POSTECH Approaches for Dialog-based English Conversation Tutoring

Article · January 2011

CITATIONS
4

READS
84

5 authors, including:

Sungjin Lee
Pohang University of Science and Technology
24 PUBLICATIONS **132** CITATIONS
SEE PROFILE

Hyungjong Noh
Pohang University of Science and Technology
19 PUBLICATIONS **94** CITATIONS
SEE PROFILE

Kyusong Lee
Carnegie Mellon University
31 PUBLICATIONS **77** CITATIONS
SEE PROFILE

Gary Geunbae Lee
Pohang University of Science and Technology
228 PUBLICATIONS **1,615** CITATIONS
SEE PROFILE

POSTECH Approaches for Dialog-based English Conversation Tutoring

Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, and Gary Geunbae Lee
Department of Computer Science and Engineering,
Pohang University of Science and Technology (POSTECH), South Korea
E-mail: {junion, nohhj, jh21983, kyusonglee, gblee}@postech.ac.kr

***Abstract*— Although there have been enormous investments into English education all around the world, not many differences have been made to change the English instruction style. Considering the shortcomings for the current teaching-learning methodology, we have been investigating advanced computer-assisted language learning (CALL) systems. This paper aims at summarizing a set of POSTECH approaches including theories, technologies, systems, and field studies. On top of the state-of-the-art technologies of spoken dialog system, a variety of adaptations have been applied to overcome some problems caused by numerous errors and variations naturally produced by non-native speakers. Furthermore, a number of methods have been developed for generating educational feedback and mining educational data from Internet. Integrating these efforts resulted in intelligent educational robots - Mero and Engkey - and virtual 3D language learning games, Pomy. To verify the effects of our approaches on students' communicative abilities, we have conducted a field study at an elementary school in Korea. The results showed that our CALL approaches can be enjoyable and fruitful activities for students. Although the results of this study bring us a step closer to understanding computer-based education, more studies are needed to consolidate the findings.**

I. INTRODUCTION

It is a fact that the private English education fee in Korea, reaching up to 16 trillion won annually, adds a great burden to Korean economy, resulting in countless articles overflowing in the media on strengthening the public education system that focuses on enhancing students' speaking ability to straighten out their hunchbacked English ability compared with the excessive grammar knowledge. This shows clear evidence for the necessity for changing our current foreign language education system in public schools which mainly focuses on vocabulary memorization and grammar-translation methodology. Although there have been enormous investments into English education all around the world, not many differences have been made to change the rote learning style in English instruction. In addition, computer-based English learning is in the center of interest, however, this method also fails to provide the opportunity for free conversation and stays at the level of simple repetition of the given text. These teaching-learning methods cannot provide persistent motivation for learners to reach the high proficiency levels in foreign language learning. Considering the shortcomings for the current teaching-learning methodology, we have been investigating English learning systems using natural language processing technology in immersion context based on the assumptions of second language acquisition theory and practice. Through the systems, foreign language learners practice English conversation in natural contexts and are provided with corrective feedback based on the error correction procedures. POSTECH and KIST's Center for Intelligent Robotics (CIR) have been cooperating in developing robots as educational assistants, called Mero and Engkey. These robots were designed with expressive faces, and have typical face recognition and speech functions allowing learners to have a more realistic and active context. Another system, Pomy (POstech iMmersive English studY), presents a virtual reality immersion environment, where learners experience the visual, aural and tactual senses to help them develop into independent learners and increase their memory and concentration abilities to a greatest extent (Fig. 1).

The remainder of this paper is structured as follows. Section 2 describes related studies. Section 3 introduces the speech and language technologies used in our approaches. Section 4 presents a detailed description of our preliminary

Fig. 1 Mero, Engkey, and Pomy

Proceedings of the Second APSIPA Annual Summit and Conference, pages 794–803,
Biopolis, Singapore, 14-17 December 2010.

field study and the results and discussion. Finally, Section 5 gives our conclusion.

II. Related Work

A. Second Language Acquisition Theory

Since the advent of Second Language Acquisition (SLA), a number of crucial factors have been revealed for improving students' productive conversational skills: 1) comprehensible input [1], 2) comprehensible output [2], 3) corrective feedback [3], and 4) motivation and attitude [4].

In relation to oral understanding, accumulated work on the process of listening suggests that comprehension can only occur when the listener places what she or he hears in context. While comprehensible input is invaluable to the acquisition process, it is not sufficient for students to fully develop their L2 proficiency. The output hypothesis claims that production makes the learner move from 'semantic processing' prevalent in comprehension to more 'syntactic processing' that is necessary for improving accuracy in their interlanguage [2]. Specifically, producing output is one way of testing one's hypotheses about the L2. Learners can judge the comprehensibility and linguistic well-formedness of their interlanguage utterances against feedback obtained from their interlocutors, leading them to recognize what they do not know, or know only partially.

On the other hand, it has been argued that corrective feedback plays a beneficial role in facilitating the acquisition of certain L2 forms which may be difficult to learn through input alone, including forms that are rare, are low in perceptual salience, are semantically redundant, do not typically lead to communication breakdown, or lack a clear form-meaning relationship.

Motivation and attitude is another crucial factor in L2 achievement [4]. For this reason it is important to identify both the types and combinations of motivation that assist in the successful acquisition of a foreign language. In order to make the language learning process a more motivating experience, researchers need to put a great deal of thought into developing programs which maintain students' interest and have obtainable short term goals. The use of an interesting computer-based method can help to increase the motivation level of students, and computer-based learning has an advantage over human-based learning in that it seems to give a more relaxed atmosphere for language learning [5], [6], [7].

There have been few serious attempts to provide students with natural contexts that embody most of the aforementioned attributes.

