The domains selected for the students were path-finding, market, post office, library, and movie theater; these domains were selected to ensure that the students practice conversations in everyday life setting (Fig. 2). To date, five missions have been developed in the game and each mission consists of three main tasks that include pre- and post-courses. Between missions, a path-finding pre-task is implemented. Before the start of the game, the students should be supposed to be familiar with the mission objective and the particular tasks; then, they are introduced to important vocabulary and useful expressions during the pre-course. For example, students should understand the meaning of some key words, including "zip code", "insure", and "over-night letter" and utter some key sentential expressions, such as "How much does it cost to send a package?" to successfully accomplish the missions in the post office. These lessons were conducted during the pre-course before the main mission was introduced. The first mission that occurred in the post office was to send a camera to an uncle in England. The package must be insured and delivered by next week. To send the package, the student must fill in the zip-code properly. After completing each task, students were asked to review what they had learned in the previous task. Then, some comprehension questions were given to check whether they fully understood the content (e.g., "If the insured package is lost, what will happen?"). To enhance the learning of essential vocabulary and expressions, students were exposed to the same words and expressions repeatedly across three stages: first, during preview stage, followed by the main task mission in virtual-reality situation, which is, in turn, followed by review. By repeating important expressions through different learning stages, the effectiveness of language learning could be maximized [24].

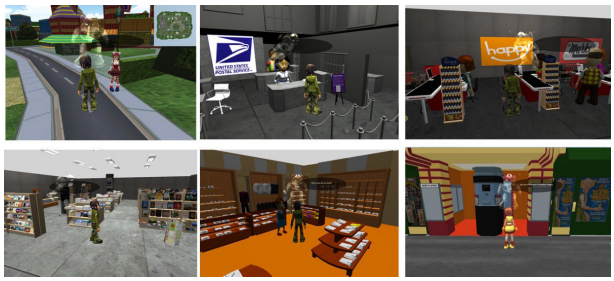


Fig. 2 Screenshot of the system.

3.2 The Tutor Character

The Tutor is a key character in the game that helps the students in various situations when they encounter difficulties. The Tutor guides them to move towards the next step throughout all the procedures of the game so that the students can complete the mission successfully. The Tutor plays several special roles. First, the Tutor cast himself in the role of an English tutor, as a player, helping the students to use more appropriate words and expressions during the game. When a student produces ungrammatical utterances, or has difficulty in speaking, the Tutor provides both implicit and explicit negative and positive feedback in a

form of recasting, which is manifested effectively in the second language acquisition processes [6]. The Tutor provided full sentences for fragmental responses of the students. Although a student may speak without errors, the Tutor sometimes gives alternative examples of expressions to help him learn various new forms. Second, the Tutor is the guide to the game. Although the explanations of the mission goals is clearly provided in the pre-course section, some students cannot remember them during the game. If this happens, the Tutor reminds them of the previous mission objective by going back to that particular objective. Third, the Tutor is an intimate partner of the student. Whenever a student speaks to the Tutor, he provides pleasing responses. Additionally, during a long interval without students' speaking, the Tutor strikes up a casual conversation unrelated to the mission (e.g., "How is the weather today?"), which gives students a chance for bonus credits during the game, and helps them practice small talk conversational skills.

4. Technical Details

4.1 Dialog Management

A ranking-based algorithm, based on the example-based dialog system [38], is used for DM. When the system utterance is chosen, the most probable user utterances are ranked based on a dialog corpus collected in advance. Likewise when a user utterance is input, the most suitable system responses are ranked. The Levenshtein distance is typically used as a string metric for measuring the difference between two sequences. This method is adopted to compare the current dialog flow (Fig. 3-(b)) with reference dialog flows in the corpus (Fig. 3-(a)) one-by-one. The method can also consider the order of nodes, so dialog history can be taken into account. High perplexity is equated with low importance and low information in the flow of discourse: i.e., a user will have many possible responses after the system gives a perplexing utterance such as "hello", but few answers will be available if the system asks a specific question like "how much is it?". The intention of "ask/price" is more important than "state/greeting" in the discourse. Therefore, the weight of all defined intentions is estimated in advance using the dialog corpus to calculate the distance. Details of the dialog-ranking procedure using the modified Levenshtein distance algorithm are illustrated in Fig. 2. The dialog history corpus with intention weights are stored in a database (Fig. 3-(a)). The current dialog flow is then compared with the dialog history corpus in run time (Fig. 3-(b)). Distances between the dialog history corpus and the current dialog are computed using the modified Levenshtein algorithm. Finally, the results are ranked according to distance score with higher scored sequences corresponding to more likely real dialogs (Fig. 3-(c)). The results of the ranking can be used for hint generation and generating various system utterance expressions for education. Details of the hint generation will be explained in the next section. The two sequences (current flow and dialog flow in corpus) do not need to have the same length, so a single sequence can be matched with many similar flows to establish a more reliable ranking. The original Levenshtein distance [39] assigns the insertion, deletion, and substitution cost as the same value (=1); however, to reflect the characteristics of discourse history, we modified these costs in the proposed system. The value is adjusted based on the perplexity of the intention, given the possible subsequent sequences available. We believe that educational DM should be able to generate diverse responses to teach various useful English expressions. Unlike information-seeking dialog management that generates only the 1-best system response, this educational system generates n-best system responses by considering the discourse history. An additional feature is used to rank the intentions, such as the ordinal position or the entity constraint feature. The ordinal position feature has a different focus than the discourse similarity feature. The discourse similarity is measured by comparing two sequences of speech acts; it considers the similarity between two distinct dialog sequences, the current dialog flow and the candidate dialog instance. Conversely, the ordinal position feature reflects the relationship between two speech acts in the current dialog flow. This feature measures the appropriateness of the order of the speech acts in the current dialog state regardless of the other dialog example. Our system focus on the causal relationships among speech acts. For example, a stranger first gets the resident's attention, and then requests the path to the destination. After learning how to get there, he asks for the distance and thanks the resident for the help. If the stranger asks for the distance before the destination is confirmed or he thanks the resident before he obtains any information about the destination from the resident, it would not be a proper dialog sequence. Thus, the ordinal relations among speech acts can be useful for predicting the next system action or for verifying the current user input. Another consideration is the relative position of each speech act in the entire dialog. In the training corpus, we have observed that two specific speech acts tend to be separated by specific intervals. For example, the system action of instructing the path to the destination is often followed immediately by the user dialog act of expressing thanks, whereas the user dialog act of getting the resident's attention tends to be separated by a significant interval from the user dialog act of expressing thanks in many conversations. The algorithm detail for the dialog management have been previously published [40].

