Vol.2011-SLP-89 No.7
2011/12/19

Spoken Language Processing for DBCALL and English Speaking Assistants

Gary Geunbae Lee[†]

Abstract

Although enormous investments have been made to develop English language education all over the world, not many changes have been made to the style of English language instruction. Keeping in mind the shortcomings of the current methodology of English teaching and learning, we have been investigating the use of advanced computer assisted language learning (CALL) systems. We have been working to develop a framework that educates students of the English language in pronunciation, prosody, and grammar. The students can then use this knowledge for developing their English speaking and writing skills. This paper summarizes a set of POSTECH approaches, including theories, technologies, systems, and field studies, and provides relevant pointers. Many errors and variations of speech are generated by non-native English speakers. Building on state of-the-art technology that uses spoken dialog systems, and assisted by the incorporation of sophisticated linguistic knowledge, a variety of adaptations have been applied to solve such problems. In addition, a number of methods have been developed for generating feedback to help students of English become more proficient. As an outcome of our efforts, we have created two intelligent educational robots named Mero and Engkey, and a virtual 3D language learning game named Pomy. To test the effectiveness of our approach on the communication skills of students, we conducted a field study at an elementary school in Korea. The results have shown that our CALL based approach can be an enjoyable and fruitful activity for students. Although the results of this study bring us a step closer to understanding computer based language education, more studies are needed to consolidate the findings of our approach.

1. INTRODUCTION

Although huge investments have been made in English language education all over the world, they have not made much of a difference to the current rote-based learning style of English instruction. In addition, although computer based English learning is a topic of central interest, the current method of English instruction fails to provide an opportunity for conversational practice and remains at the level of simple repetition of textbook material. Such methods of instruction cannot offer any meaningful motivation for students to become proficient in English. Considering the shortcomings of the current teaching methodology, we have been investigating English teaching systems using natural language processing technology in an immersive way. Our systems are based on the assumptions of second language acquisition theory and practice. Using these systems, students learing English as a foreign language can practice English conversation in a natural context. They are provided with corrective feedback based on our error correction procedures. POSTECH and KIST's Center for Intelligent Robotics (CIR) have been cooperating in developing robots as educational assistants. The robots are named Mero and Engkey. They were designed with expressive faces, and have typical face recognition and speech functions that allow learners to get a more realistic and active experience. Another system, Pomy (POstech iMmersive English studY), presents a virtual reality environment for immersive study. Here, students can experiment with the use of their visual, aural and tactile senses to help them become more independent in their study of the English language. This system also helps them increase their memory and concentration (Fig. 1). To assess the English speaking and writing skills of the students, and to help improve their skills, a new framework was developed. This framework consists of two parts: the first is for assessing skills, and the second is for providing feedback. The framework has been applied to teaching pronunciation, prosody, and grammar.

Fig. 1 Mero, Engkey, and Pomy

[†]Department of Computer Science and Engineering
Pohang University of Science and Technology (POSTECH), South Korea
gblee@postech.ac.kr

2. DBCALL – DIALOG BASED COMPUTER ASSISTED LANGUAGE LEARNING

2.1 Automatic Speech Recognition

Speech recognition is performed by the DARE recognizer [1], a speaker independent real-time speech recognizer. Because the data for a fully trained acoustic model for a specific accent is expensive, we have used a small amount of transcribed Korean children's speech (17 hours). We have used this to adapt acoustic models that were originally trained on the Wall Street Journal corpus. Standard adaptation techniques were used, including both maximum likelihood linear regression (MLLR) [2] and maximum a posteriori (MAP) adaptation [3]. The occurrence of variations in pronunciation was detected with the help of a speech recognizer in forced alignment, using a lexicon expanded keeping in mind all the possible substitutions between phonemes that can possibly be confused with each other.

2.2 Language Understanding

Since language learners commit a numerous and diverse set of errors, a language teaching system should be able to understand what learners say, in spite of the obstacles involved. 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 [4, 5, 6]. Therefore, we use statistics to infer the intention of the learner. We do this by taking into account not just the utterance itself but also the dialog context into consideration, as human tutors do. The intention recognizer is a hybrid model of the dialog state model and the utterance model [7].

2.3 Dialog Management

The dialog manager generates system responses according to a student's intention and generates corrective feedback if needed. Our approach is implemented based on the example based dialog management (EBDM) framework, a data driven dialog modelling framework that was inspired by example based machine translation (EBMT) [8]. EBMT is a translation system in which the source sentence can be translated using similar example fragments within a large parallel corpus, without any knowledge of the language's structure. The idea of EBMT can be extended to determine the system's next 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 turn) in the dialog corpus is represented as one dialog example. The relevant examples are initially grouped using 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. The model puts data into groups 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. After that, the possible system actions are selected by finding semantically relevant user utterances to the current dialog state. The best system action can be expected to maximize a certain similarity metric.