B. Related Research Projects

Many research projects have tested the idea of providing pronunciation training using a speech recognizer in a forced recognition mode [8], [9], but a few systems exist that allow the user to engage in some form of meaningful dialogue.

DEAL, developed at KTH, is a roleplay dialogue system for second language learners, using a spoken dialogue system [10]. It is intended as a multidisciplinary research platform, particularly in the areas of human-like utterance generation, game dialogue, and language learning. The domain is the trade domain, specifically flea market situation. DEAL provides hints about things the user might try to say if he or she is having difficulties remembering how things are called, or if the conversation has stalled for other reasons.

Another system is the Spoken Electronic Language Learning (SPELL) system [11]. It provides opportunities for learning languages in functional situations such as going to a restaurant, expressing (dis-)likes, etc. Recast feedback is provided if the learner's response is semantically correct but has some grammatical errors. This system combines semantic interpretation and error checking in the speech recognition process. Thus, it uses a special speech recognition grammar to cover both normal speech and erroneous speech.

Spoken Conversational Interaction for Language Learning (SCILL) was developed based on the spoken dialogue system of MIT [12]. This system covers the topics of weather information and hotel booking. They implemented the simulated user to produce example dialogs to expose language learners to language use and to expand the training corpus for the system. It decides stochastically what to say on the basis of the system's previous reply [13].

The Let's Go system [14] is a spoken dialog system that provides bus schedule for the area around Pittsburgh, PA, U.S.A. This system is an extension of a previously developed system [15]. Raux and Eskenazi adapted non-native speakers' data for the use of language learning. They modified the grammar for the native speaker. Modifications include the addition of new words, new constructs and the relaxation of some syntactic constraints to accept ungrammatical sentences.

In Japan, the educational use of robots has been studied, mostly with Robovie in elementary schools, focusing on English language learning. Robovie has behavior episodes with some English dialogues. To identify the effects of a robot in English language learning, the researchers placed a robot in an elementary school, and compared the frequency of students' interaction with the English test score. The students who showed a lot of interest at the starting phase had a significantly elevated English score. This implies that robot-aided English learning can be effective for students' motivation [16].

IROBI was recently introduced by Yujin Robotics in Korea. IROBI is both an educational and home robot, containing many features. IROBI was used in [17] to compare the effects of non-computer-based media and web-based instruction with the effects of robot-assisted learning for children. Robot-assisted learning is thought to improve children's concentration, interest, and academic achievement. It is also thought to be more user-friendly than other types of instructional media.

Studies on dialogue-based computer-assisted language learning (DB-CALL) are still relatively new and most are in the early stages in a starting phase. Therefore, many attempts need to be made to investigate the effects of their use.

The following section gives an account of the speech and language technologies which have been used in our systems.

III. SPEECH AND LANGUAGE TECHNOLOGY

We have constructed DB-CALL systems, including speech recognition, language understanding, and dialog management modules, which can perceive the utterances of learners, especially Korean learners of English, to provide effective feedback and the opportunities for practicing free conversation.

A. Automatic Speech Recognition

Speech recognition is performed by the DARE recognizer [18], a speaker independent real-time speech recognizer. Since data is costly for a fully trained acoustic model for a specific accent, we have used a small amount of transcribed Korean children's speech (17 hours) to adapt acoustic models that were originally trained on the Wall Street Journal corpus using standard adaptation techniques, both of maximum likelihood linear regression (MLLR) [19] and maximum a posteriori (MAP) adaptation [20]. The occurrence of pronunciation variants was detected with a speech recognizer in forced-alignment using a lexicon expanded according to all the possible substitutions between confusable phonemes. Korean speakers tend to replace the following consonants with the correspondingly similar consonants and the following eight pronunciation variants of vowels are common to Korean speakers (Table 1).

B. Language Understanding

Since language learners commit numerous and diverse errors, a system should be able to understand language learners' utterances in spite of these obstacles. To accomplish this purpose, rule-based systems usually anticipate error types and hand-craft a large number of error rules, but this approach makes these methods sensitive to unexpected errors and diverse error combinations [11], [14], [21]. Therefore we statistically infer the actual learners' intention by taking not only the utterance itself but also the dialog context into consideration, as human tutors do (Fig. 2). The intention recognizer is a hybrid model of the dialog state model and the utterance model.

To predict the user intention from the utterance itself, we use the maximum entropy model [22] trained on linguistically-motivated features:

- **Lexical word features:** Lexical word features consist of lexical tri-grams using current, previous, and next lexical words. They are important features, but the lexical words appearing in training data are limited, so data sparseness problems can arise.
- **POS tag features:** POS tag features also include POS tag tri-grams matching the lexical features. POS tag features provide generalization power over the lexical features.

TABLE 1: LIST OF POSSIBLE SUBSTITUTIONS

Consonant	Vowel
CH → T	IH → IY
DH → D	OY → IY
TH → T	ER → R
TH → S	UH → OW
ZH → JH	EH → AE
F → P	AA → AO
R → L	AO → OW
V→ B	AH → AA

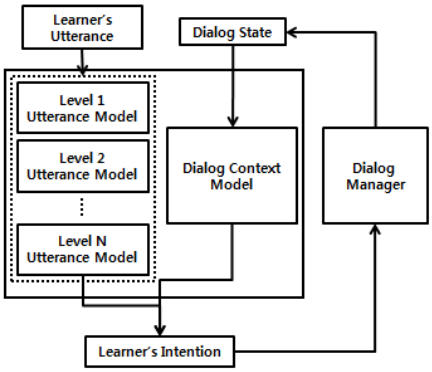

Fig. 2 Hybrid model of language understanding

The task of predicting the probable user intention in a given dialog context can be viewed as searching the dialog context space for ones that are similar to the current one and then inferring the expected user intention from the user intentions of the dialog contexts found. Therefore, we can formulate the task as the k-nearest neighbors (KNN) problem [23] with some enhancements. Our representation of a dialog context consists of diverse pieces of discourse and subtask information as shown in Table 2.