4.2 Hint Generation

In the pilot tests, students often had difficulty with providing proper responses expected to proceed with the game successfully, but they felt bored and did not want to continue playing the game if the answers were given directly. Therefore, instead of revealing the answers, hints were provided to enable the students to speak properly on their own. The ranking-based DM made it possible to generate hints based on the most probable user utterances. N-best results for both user and system utterances are extracted by

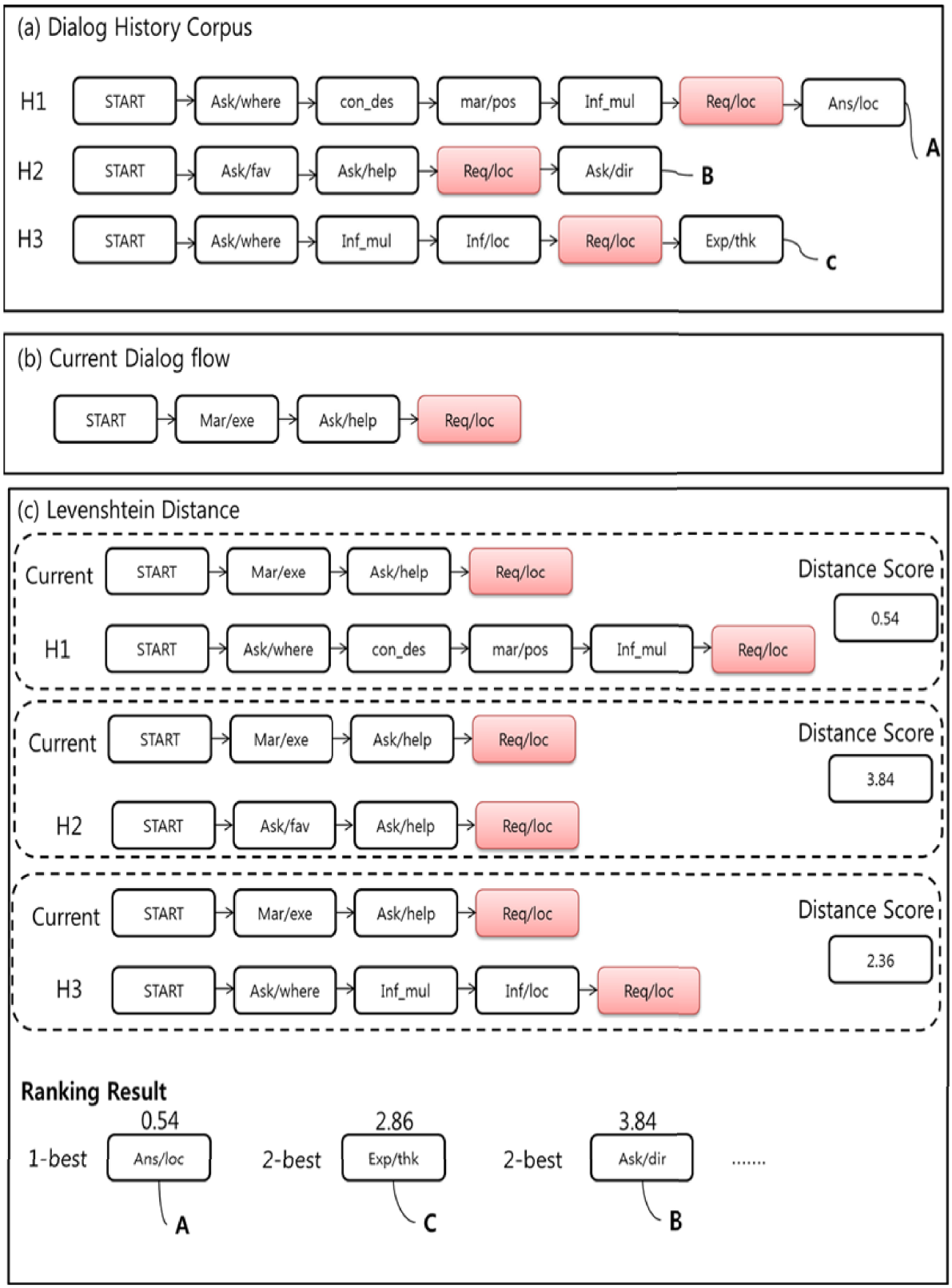


Fig. 3 Overall process of dialog management.