2.4 Grammar Error Simulation

We have developed a new method for the generation of realistic grammar errors. It provides an effective way to merge a statistical approach with a Markov Logic based approach that uses expert knowledge about the grammar error characteristics of language learners. 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 consists of 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, and (3) Creating an ill-formed output sentence by realizing the chosen error types [9].

2.5 Grammar Error Correction

We have also developed a grammatical error correction system which detects and corrects grammatical errors in response to a learner's utterances. The grammatical error correction system takes a confusion network (CN) as an input, checks the grammatical accuracy of each word using pattern matching and the support vector machine (SVM), and classifies its error types. Once a CN comes in, the system extracts a feature vector made of the CN-scores of words matching the words in error patterns generated by the grammatical error simulator. Error patterns are five word sequences with their corresponding corrections and error types extracted from ill formed sentences generated by the grammatical error simulator. With the extracted feature vector, we then check whether the word sequence is grammatically correct using a binary SVM. Error types are classified after a grammar check using the weighted sum of the score and the error type frequency of each error pattern [9].

3. PESAA – POSTECH ENGLISH SPEAKING ASSESSEMENT AND ASSISTANT

3.1 Pronunciation Education

We have developed a new framework for assessing students' pronunciation and providing

them with appropriate feedback. The framework is designed to provide a training environment for improvised speech within a vocabulary that is limited to the inventory of words known to the students. The framework consists of three parts: (1) a pronunciation simulation part that learns and produces the pronunciation of non-native speakers from actual pronunciations of non-native speakers and canonical pronunciations; (2) a speech recognition part that internally generates word and phoneme level recognition results using two different ASRs and marks the mismatching phonemes as candidate errors by comparing two recognition results; and (3) an error detection and feedback part that detects pronunciation errors from error candidates and generates proper feedback relying on the ASR confidences, error classification confidences, and some pre-defined feedback preferences.

The process begins with building an extended pronunciation dictionary (EPD) that contains expected non-native pronunciation variants as well as canonical pronunciations. The pronunciations in the EPD comprise the search space of the ASR decoders in the second part. When a student utters a speech segment, the segment is recognized and converted into a word level transcription. The words in the transcription are expanded into multiple pronunciations by EPD and the phonemes of the pronunciations are connected into an extended recognition network (ERN). The phoneme level ASR recognizes the input speech again within the ERN and produces a phoneme level recognition result of the speech. The comparison module compares the phoneme recognition result with the canonical pronunciation corresponding to the word level transcription using a sequence alignment algorithm such as Levenshtein distance, to find out candidate errors. The error detection module classifies each candidate error into two classes: feedback and non-feedback, considering both the significance of the error and the necessity of feedback. The feedback generation module finally generates appropriate feedback based on the classification and presents it to the student.

3.2 Prosody Education

The components of prosody are rhythm, stress, and intonation. Of these components, we have focused on stress so far. Every word spoken in isolation has a stress. However, when words are put together in a sentence, only some words are stressed. Sentence stress emphasizes the portion of the utterance that is more important for the speaker or that the speaker wants the listener to concentrate on. The words which are likely to be more prominent and to carry a stress are those which are the most important for meaning. The system contains two models which require training data: (1) a sentence stress prediction model, and (2) a sentence stress detection model. To train the prediction model which predicts an appropriate stress pattern for a given sentence, the Boston University radio news corpus [10] was used, in which native speakers' stress patterns are reflected. This corpus consists of seven hours of speech spoken by seven native announcers along with orthographic transcription, phonetic alignments, part-of-speech tags, and prosodic labels. From the prosodic labels written in the ToBI system [11], the pitch accent with an asterisk is considered as a stressed syllable [12]. To train the detection model which detects a stress pattern for given leaners' speech, an in-house sentence stress labelled corpus was used, in which non-native speakers' stress patterns are reflected. This corpus consists of six hours of speech spoken by 72 Korean speakers along with orthographic transcription and sentence stress marks. The stress marks were manually labelled by linguists and were cross-checked. Prediction and detection models have been developed using conditional random fields. Using sophisticated linguistic rules, we have incorporated such rules into machine learning features. Another model, which provides corrective feedback, uses output probabilities of sentence stress prediction and detection models. It is set to minimize incorrect feedback. For example, if the difference of output probabilities between predicted sentence stress and detected sentence stress is high enough, the feedback model generates feedback to students to let them know whether they are right or wrong. If not, the feedback model does not provide feedback, because it is uncertain whether it is right or wrong.

3.3 Grammar Education

We include the grammatical error correction system, explained in section II.E, to assess and assist language learners. A student's grammatical ability as a primary language skill can be measured by detecting errors. Our system helps students improve their grammar skills by giving corrective feedback.

4. FIELD STUDY

We performed a field study at a Korean elementary school to investigate the results of our approach using the educational robots, Mero and Engkey.