When the learner speaks, the utterance model elicits n-best hypotheses of the learner's intention which are then re-ranked by the results of the dialog state model. The detailed algorithm is described in [24].

To evaluate the proposed model, instead of involving real language learners, we simulated them by injecting grammar errors into clear utterances generated using the user simulation method described in [25]. In the first step of the error generation procedure, we set the Grammar Error Rate (GER) between 0 % ~ 100 % and determined error counts to be produced based on the GER. Then, we distributed the errors among categories and error types according to the percentages in the error types list. To verify the effectiveness of the dialog state-awareness, we compared the hybrid model

TABLE 2: REPRESENTATION OF DIALOG CONTEXT AND AN EXAMPLE FOR THE SHOPPING DOMAIN

PREV_SYS_INT	Previous system intention Ex) request(quantity)
PREV_USR_INT	Previous user intention Ex) inform(item)
SYS_INT	Current system intention Ex) confirm(quantity)
INFO_EX_STAT	A list of exchanged information states which is essential to successful task completion; (c) denotes *confirmed*, (u) *unconfirmed* Ex) [item='apple'(c), quantity=?(u)]
DB_RES_NUM	Number of database query results Ex) 0

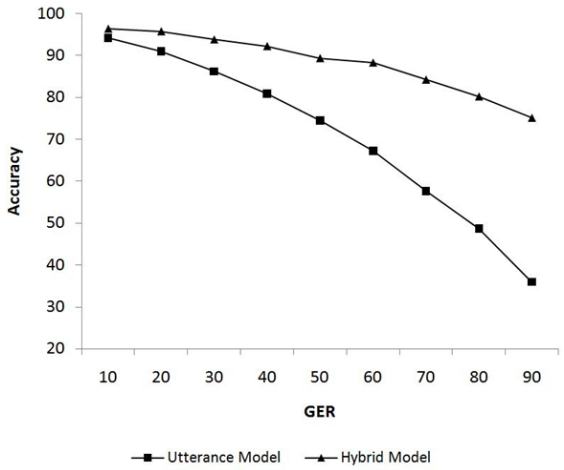

Fig. 3 Comparison between the hybrid model and the utterance only model

with the utterance-only model. The utterance-only model just omits the dialog context model from the hybrid model. We conducted 200 dialogs for each model per 10 % GER intervals. The hybrid model significantly outperformed the utterance only model for overall range of GER. As the GER increased, the performance of the utterance only model decreased dramatically, whereas the performance of the hybrid model decreased smoothly (Fig. 3). It verifies the effectiveness of dialog state-awareness through our hybrid approach.

C. Dialog Management

The dialog manager generates system responses according to the learner's intention and generates corrective feedback if needed. Our approach is implemented based on example-based dialog management (EBDM) framework, a data-driven dialog modeling, which was inspired by example-based machine translation (EBMT) [26], a translation system in which the source sentence can be translated using similar example fragments within a large parallel corpus, without knowledge of the language's structure. The idea of EBMT can be extended to determine the next system actions by finding similar dialog examples within an annotated dialog corpus. A dialog example is defined as a set of tuples that have the same semantic and discourse features. Each turn pair (one user turn and the corresponding system turns) in the dialog corpus is represented as one dialog example. The relevant examples are first grouped by a set of semantic and discourse features to represent the dialog state. The dialog examples are mapped into the relevant dialog state using a relational model that groups data using common attributes found in the data set because structured query languages (SQLs) can be easily manipulated to find and relax the dialog examples with some features. Then, the possible system actions are selected by finding semantically relevant user utterances with the current dialog state. The best system action can be expected to maximize a certain similarity metric.

A relational database is automatically built by first collecting a human–human dialog corpus related to pre-

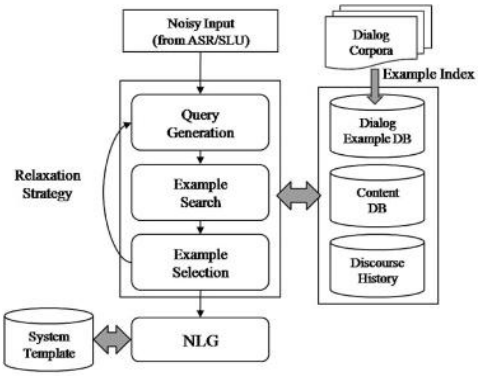

Fig. 4 Overall strategy of the EBDM framework

defined scenarios in each task. Then semantic tags (e.g., dialog act, main goal, and slot entity) are manually annotated to the user utterances, and system action tags to the system utterance. A hand-crafted automatic system is also used to extract discourse contextual features (e.g., previous intention and slot-filling status) by keeping track of the dialog states for each point in the dialog. After that, a dialog example database (DEDB) is semantically indexed to generalize the data; here the indexing keys can be determined according to state variables chosen by a system developer for domain-specific applications. Each turn pair (user turn, system turn) in the dialog corpus is mapped to semantic records in the DEDB. The index constraints represent the state variables which are domain-independent attribute. Our basic constraints consist of general features to define the dialog state such as the current user intention (dialog act and main goal), slot flags, discourse history vector, and lexico-semantic string of the current utterance. To determine the next system action, the EBDM framework uses the three following processes (Fig. 4):

- Query generation: DM generates an SQL statement using discourse history and the current dialog frame.
- Example search: DM searches for semantically similar dialog examples in the DEDB given the current dialog state. If no example is retrieved, some features can be ignored by relaxing particular features according to the level of importance given the dialog's domain.
- Example selection: DM selects the best example to maximize the example score based on lexico-semantic similarity and discourse history similarity.