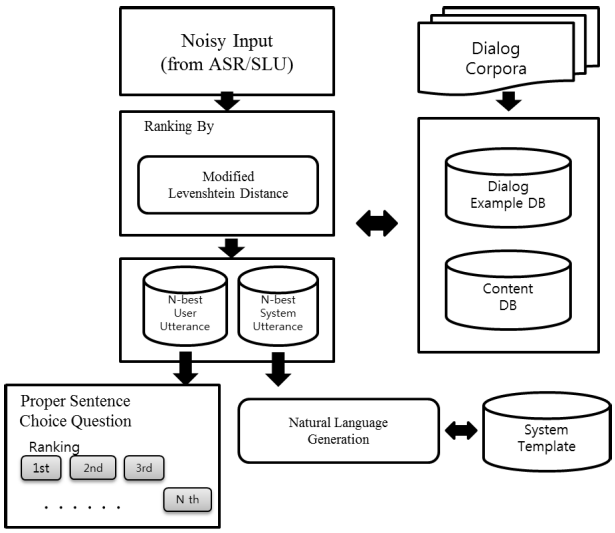
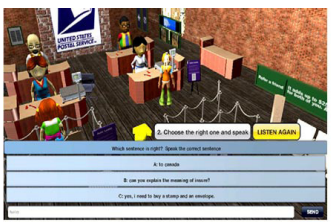


Fig. 4 Process of hint generation.



a proper sentence choice question

A: to Canada

B: Can you explain the meaning of insure?

C: yes, I need to buy a stamp and an envelope

Fig. 5 Screenshot of hint generation.

the ranking-based DM. N-best results of the user utterances are used for HG (Fig. 4). The proper sentence choice questions were provided using the ranking-based DM. The highest ranked result was the most proper answer in the given situation, and the lowest ranked utterance would be an inappropriate utterance. Therefore, one contextually proper answer and several inappropriate utterances are offered to students upon their request when they hesitate to speak or need to be helped. The given instruction is "Which sentence is right? Speak the correct one", and students can choose the right answer and speak. Figure 5 demonstrates a sample of Hint Generation (HG) during game situation taking place in the post office. Generating hints is very important in language learning dialog systems for the following reasons: Language learning dialog systems are more similar to chat-oriented dialog systems rather than to information seeking dialog system. Chat-oriented dialog systems have lower performance than information-seeking dialog systems because the perplexity of the dialog is higher in chat-oriented dialog systems, because the domain is already decided in the information-seeking dialog systems so that domain selection is not an issue in this system. However, as chat-oriented dialog system does not have a specific domain, there are many cases that the user's utterance is not present in the training corpus. Therefore, having a back off strategy is important in a chat system. It is the case that chat-bots in current technology often respond with answers that are not related to the user's state. Having the most proper answer is important not only in a chat system, but in an educational dialog system. Therefore, instead of giving back off utterances when receiving unexpected utterances from users, the present system guides them to speak a correct utterance that can fit the system.

	Grammatical Error Detection	I	am	here	at	business
1)	Grammaticality Checking	0	0	0	1	0
2)	Error Type Classification	None	None	None	PRP_LXC	None

Fig. 6 Grammar Error Detection (PRP_LXC indicates a preposition lexical error).

4.3 Grammar Error Detection and Feedback

We developed a grammar error detection module to implement the Tutor who provides feedback to the user regarding ungrammatical errors in his or her utterance. It is not a trivial task to detect grammatical errors in oral conversations because of the unavoidable errors of the ASR system. To date, few studied have been conducted on grammar error detection of spoken dialog. SPELL and DISCO detect grammatical errors using finite state network (FSN) recognition grammars that include both correct and incorrect responses in ASR. This approach has its limitation. However, recognition performance is low because the number of grammar instances increases exponentially, and it is impossible to design every ungrammatical response. Moreover, previous research only considers the 1-best result to detect the grammatical error on learners' speech. Therefore, we investigate a method to use a confusion network (CN) [41] to consider multiple hypotheses based on confidence scores. Another reason that predicting grammatical errors is difficult is that there are more grammatical words than ungrammatical errors in the data. Therefore, imbalanced data distribution must be considered when applying a machine learning technique. When accuracy is the performance measure, using the classifier trained on the highly imbalanced data simply produces the majority class for all test data to achieve the best performance. In addition, the number of error types to classify is relatively large, which can make the model learning and selection procedure complicated. Therefore, to cope with these difficulties, the grammatical error detection model is divided into two sub-models: the grammaticality-checking model and the error-type classification model (Fig. 6). The technology of grammar error detection and feedback will be presented herein. Details about the grammar error detection and feedback module have been previously published in [42], [43].