The content database (DB) contains several contents which denote a set of DB results (e.g., building information, person information) returned by the current dialog frame. The slot names and the slot values in the current dialog frame are transformed into a set of query constraints to find the user's desired contents. The discourse history stores the previous discourse information (e.g., the previous dialog frame, the previous contents, and the previous system action). This information has a stack structure of the previous discourse information from the dialog's start to the current turn. The

previous discourse information represents the dialog state as the discourse features.

The EBDM framework is a simple and powerful approach to rapidly develop spoken dialog systems for multi-domain dialog processing [27]. However, this framework must solve three problems for practical dialog systems for domain-specific tasks: (1) Keeping track of the dialog state to ensure steady progress towards task completion, (2) Supporting n-best recognition hypotheses to improve the robustness of dialog manager, and (3) Enabling error handling to recover ASR and SLU errors. Consequently, we sought to solve these problems by integrating the agenda graph as prior knowledge to reflect the natural hierarchy and order of subtasks needed to complete the task. The graph is used to both keep track of the dialog state and to select the best system action using multiple recognition hypotheses for augmenting the previous EBDM framework. Dynamic help generation was also adopted as an error recovery strategy that provides immediate help messages using the agenda graph and dialog examples. Our error recovery strategies can use the discourse information to provide an intelligent guidance based on the agenda graph, and the help delivered may reflect what the user was trying to achieve at the current turn. The detailed algorithm is described in [28].

When it is desirable to offer corrective feedback, the dialog manager provides fluent utterances which realize the learner's intention. Corrective feedback generation takes two steps (Fig. 5): 1) Example Search: the dialog manager retrieves example expressions by querying Example Expression Database (EED) using the learner's intention as the search key. 2) Example Selection: the dialog manager selects the best example which maximizes the similarity to the learner's utterance based on lexico-semantic pattern matching. If the example expression is not equal to the learner's utterance, the dialog manager suggests the example as recast feedback and conducts a clarification request to induce learners to modify their utterance. Sometimes, students have no idea about what to say and they cannot continue the dialog. In such a case, timeout occurs and the utterance model does not generate hypotheses. Hence, the dialog manager searches EED with only the result of the dialog state model and suggests the retrieved expression so that students can use it to continue a

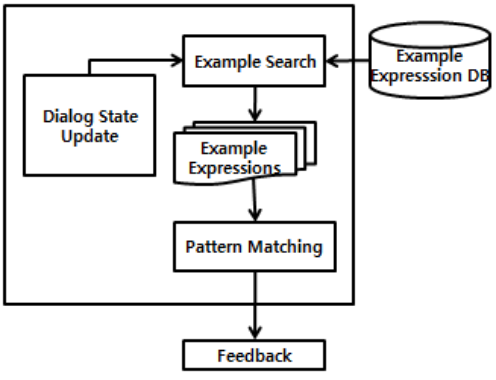

Fig. 5 Corrective feedback generation procedure

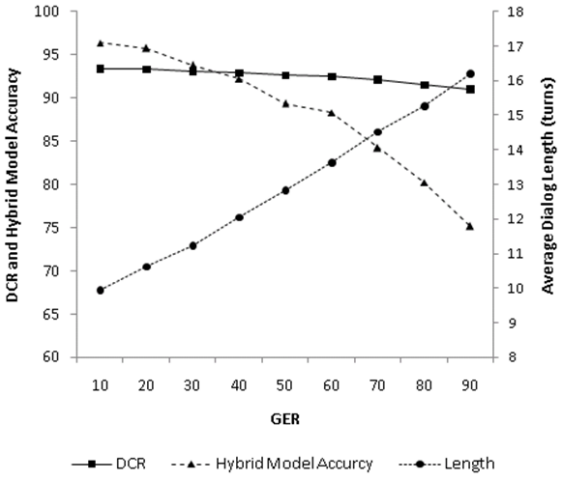

Fig. 6 The relation between Dialog Completion Rate and the performance of the hybrid model and the Average Dialog Length

conversation [24].

To evaluate the appropriateness of the feedback, we conducted 200 dialogs per 10 % GER intervals from 10 % to 90 %, and observed the Dialog Completion Rate (DCR) as the GER increased. As the GER increased, the performance of accuracy of the language understanding module (hybrid model) decreased, whereas the DCR decreased very slightly (Fig. 6). Because of the clarification sub-dialogs, the average dialog length increased as the GER increased. Based on this result, we can conclude that our method is suitable to produce appropriate feedback even when the inferred intention is not the same as the actual one. This is because the dialog context model effectively confines candidate intentions within the given context.

D. Grammar Error Simulation

Recent topics of computer-assisted language learning (CALL) research include a number of advanced technologies: generating corrective feedback in DB-CALL systems, simulating a language learner for learning tutoring strategies, and generating grammar quizzes as a game-play in educational games. This section investigates a common component of those technologies, namely grammar error simulation (GES), which plays a crucial role within them. We have developed a new method for generation of realistic grammar errors which provides an effective way to merge a statistical approach with expert knowledge about the grammar error characteristics of language learners via Markov Logic. Markov logic enables concise specification of very complex models. The task of grammar error simulation is to generate an ill-formed sentence when given a well-formed input sentence. The generation procedure involves three steps: 1) Generating probability over error types for each word of the well-formed input sentence through Markov Logic Network (MLN) inference 2) Determining an error type by sampling the generated probability for each word 3) Creating an ill-formed output sentence by realizing the chosen error types (Fig. 7).