4.3.1 Grammaticality Checking Model

To detect the grammaticality, we first extract error patterns from the simulated ungrammatical responses using a gram-

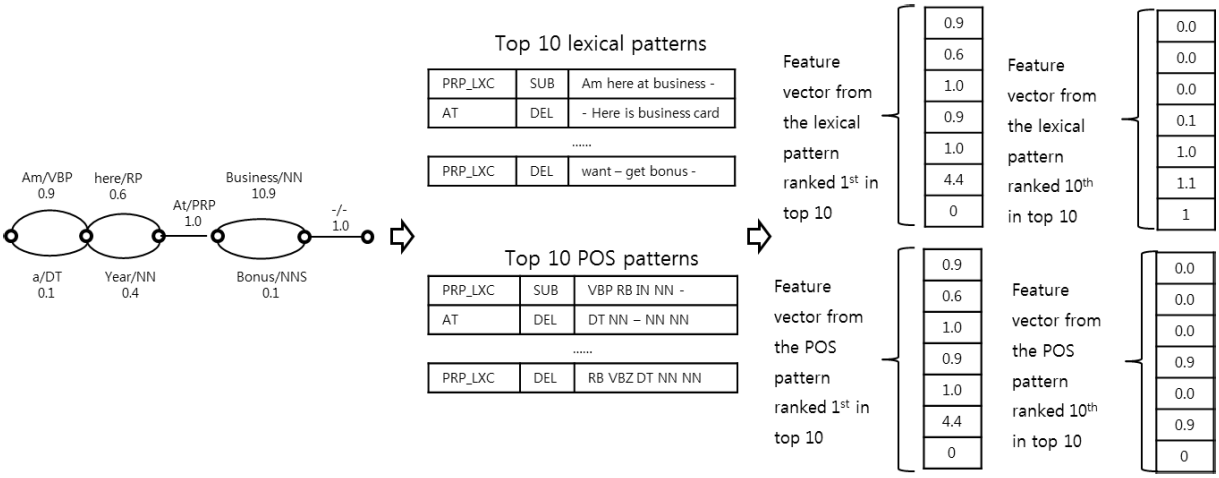


Fig. 7 Feature extraction process.

mar error simulation [44]. The error pattern is the 5-tuple, which consists of the erroneous word and its two left and two right neighbor words. For example, the error pattern for the preposition error at 'at' for the utterance 'I am here at business' will be the 5-tuple ['am', 'here', 'at', 'business', '-']. The error pattern is also tagged with the error type and structural deviation (e.g., deletion or substitution) for the error-type classification task. When a speech is recognized, at each position in the CN, a feature vector is extracted by comparing the error patterns with the segment of the CN consisting of the target position and the two left and right neighboring positions. Seven features are extracted for each error pattern, such as confidence score of the hypothesis matching the first, the second, the third, the fourth, and the fifth word in the error pattern, total score (TS), and indicator of structural error type (1 for deletion and 0 for substitution). For example, if the first word in the error pattern exists among the competing word hypotheses at the first position in the CN, then the confidence score of the matched word hypothesis is used as the feature. If no matched word hypothesis is used the feature is simply set to zero. The higher the matching scores an error pattern has, the more likely the recognized result has the relevant error in it. Because the number of error patterns is very large and likely uninformative, only the features extracted from the top 10 error patterns ranked by the TS feature are used. A similar feature extraction process is performed at the part-of-speech (POS) level. POS tagging is applied to both the recognition result and the error patterns to get additional features from the top 10 POS-level error patterns. The POS-level features contribute to raising the recall rate by alleviating the data sparseness problem of lexical-level features. Figure 7 depicts the aforementioned feature extraction process. The LIBSVM [45] Support Vector Machine classifier is used to produce a model that predicts grammaticality. A radial basis function (RBF) is used as the kernel because unlike linear kernels, an RBF kernel can handle nonlinear interactions between attributes and relationships between class labels and attributes.

4.3.2 Error Type Classification Model

To provide the feedback, the error type must be identified. Error-type classification is performed for the words that are determined ungrammatical by the grammaticality-checking model. The simplest way to classify the error type is to choose the error type associated with the top ranked error pattern. To break tied error patterns, error frequency is considered. Error patterns are reordered according to the equation as follows:

$$\text{Score(e)} = \text{TS(e)} + \alpha * \text{EF(e)}$$

where TS is the TS feature of the error pattern e and EF (error frequency). The constant α is 0.1 for this study.

5. Experiments and Result

In this section, we describe the student reactions in the field tests and design principle for DB-LLG. We also present each performances of our system such as dialog management and grammar error detection. We want to show that the performance of our methods is proper for language learning dialog systems.

5.1 Dialog Management

In POMY, users have to ask directions in a path-finding mission by talking with Non player characters. Learners can initiate any dialog they desire. These types of dialog are different from traditional task-oriented dialog systems. The conversation includes many colloquial utterances that are not directly related to task completion. For example, utterances such as "Calm down", "Don't worry", "Great!" and "Yes, I do" are relatively colloquial in style and make the conversation more natural. The utterances are relatively diverse in the POMY system. Our goal was to achieve the following characteristics in dialog management:

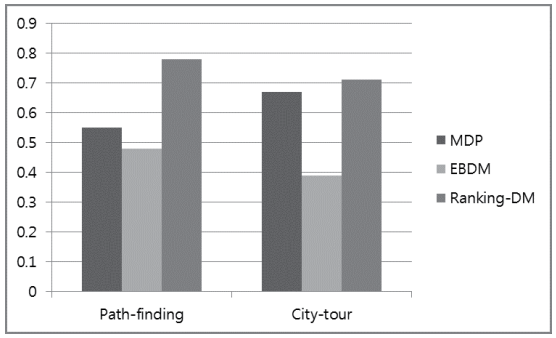


Fig. 8 Task completion rate in the path-finding and city-tour domain.