		He	wants	to	go	to	a	movie	theater	
	v_agr_sub	0.000	**0.371**	0.000	0.000	0.000	0.000	0.000	0.000	*1 step*
	prp_lex_del	0.000	0.000	**0.284**	0.000	0.269	0.000	0.000	0.000	
Inference	*at_del*	0.000	0.000	0.000	0.000	0.000	**0.355**	0. 000	0.000	
	...	...	...	...	...	...	...	...	...	
	none	0.921	0.449	0.604	0.866	0.605	0.506	0.781	0.798	*2 step*
Sampling		none	v_agr_sub	prp_lex_del	none	none	at_del	none	none	*3 step*
Realization		He	***want***		go	to		movie	theater	

Fig. 7 An example process of grammar error simulation

Our MLN implementation consists of three components: 1) Basic formulas based on parts of speech, which are comparable to the previous study [29] 2) Analytic formulas drawn from expert knowledge obtained by error analysis on a learner corpus 3) Error limiting formulas that penalize statistical model's over-generation of nonsense errors. The analytic formulas add concrete knowledge of realistic error characteristics of language learners. Error analysis and linguistic differences between the first language and the second language can identify various error sources for each error type. For example, English learners often commit pluralization error with irregular nouns. This is because they over-generalize the pluralization rule, i.e. attaching 's/es', so that they apply the rule even to irregular nouns such as 'mice' and 'feet' etc. This characteristic is captured by the simple formula:

- $IrregularPluralNoun(s,i) \wedge PosTag(s,i,NNS) \Rightarrow ErrorType(s,i,N\_NUM\_SUB)$

where IrregularPluralNoun(s, i) is true if and only if the ith word of the sentence s is an irregular plural and N_NUM_SUB is the abbreviation for substitution by noun number error. More detailed algorithm is described in [30].

Experiments used the NICT JLE Corpus, which is speech samples from an English oral proficiency interview test, the ACTFL-ALC Standard Speaking Test (SST). 167 of the files are error annotated. The error tagset consists of 47 tags that are described in [31]. We appended structural type of errors (substitution, addition, deletion) to the original error types because structural type should be determined when creating an error. For example, V_TNS_SUB consists of the original error type V_TNS (verb tense) and structural type SUB (substitution).

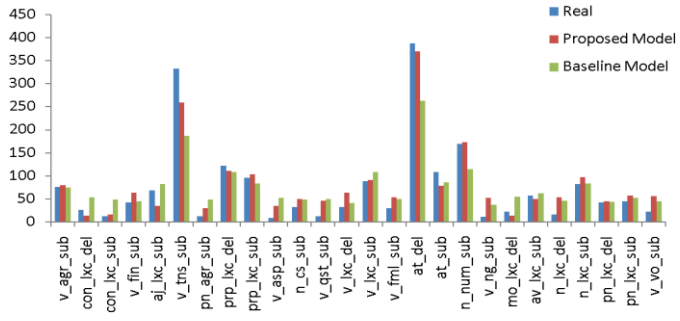

Fig. 8 Comparison between the distributions of the real and simulated data

The grammar error simulation was compared with real learners' errors and the baseline model using only basic formulas, with 10-fold cross validations performed for each group. The validation results were added together across the rounds to compare the number of simulated errors with the number of real errors. Error types that occurred less than 20 times were excluded to improve reliability. Result graphs suggest that the distribution of simulated grammar errors generated by the proposed model using all formulas is similar to that of real learners for all level groups and the proposed model outperforms the baseline model using only the basic formulas (Fig. 8). Two human judges verified the overall realism of the simulated errors. They evaluated 100 randomly chosen sentences consisting of 50 sentences each from the real and simulated data. The sequence of the test sentences was mixed so that the human judges did not know whether the source of the sentence was real or simulated. They evaluated sentences with a two-level scale (0: Unrealistic, 1: Realistic). The result shows that the inter evaluator agreement (kappa) is moderate and that both judges gave relatively close judgments on the quality of the real and simulated data (Table 3).

E. Educational Data Mining

Just using conventional dialog systems in a foreign language would not be beneficial because language learners do not have good comprehension ability in general. It highlights the need of feedback providing cultural and contextual information to help learners understand what the system says. This section presents our investigation on developing a web-mining method to collect information for comprehension-aid feedback from Internet.

In particular, we focus on mining English as a second language (ESL) podcast related sites [32]. An ESL podcast document consists of two parts, the script part and the description part. These documents are transcriptions of audio files recorded by native speakers. The script part can be a

TABLE 3: HUMAN EVALUATION RESULTS

	Human 1	**Human 2**	**Average**	**Kappa**
Real	0.84	0.8	0.82	0.46
Simulated	0.8	0.8	0.8	0.5

dialog or a short article. The description part explains the content of the script part. Usually, ESL podcast documents contain rich descriptions about the cultural and contextual information in which some expressions that the student is trying to learn are used. To facilitate the access to such information by DB-CALL systems, we are developing a web-mining method which performs automatic extraction of pairs of the expression to learn and the corresponding description. The detailed process takes the following steps. First, effective linguistic patterns are mined from ESL podcast documents which play crucial roles in characterizing descriptions which contain expressions to explain. Having obtained linguistic patterns, we train a classifier with binary features representing the existence of the patterns. By applying the classifier to ESL podcast documents, descriptions for key expressions to learn can be detected. From the descriptions detected, we extract the expression segments using a simple alignment technique. To raise the accuracy, we confirm the extracted expressions by matching them with the sentences in the script part.