- The dialog participant is a novice user
- The dialogs include colloquial small talk as well as task-oriented conversations
- The user's utterances can include varied speech acts and expressions
- The sequences of speech acts in the dialogs are relatively diverse
- Whereas a user who wants a specific service and a service provider (or agent) participates in a traditional task-oriented dialog, a chat-like dialog occurs between persons

However, previous dialog systems, such as example dialog system (EBDM) and Markov Decision Process (MDP), have encountered difficulties in responding to such complex dialog demands, because they were developed to provide a specific goal, such as explaining the bus schedule or asking for phone numbers. The number of dialog acts in a task oriented dialog system is around 10; however, the number of dialog acts in our DM can cover more than 40 to 50 in each domain. The dialogs we used in POMY, i.e., Path-finding and City-tour dialogs, are neither task-oriented nor chat-oriented dialogs. We found that EBDM and MDP have a lower task completion rate than our ranking-DM (Fig. 8). To evaluate the performance of the DM, we compared it with MPD system that was implemented based on MDP policy [46]. It was trained with dynamic algorithm [47], one of the reinforcement learning algorithms. The dialog states are defined as combinations of the last user speech act and the slot-filling status. To select the next best system action, reward function is defined according to the transaction probability. Every transaction that appears in the training corpus received a positive score, whereas any transaction that does not appear in the training corpus is received with a negative score. The system is also compared with the EBDM system. To conduct an automated evaluation of the spoken dialog system, we implemented a user simulator that can simulate a plausible utterance when a dialog state is given. Using the user simulator avoids the evaluation problems that can arise with human subjects and experiments can be conducted effectively under various different conditions without changing other control variables. The user simulator was implemented based on the previous work [48] and includes user intention modeling using a linear-chain conditional random field (CRF), data-driven user utterance simulations, and ASR channel simulations which uses linguistic knowledge. When a dialog state is given, the user simulator generates a user intention based on the CRF model. The CRF model is trained using dialog history information from the training corpus. From the selected user intention, the corresponding user utterance is generated statistically according to the characteristics of the training corpus.

Table 4 Experimental result on the grammaticality-checking task.

Model	Precision	Recall	F-score	False Positive Rate
FSN	19.30	18.60	18.94	6.25
EPM	**97.44**	19.64	32.69	**0.04**
Proposed	91.82	**63.82**	**75.30**	0.46

5.2 Grammar Error Detection and Feedback

The current version of the error tagset targets morphological, grammatical, and lexical errors and can describe diverse grammatical errors. The error tagset currently includes 46 tags. The full list of error types are explained in [49]. Lexis errors related to open-word class (i.e., noun lexical, verb lexical, adjective lexical, and adverb lexical), were excluded in this experiment because realizing such errors without encountering the data sparseness problem requires a huge amount of learner data. Some other errors (i.e., collocation, misordering of words, unknown type errors, unintelligible utterance) were also excluded because these error categories have not yet been clearly analyzed for practical applications. Error categories that occurred less than five times were also excluded to improve reliability. This results in a total of 23 error types. Korean and Japanese speakers learning English have very similar error characteristics because the two languages are grammatically similar. In addition, we did not explicitly generate insertion errors, because many insertion errors appear implicitly as replacement errors in the NICT Japanese Learner English corpus. The insertion errors, which are not covered in this model, usually related to vocabulary of open-word class or are highly unpredictable even when linguistic context is taken into account. The experimental results showed that the proposed model largely outperformed the baseline FSN model for all metrics (Table 4). This is because the FSN-based Viterbi-decoding exhibits a very low sentence-level recognition performance because of the relatively large size of the recognition grammar consisting of many similar variants for various grammatical errors. This large size of recognition grammar affects not only the precision and recall but also the false-positive rate, which can be detrimental for language tutoring because it may frustrate learners about wrong instructions. The proposed method also surpasses the exact pattern matching (EPM) model in F-scores, which is attributed to the large gain in the recall rate. The proposed method achieves a far higher recall rate than that of the EPM model by exploiting a soft pattern match based on the confidence score. Furthermore, the proposed method lost little precision from the SVM model optimization used to satisfy the

constraints on the precision and false positive rate. Both the EPM model and proposed model showed a very low false positive rate. This finding implies that the proposed method is very suitable for educational applications. For the error-type classification task, the baseline method that does not consider the error frequency showed an accuracy of 95.55%. The proposed method improved the baseline performance by 4.05%. The baseline model is fairly accurate, but the incorporation of error frequency into the model improves the accuracy.

5.3 Field Test

5.3.1 Field Test Setting

The subjects in our study were 25 (10 male and 15 female) elementary school students in Korea who were recruited by the school's teachers. All subjects were either in the 5th or 6th grade (i.e., 11 or 12 years old), and all of them had been born and raised in Korea, speaking Korean as their native language. They began learning English as a foreign language at the age of seven or eight, and none of them had lived in an English speaking country for more than six months. Because no standardized test score was available at the time of test, we divided the students into two proficiency groups based on two sources: their academic grade in their English classes in school and the oral interview conducted with the researchers of the present study at the time of the experiment. Our students used English only in a classroom setting and the system was used during the class sessions, which were held for 45 minutes, three times a week for four weeks. Students received stationery or a gift certificate equivalent to $5.00 as rewards. The class was conducted in a computer room in which each student had access to a laptop computer and a headset that was used to record students' utterances. The scenario used in the present experiment contained four situations: Happy Market, Post Office, Bluebird Library, and Jina's House, each of which was covered over three class sessions.

5.3.2 Design Principles

Several pilot tests were performed using elementary school students. Subsequently, we analyzed the students' behaviors, reactions, and answers from video and log data recorded during the tests. The following design principles for DB-LLG were developed based on students' oral production, behavior and interviews after the game-like experiments will be described. Principle 1: Familiarization. It was important to make students familiar with interactive games. Some students failed to interact with the computer in speech modality. Many elementary school students already play many computer games at home using mainly the keyboard and the mouse, hence during the experiments, students were focusing only on the keyboard and the mouse to control their game characters, without listening and speaking. Therefore, we developed voice commands to control the user's avatar (e.g., "Go straight", "Turn left", and "Turn right") to make learners familiar with the speech modality. Principle 2: Educational purpose. It is important to avoid making a game that students play only for enjoyment, neglecting the educational purpose. Although they strongly wanted to accomplish the missions with a high success rate, students sometimes simply looked for shortcuts to finish the missions without having a conversation with the computer. For example, they just said "Yes" even though they did not understand the question exactly. The game will be designed to prevent this problem by checking their understanding strictly. Principle 3: Transcription of speech. It is important to vary the usage of the on-screen speech bubbles according to the students' proficiency level. For lower level students, the speech bubbles appear more frequently and for a longer duration to assist their understanding of the spoken dialog. However, we found them to be distracting for high level students who wanted to concentrate on listening. Hence, the speech bubbles are provided only when students have problems with understanding the system utterances.

5.3.3 Field Test Result

Some students were fluent in English, so they preferred to speak without viewing the hint. However, errors of automatic speech recognition cannot be avoided. Moreover, the current technology of DM cannot cover every utterance spoken by the user. Thus, fluent speakers had to rely on the hint in about 20% of their turns. About 5 students relied on the hint every turn, which suggests that most students had a low level. But even if they could not respond by themselves, they learn English by playing the POMY with hints. Thus, the hint generation was essential for L2 learners to maintain a continuous dialog with the system. We recorded the number of words uttered by the students to examine whether they could become more fluent by producing more words after they practiced speaking for four weeks. As expected, students tried to include more words in their speech, using a significantly greater number of words in the post-test than in the pre-test. Some students answered questions in full sentence form in the post-test, whereas they had answered the same question with only a single word in the pre-test. This suggests that students gained confidence in their speaking after considerable practice through classes in which they communicated with the system to carry out various tasks in the game environment. In L2 learning processes, grammatical errors tend to appear as the number of learner utterances increases and the utterances become longer and more complex. Accordingly, students in this study also made more grammatical errors as they spoke more as shown in Table 5. This is related to the types of grammatical errors students made: some errors could be easily corrected, whereas others could not. After transcribing the students' utterances, we divided errors into three categories: 1) morphological errors (e.g., He go to school. I want two apple.), 2) lexical errors (e.g., I look the picture) and 3) word order errors (e.g., You buy what dress?). Although many of the lexical and

Table 5 Results of pre-test and post-test regarding language skills (GE indicates grammatical error).

		Pre-test		Post-test			
Category	N	Mean	SD	Mean	SD	Diff	P
No. of Words	25	136.3	55.3	170.0	80.8	33.7	<0.01*
No. of GE	25	42.0	6.8	44.4	6.8	2.4	<0.01*

word order errors could be corrected during the course, and fewer errors occurred in the post-test, morphological errors recurred throughout the course and lingered into the post-test. The number of morphological errors actually increased slightly in the post-test because students made more utterances in the post-test than in the pre-test. Morphological errors are observed in the production data (speaking and writing) of even advanced English learners, although they know the grammar. Because morphological errors have been often considered performance errors in SLA, they can occur even after learners acquire grammatical competence [50].

6. Conclusion

In this paper, we introduced the POSTECH immersive English study (POMY) system which is the dialog-based language learning system. We described a set of technologies were used to implement the educational 3D virtual game. Our approach applies dialog system technology and machine learning techniques for grammatical error detection. Based on the field test results, we suggest the design principles of dialog-based language learning system, such as the hint generation system for beginner in English. We also report that our technologies show state-of-the-art performance for language learning dialog systems and grammatical error detection. The results of this study bring us a step closer to understanding computer-based education.

Acknowledgements

This work was supported by the Industrial Strategic technology development program, 10035252, Development of dialog-based spontaneous speech interface technology on mobile platform funded by the Ministry of Trade, Industry and Energy (MI, Korea).