However, it is difficult for simple statistical methods to capture such effective sequential patterns. Although N-gram language model is considered to be effective in characterizing sequential patterns, it becomes very expensive if N > 3 and N-grams only consider continuous sequence of words, which is unable to detect non-continuous patterns, e.g., "A(n) ... is ... that" in the following sentences.

- "**An** organ **is** a musical instrument **that** has long pipes that play different notes."
- "**A** resource **is** anything **that** you can use to get something done."

Therefore, we adapt labeled sequential patterns to effectively characterize the features of description sentences. The labeled sequential patterns (LSP) are proven very effective to the problems where non-continuous sequential patterns are needed such as identifying erroneous sentences [33]. Please refer to [34] for detailed description of the algorithm.

To evaluate the proposed method, we have acquired 200 documents from ESL podcast web site. We randomly selected 160 documents as the train data set, and the other 40 documents as the test data set. All documents have been annotated by a total of eight annotators. The annotators have tagged descriptions in accordance with the sentences in the script part and every expression which is explained by the descriptions where it is belonging. The train data set consists of 14112 sentences. The test data set includes 3429 sentences.

Unfortunately, there have been no previous studies targeting the same task with the one this research focuses on. Therefore we set up three different classification models to compare with each other: 1) trigger-words model, 2) N-grams model, and 3) the proposed method. Trigger-words model which exploited a total of 18 trigger-words (e.g., mean, describe, and express) classifies a sentence as description if the sentence contains at least one trigger-word. N-grams model used all unigrams and bigrams of the train data set as its features for training a support vector machine (SVM) model. Finally, the proposed method was taking LSPs in addition to the unigrams and bigrams to further cover discontinuous sequential patterns.

TABLE 4: THE PERFORMANCE OF THE THREE CLASSIFICATION MODELS

Method	Precision	Recall	F1-score
Trigger-words	0.8460	0.4727	0.6065
N-grams	0.8026	0.7748	0.7885
Proposed method	0.8086	0.8857	0.8454

Table 4 shows the experimental results of the three classification models. The performance of the trigger-words model was 0.6065 in terms of F1 score mainly due to the low recall. As we expected, the N-grams model improve the performance of recall leading to much higher F1-score 0.7885. However, we could still obtain more performance gain by exploiting LSPs which enable the classification model to take into consideration discontinuous sequential patterns. The performance of the proposed method showed precision 0.8026, recall 0.8857, and F1-score 0.8454. Note that there was a large gap between the recall of the proposed method and that of the N-grams model.

IV. FIELD STUDY

We performed a field study at a Korean elementary school to investigate the educational effects of our approaches using the educational robots, Mero and Engkey. The following subsections describe the method of the study in more detail.

A. Setting and Participants

A total of 24 elementary students were enrolled in English lessons two days a week for a total of about two hours per day and had chant and dance time on Wednesdays for eight weeks during the winter vacation. However, three students left the study, resulting in a total of 21 students. The students in this study were recruited by the teachers of the school and divided into beginner-level and intermediate-level groups, according to the pre-test scores. They ranged from second to sixth grade; in general, there are six grades in a Korean elementary school. All of them were South Korean, spoke Korean as their first language and were learners of English as a foreign language. None of the participants had stayed in an English-speaking country, such as the United States and United Kingdom, for more than three months, which may indicate that this group had limited English proficiency. Fig. 9 shows the layout of the classroom: 1) PC room where students took lessons by watching digital contents, 2) Pronunciation training room where the Mero robot performed automatic scoring of pronunciation quality for students' speech and provided feedback, 3) Fruit and vegetable store, and 4) Stationery store where the Engkey robots acted as sales clerks and the students as customers.

B. Material and Treatment

The researcher produced training materials including a total of 68 lessons, with 17 lessons for each combination of the

Fig. 9 Students interacting with Mero and Engkey

level, beginner and intermediate, and the theme, fruit and vegetable store and stationery store. Among other things, the course involves small talks, homework checking, purchases, exchanges, refunds, etc. Participants in this course should become thoroughly trained in various shopping situations. With this aim in mind, when dealing with task assignment, the instructors proceeded in subtle gradations, moving from the simple to the complex. Throughout the course of the study, each student was asked to enter the four rooms in the order of PC room, Pronunciation training room, Fruit and vegetable store, and Stationery store so that students were gradually exposed to more active oral linguistic activities.

C. Data Collection and Analysis

In order to measure the cognitive effects, i.e., improvement of listening and speaking skills, all students took a pre-test at the beginning of the study and a post-test at the end. For the listening skill test, 15 multiple-choice questions were used which were developed by experts in evaluation of educational programs. The items in the test were mainly selected from the content taught during the course. The test was used as the assessment tool in both the pre-test and the post-test phases of the study. The speaking skill test consisted of 10 1-on-1 interview items. The topics of the interviews were selected from the content taught. The evaluation rubric measured speaking proficiency on a five point scale in four categories: pronunciation, vocabulary, grammar, and communicative ability. A paired t-test was performed using the mean scores and standard deviations to determine if any significant differences occurred.

In order to investigate the effects on affective factors such as satisfaction in using robots, interest in learning English, confidence with English, and motivation for learning English, a questionnaire was designed by 10 teachers and experts in evaluation of educational programs. It consisted of some personal information and 52 statements in accordance with four-point Likert scale, which had a sliding answer scale of 1-4, ranging from "strongly disagree" to "strongly agree", without a neutral option. Mean and standard deviation were used to evaluate the effect on students' satisfaction, whereas a pre-test/post-test method was used for other factors.

D. Results and Discussion

According to the findings, there were large improvements of the speaking skills in the beginner-level participants' achievement on the post-test. The score in the post-test is significantly better than that of the pre-test. The listening skill, however, showed no significant difference. Significant differences of the speaking skill were also found in the result of the intermediate group and the effect sizes are also large, whereas the listening skill showed a significantly negative effect. The combined results of both groups showed no significant differences in the listening skill (Table 5). This finding can be explained by a number of factors such as the unsatisfactory quality of the text-to-speech component and hindrance of robots' various sound effects. The large improvement of speaking skill in the overall results agrees with the findings of previous studies in general. Specifically, the gain in the vocabulary area indicates that the authentic context facilitated form-meaning mapping and the vocabulary acquisition process. The improved accuracy of pronunciation and grammar supports the output hypothesis and the effects of corrective feedback. Learners had feedback at any related point which made them reflect on their erroneous utterances. The increase of communicative ability shows that learners were getting accustomed to speaking English. It also can be attributed to the fact that when using robot-assisted learning the student gained confidence in a relaxed atmosphere. A lack of confidence and a feeling of discomfort were more related to students' participation in face-to-face traditional discussions, and less to participation in computer-based learning. Please refer to [35] for detailed information of the

TABLE 5: COGNITIVE EFFECTS ON ORAL SKILLS FOR OVERALL STUDENTS

Category		N	Pre-test		Post-test		Mean difference	t	df	Effect size
			Mean	SD[a]	Mean	SD[a]				
Listening		21	10.95	3.2	10.67	1.91	-0.29	-0.55	20	0.12
Speaking	Pronunciation	21	32.14	8.86	45.62	4.28	13.48	9.48*	20	0.90
	Vocabulary	21	32.95	8.21	42.38	5.31	10.43	8.00*	20	0.87
	Grammar	21	31.62	7.96	40.62	4.43	9.00	7.59*	20	0.86
	Communicative ability	21	33.57	9.83	47.48	3.06	13.91	7.60*	20	0.86
	Total	21	123.13	34.13	176.1	16.53	46.81	8.48*	20	0.88

* $p < .01$, SD[a] = Standard Deviation

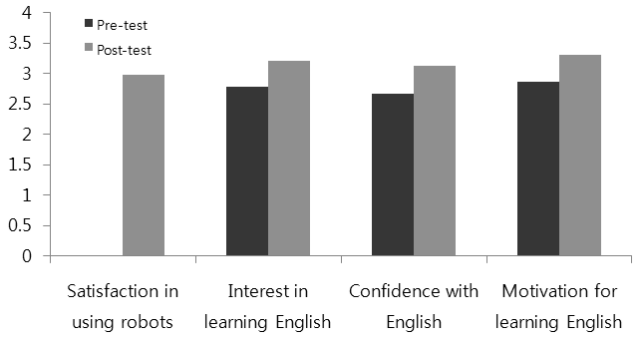

Fig. 10 The effects on affective factors

cognitive effects.

As it is shown in Fig. 10, the students were highly satisfied in using robots for language learning. But, some questions showed the need to develop a more anthropomorphic appearance and a natural voice. The students' responses to the questions about interest in learning English on pre- and post-test showed a large improvement of interest with significance level of 0.01. But the lower score of the question about increase of familiarity with English might reflect that engaging in studying English for only two months is not enough to get familiar with listening and speaking English. A significantly large increase of confidence was found in the responses to the questions about confidence in English on pre- and post-test with significance level of 0.01. This can be attributed to the fact that using robot-assisted learning allowed the students to make academic achievement and get confidence through repeated exercises in a relaxed atmosphere. However, relatively low scores were given to the questions related to individual level of fear or anxiety associated with either real or anticipated communication with another person or persons. The responses to the questions about motivation for learning English presented a large enhancement of motivation, with significance level of 0.01. The low score of the questions related to preparing to study English may illustrate that traditional education doesn't work for the new generation of children. The popularity of e-Learning in Korea is promoting the increasing disengagement of the "Net Generation" or "Digital Natives" from traditional instruction.

V. Conclusion

In this paper, we described the rationale of POSTECH approaches for CALL from a theoretical view of language learning and briefly introduced a set of technologies that we used to implement the educational assistant robots and 3D virtual language learning games. Our approaches basically apply many adaptations to the state-of-the-art technologies of spoken dialog system to overcome some problems caused by numerous errors and variations of non-native speakers. Furthermore, a number of methods have been developed for generating educational feedback and mining educational data from Internet. In addition, to investigate the cognitive and affective effects of our approaches, a course was designed in which students had meaningful interactions with intelligent robots in an immersive environment. The result showed no significant difference in the listening skill, but the speaking skills improved with a large effect size. Also, it showed that the systems promote and improve students' satisfaction, interest, confidence, and motivation. The results showed that our CALL approaches can be an enjoyable and fruitful activity for students. Although the results of this study bring us a step closer to understanding computer-based education, more studies are needed to consolidate/refute the findings of this study over longer periods of time using different activities with samples of learners of different ages, nationalities, and linguistic abilities.

Acknowledgment

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 Knowledge Economy (MKE, Korea), and the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-C1090-1031-0009), funded by the Ministry of Knowledge Economy (MKE, Korea).

References

[1] S.D. Krashen, "The input hypothesis: Issues and implications," New York, 1985.

[2] M. Swain, "Communicative competence: Some roles of comprehensible input and comprehensible output in its development," Input in second language acquisition, vol. 15, 1985, pp. 165–179.

[3] M.H. Long, "Focus on form in task-based language teaching," Language policy and pedagogy: Essays in honor of A. Ronald Walton, 2000, pp. 179–192.

[4] A.M. Masgoret and R.C. Gardner, "Attitudes, Motivation, and Second Language Learning: A Meta-Analysis of Studies Conducted by Gardner and Associates," Language Learning, vol. 53, 2003, pp. 167–210.