References

[1] R. Wallace, "Alice-artificial linguistic internet computer entity-the alice ai. foundation," Disponıvel em http://www.alicebot.org. Acesso em, vol.18, 1995.
[2] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," Commun. ACM, vol.9, no.1, pp.36–45, 1966.
[3] A. Kerly, P. Hall, and S. Bull, "Bringing chatbots into education: Towards natural language negotiation of open learner models," Knowledge-Based Systems, vol.20, no.2, pp.177–185, 2007.
[4] B. Heller, M. Proctor, D. Mah, L. Jewell, and B. Cheung, "Freudbot: An investigation of chatbot technology in distance education," World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp.3913–3918, 2005.
[5] S.D. Krashen, The Input Hypothesis: Issues and Implications, Longman London, 1985.
[6] M.H. Long, "Input, interaction, and second-language acquisition," Annals of the New York Academy of Sciences, vol.379, no.1, pp.259–278, 1981.
[7] R.W. Schmidt, "The role of consciousness in second language learning1," Applied Linguistics, vol.11, no.2, pp.129–158, 1990.
[8] R. Schmidt, "Awareness and second language acquisition," Annual Review of Applied Linguistics, vol.13, no.1, pp.206–226, 1993.
[9] R. Bley-Vroman, "What is the logical problem of foreign language learning," Linguistic Perspectives on Second Language Acquisition, vol.4, pp.1–68, 1989.
[10] S. Carroll and M. Swain, "Explicit and implicit negative feedback," Studies in Second Language Acquisition, vol.15, no.03, pp.357–386, 1993.
[11] N.C. Ellis, "At the interface: Dynamic interactions of explicit and implicit language knowledge," Studies in Second Language Acquisition, vol.27, no.02, pp.305–352, 2005.
[12] M.H. Long, S. Inagaki, and L. Ortega, "The role of implicit negative feedback in sla: Models and recasts in Japanese and Spanish," The Modern Language Journal, vol.82, no.3, pp.357–371, 1998.
[13] R. Lyster, "Recasts, repetition, and ambiguity in l2 classroom discourse," Studies in Second Language Acquisition, vol.20, no.01, pp.51–81, 1998.
[14] Y. Sheen, "Corrective feedback and learner uptake in communicative classrooms across instructional settings," Language Teaching Research, vol.8, no.3, pp.263–300, 2004.
[15] Z. Dörnyei, "New themes and approaches in second language motivation research," Annual Review of Applied Linguistics, vol.21, pp.43–59, 2001.
[16] J.M. O'malley and A.U. Chamot, Learning strategies in second language acquisition, Cambridge University Press, 1990.
[17] R.L. Oxford and J.A. Burry-Stock, "Assessing the use of language learning strategies worldwide with the esl/efl version of the strategy inventory for language learning (SILL)," System, vol.23, no.1, pp.1–23, 1995.
[18] L. Lee, "Synchronous online exchanges: A study of modification devices on non-native discourse," System, vol.30, no.3, pp.275–288, 2002.
[19] L. Lee, "Learners' perspectives on networked collaborative interaction with native speakers of spanish in the us," Language Learning & Technology, vol.8, no.1, pp.83–100, 2004.
[20] Z.I. Abrams, "The effect of synchronous and asynchronous cmc on oral performance in german," The Modern Language Journal, vol.87, no.2, pp.157–167, 2003.
[21] J. Pellettieri, "Negotiation in cyberspace: The role of chatting in the development of grammatical competence," Network-based Language Teaching: Concepts and Practice, pp.59–86, 2000.
[22] M. Warschauer, "Comparing face-to-face and electronic discussion in the second language classroom," CALICO journal, vol.13, no.2&3, pp.7–26, 1995.
[23] M. Warschauer, "The paradoxical future of digital learning," Learning Inquiry, vol.1, no.1, pp.41–49, 2007.
[24] S.M. Gass and E.M. Varonis, "Input, interaction, and second language production," Studies in second language acquisition, vol.16, no.03, pp.283–302, 1994.
[25] M.H. Long, "Task, group, and task-group interactions.," 1990.
[26] M.H. Long, "Input, interaction, and second-language acquisition," Annals of the New York Academy of Sciences, vol.379, no.1, pp.259–278, 1981.
[27] S. Lee, H. Noh, J. Lee, K. Lee, G.G. Lee, S. Sagong, and M. Kim, "On the effectiveness of robot-assisted language learning," ReCall, vol.23, no.1, pp.25–58, 2011.
[28] D.O. Neville, B.E. Shelton, and B. McInnis, "Cybertext redux: Using digital game-based learning to teach l2 vocabulary, reading, and culture," Computer Assisted Language Learning, vol.22, no.5, pp.409–424, 2009.
[29] N.B. Sardone and R. Devlin-Scherer, "Digital games for english classrooms," Teaching English with Technology, vol.10, no.1,