[5] A. Liang and R.J. McQueen, "Computer Assisted Adult Interactive Learning in a Multi-Cultural Environment.," Adult Learning, vol. 11, 1999.

[6] J. Roed, "Language learner behaviour in a virtual environment," Computer Assisted Language Learning, vol. 16, 2003, pp. 155–172.

[7] H. Yi and J. Majima, "The teacher-learner relationship and classroom interaction in distance learning: A case study of the Japanese language classes at an American high school," Foreign Language Annals, vol. 26, 1993, pp. 21–30.

[8] J. Dalby, "Explicit pronunciation training using automatic speech recognition technology," Research in technology and second language education: developments and directions, 2005, p. 379.

[9] A. Neri, C. Cucchiarini, and H. Strik, "Effective feedback on L2 pronunciation in ASR-based CALL," Proc. of the workshop on Computer Assisted Language Learning, 2001, pp. 40–48.

[10] J. Brusk, P. Wik, and A. Hjalmarsson, "DEAL: A Serious Game for CALL Practicing Conversational Skills in the Trade Domain," The Proceedings of SlaTE-Workshop on Speech and Language Technology in Education, Pennsylvania, USA, 2007.

[11] H. Morton and M.A. Jack, “Scenario-based spoken interaction with virtual agents,” Computer Assisted Language Learning, vol. 18, 2005, pp. 171–191.
[12] V.W. Zue and J.R. Glass, “Conversational interfaces: Advances and challenges,” Proceedings of the IEEE, vol. 88, 2000, pp. 1166–1180.
[13] S. Seneff, C. Wang, and J. Zhang, “Spoken conversational interaction for language learning,” InSTIL/ICALL Symposium 2004, 2004.
[14] A. Raux and M. Eskenazi, “Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges,” InSTIL/ICALL Symposium 2004, 2004.
[15] A.I. Rudnicky, C. Bennett, A.W. Black, A. Chotomongcol, K. Lenzo, A. Oh, and R. Singh, “Task and domain specific modelling in the Carnegie Mellon Communicator system,” Sixth International Conference on Spoken Language Processing, 2000.
[16] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, “Interactive robots as social partners and peer tutors for children: A field trial,” Human-Computer Interaction, vol. 19, 2004, pp. 61–84.
[17] J. Han, M. Jo, S. Park, and S. Kim, “The educational use of home robots for children,” IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005, 2005, pp. 378–383.
[18] D.H. Ahn and M. Chung, “One-Pass Semi-Dynamic Network Decoding Using a Subnetwork Caching Model for Large Vocabulary Continuous Speech Recongnition,” IEICE TRANSACTIONS on Information and Systems, vol. 87, 2004, pp. 1164–1174.
[19] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer speech and language, vol. 9, 1995, p. 171.
[20] G. Zavaliagkos, R. Schwartz, and J. McDonough, “Maximum a posteriori adaptation for large scale HMM recognizers,” Proceedings of the Acoustics, Speech, and Signal Processing, 1996, pp. 725–728.
[21] D. Schneider and K.F. McCoy, “Recognizing syntactic errors in the writing of second language learners,” Proceedings of the 17th international conference on Computational linguistics-Volume 2, 1998, pp. 1198–1204.
[22] A. Ratnaparkhi, “Maximum entropy models for natural language ambiguity resolution,” Ph.d. Dissertation, University of Pennsylvania, 1998.
[23] B.V. Dasarathy, Nearest Neighbor: Pattern Classification Techniques, Ieee Computer Society, 1990.
[24] S. Lee, C. Lee, J. Lee, H. Noh, and G.G. Lee, “Intention-based Corrective Feedback Generation using Context-aware Model,” Proceedings of International Conference on Computer Supported Education, 2010.
[25] 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, 2009, pp. 479–509.
[26] M. Nagao, “A framework of a mechanical translation between Japanese and English by analogy principle,” Readings in machine translation, 2003, p. 351.
[27] C. Lee, S. Jung, S. Kim, and G.G. Lee, “Example-based dialog modeling for practical multi-domain dialog system,” Speech Communication, vol. 51, 2009, pp. 466–484.
[28] C. Lee, S. Jung, K. Kim, and G.G. Lee, “Hybrid approach to robust dialog management using agenda and dialog examples,” Computer Speech & Language, vol. 24, Oct. 2010, pp. 609-631.
[29] J. Foster, “Treebanks gone bad: Generating a treebank of ungrammatical English,” Proc. IJCAI Workshop on Analytics for Noisy Unstructured Data, 2007.
[30] S. Lee and G.G. Lee, “Realistic grammar error simulation using Markov Logic,” Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009, pp. 81–84.
[31] E. Izumi, K. Uchimoto, and H. Isahara, “Error annotation for corpus of Japanese Learner English,” Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005), 2005, pp. 71–80.
[32] http://www.eslpod.com
[33] G. Sun, X. Liu, G. Cong, M. Zhou, Z. Xiong, J. Lee, and C.Y. Lin, “Detecting erroneous sentences using automatically mined sequential patterns,” Annual Meeting-Association for Computational Linguistics, 2007, p. 81.
[34] H. Noh, S. Lee, J. Lee, K. Lee, and G.G. Lee, “Extracting Expression-Description Pairs from ESL Podcast Documents for Dialog Based Computer-Assisted Language Learning,” Proceedings of Interspeech Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology (to be published), 2010.
[35] S. Lee, H. Noh, J. Lee, K. Lee, and G.G. Lee, “Cognitive Effects of Robot-Assisted Language Learning on Oral Skills,” Proceedings of Interspeech Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology (to be published), 2010.