pp.35–50, 2010.
[30] D. Zheng, M.F. Young, M. Wagner, and R.A. Brewer, "Negotiation for action: English language learning in game-based virtual worlds," The Modern Language Journal, vol.93, no.4, pp.489–511, 2009.
[31] M.C. Mayrath, T. Traphagan, L. Jarmon, A. Trivedi, and P. Resta, "Teaching with virtual worlds: Factors to consider for instructional use of second life," J. Educational Computing Research, vol.43, no.4, pp.403–444, 2010.
[32] R.W. Hill Jr, J. Gratch, S. Marsella, J. Rickel, W.R. Swartout, and D.R. Traum, "Virtual humans in the mission rehearsal exercise system.," KI, vol.17, no.4, p.5, 2003.
[33] J. Gustafson, J. Boye, M. Fredriksson, L. Johanneson, and J. Königsmann, "Providing computer game characters with conversational abilities," Intelligent Virtual Agents, pp.37–51, Springer, 2005.
[34] W.L. Johnson, N. Wang, and S. Wu, "Experience with serious games for learning foreign languages and cultures," Proc. SimTecT Conference, 2007.
[35] H. Morton and M.A. Jack, "Scenario-based spoken interaction with virtual agents," Computer Assisted Language Learning, vol.18, no.3, pp.171–191, 2005.
[36] J. Brusk, P. Wik, and A. Hjalmarsson, "Deal a serious game for call practicing conversational skills in the trade domain," Proc. SLATE 2007, 2007.
[37] S. Jung, C. Lee, and G.G. Lee, "Using utterance and semantic level confidence for interactive spoken dialog clarification," JCSE, vol.2, no.1, pp.1–25, 2008.
[38] C. Lee, S. Jung, S. Kim, and G.G. Lee, "Example-based dialog modeling for practical multi-domain dialog system," Speech Commun., vol.51, no.5, pp.466–484, 2009.
[39] V. Lcvenshtcin, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics-Doklady, vol.10, p.707, 1966.
[40] H. Noh, S. Lee, K. Kim, K. Lee, and G.G. Lee, "Ranking dialog acts using discourse coherence indicator for language tutoring dialog systems," Proc. Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop, pp.203–214, 2011.
[41] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," Computer Speech & Language, vol.14, no.4, pp.373–400, 2000.
[42] S. Lee, H. Noh, K. Lee, and G.G. Lee, "Grammatical error detection for corrective feedback provision in oral conversations," AAAI, 2011.
[43] C. Lee, S. Jung, K. Kim, D. Lee, and G.G. Lee, "Recent approaches to dialog management for spoken dialog systems," JCSE, vol.4, no.1, pp.1–22, 2010.
[44] S. Lee and G.G. Lee, "Realistic grammar error simulation using markov logic," Proc. ACL-IJCNLP 2009 Conference Short Papers, pp.81–84, Association for Computational Linguistics, 2009.
[45] C.C. Chang and C.J. Lin, "Libsvm: A library for support vector machines," ACM Trans. Intelligent Systems and Technology (TIST), vol.2, no.3, p.27, 2011.
[46] M.L. Puterman, Markov decision processes: discrete stochastic dynamic programming, John Wiley & Sons, 2009.
[47] R.S. Sutton, "Planning by incremental dynamic programming," ML, pp.353–357, Citeseer, 1991.
[48] S. Jung, C. Lee, K. Kim, M. Jeong, and G.G. Lee, "Data-driven user simulation for automated evaluation of spoken dialog systems," Computer Speech & Language, vol.23, no.4, pp.479–509, 2009.
[49] E. Izumi, K. Uchimoto, and H. Isahara, "Error annotation for corpus of japanese learner english," Proc. Sixth International Workshop on Linguistically Interpreted Corpora, pp.71–80, 2005.
[50] D. Lardiere, "Dissociating syntax from morphology in a divergent l2 end-state grammar," Second Language Research, vol.14, no.4, pp.359–375, 1998.

Kyusong Lee is a Ph.D. student at the Department of Computer Science and Engineering at POSTECH, Pohang, South Korea. He received his B.S. degree at the Department of Computer Science and Engineering at Soongsil University, Seoul, South Korea. His research interests include dialog-based computer assisted language learning, grammar error, and student modelling.

Soo-ok Kweon is currently associate professor at POSTECH in Korea. She received her Ph. D degree in linguistics from the University of Hawaii at Manoa. Her primary research interests include SLA theory and practice, psycholinguistics, and cognitive neuro science. She is currently working on language variation and thought using eye tracking.

Sungjin Lee received the B.S. and Ph.D. degrees in Computer Science Engineering from POSTECH in 2006 and 2012. He is currently a post-doctoral fellow in the Language Technologies Institute at Carnegie Mellon University. His research interests include statistical dialog modeling, belief state tracking, dialog strategy learning, and machine learning. He is also interested in applying spoken language technolgies to computer-assisted language learnign setting.

Hyungjong Noh is a Ph.D. student at the Department of Computer Science and Engineering at POSTECH, Pohang, South Korea. He received his B.S. degree at the Department of Computer Science and Engineering at POSTECH. His research interests include spoken dialog system, non-task-oriented dialog management, and dialog-based computer assisted language learning.

Gary Geunbae Lee received his B.S. and M.S. degrees in Computer Engineering from Seoul National University in 1984 and 1986 respectively. He received Ph.D. degree in Computer Science from UCLA in 1991 and was a research scientist in UCLA from 1991 to 1991. He has been a professor at the CSE department, POSTECH in Korea since 1991. He is a director of the Intelligent Software laboratory which focuses on human language technology researches including natural language processing, speech recognition/synthesis, and speech translation. He authored more than 100 papers in international journals and conferences, and has served as a technical committee member and reviewer for several international conferences such as ACL, COLING, IJCAI, ACM SIGIR, AIRS, ACM IUI, Interspeech-ICSLP/EUROSPEECH, EMNLP and IJCNLP. He is currently leading several national and industry projects for robust spoken dialog systems, computer assisted language learning, and expressive TTS.