4.1 Setting and Participants

A total of 21 elementary students were enrolled in English lessons two days a week for about two hours per day and had chanting and dancing sessions for eight weeks. The students were recruited by teachers of the school and divided into beginner level and intermediate level groups, according to the pre-program test scores. The students ranged from second grade to sixth grade. All of them were South Korean, spoke Korean as their first language and were students of English as a foreign language. None of the participants had stayed in an English speaking country for more than three months, which may indicate that this group had limited English proficiency. Fig. 2 shows the layout of the classroom: (1) A PC room where students

Fig. 2 Students interacting with Mero and Engkey

took lessons by watching digital contents, (2) A pronunciation training room where the Mero robot performed automatic scoring of pronunciation quality for students' speech and provided feedback, (3) A fruit and vegetable store, and (4) Stationery store where the Engkey robots acted as sales clerks with the students as customers.

4.2 Results and Discussion

There were large improvements in the speaking skills of beginner level participants in the post-program test. The scores in the post-program test were significantly better than that of the pre-program test. The listening skills, however, showed no significant difference. Significant differences in speaking skill were also found in the results of the intermediate group and the effect sizes were also large, whereas the listening skill showed a significantly negative effect. The combined results of both groups showed no significant differences in listening skill (Table 1). These findings can be explained by a number of factors, such as the unsatisfactory quality of the text-to-speech component and problems with the robots' various sound effects. The large improvement of speaking skill in the overall results agrees with the findings of previous studies. Specifically, the improvements in vocabulary indicate that the authentic context facilitated the mapping of form to meaning and to the vocabulary acquisition process. The improved results in pronunciation and grammar support our hypothesis about the effects of corrective feedback. Learners had access to feedback at any relevant point in time, which allowed them to work on their errors in speech. The

TABLE 1: COGNITIVE EFFECTS ON ORAL SKILLS FOR OVERALL STUDENTS

Category		N	Pre-test		Post-test		Mean	t	df	Effect
			Mean	SD[a]	Mean	SD[a]	difference			size
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

improvement in communication ability shows that students were getting accustomed to speaking English. It can also be attributed to the fact that when using robot-assisted learning, the student gained confidence in a relaxed atmosphere. Any lack of confidence or feeling of discomfort was more likely when students participated in traditional face-to-face discussions, and less to participation in computer-based learning. Please refer to [13] for detailed information about the cognitive effects. As is shown in Fig. 3, the students were quite satisfied with using robots for language learning. But some questions highlighted the need to develop a more anthropomorphic appearances and more natural voices. The responses of the students to questions regarding interest in learning English before and after the tests showed a significant improvement of interest, with a significance level of 0.01. But the lower score in answer to the question regarding an increase in familiarity with English might reflect the possibility

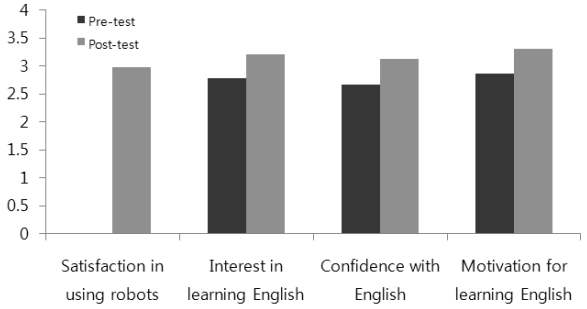

Fig. 3 The effects on affective factors

that studying English for only two months is not enough to become familiar with listening to and speaking in English. A significant increase in confidence was found in the responses to questions about confidence in English on the pre and post-program tests, with a significance level of 0.01. This can be explained by the observation that using robot-assisted learning allowed students to do well in academics, and acquire confidence via repeated exercises in an informal atmosphere. However, relatively low scores were given to questions related to individual levels of fear or anxiety, associated with either real or anticipated communication with other people. Responses to questions about individual motivations for learning English presented a significant enhancement of motivation, with a significance level of 0.01. The low scores in answer to questions related to how to prepare to study English may illustrate the possibility 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.

5. CONCLUSIONS

Our approach applies a number of adaptations to state-of-the-art technologies relevant to spoken dialog systems to overcome problems caused by numerous errors and variations in the speech of non-native speakers. Furthermore, a number of methods have been developed for generating educational feedback. In addition, to investigate the cognitive and practical effects of our approaches, a course was designed in which students had meaningful interactions with intelligent robots in an immersive environment. The results showed no significant difference in listening skills, but the speaking skills showed a marked improvement. Also, it demonstrated that the systems we have developed 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 or 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.

REFERENCES

[1] 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.
[2] 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.
[3] 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.
[4] H. Morton and M.A. Jack, "Scenario-based spoken interaction with virtual agents," Computer Assisted Language Learning, vol. 18, 2005, pp. 171–191.
[5] A. Raux and M. Eskenazi, "Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges," InSTIL/ICALL Symposium 2004, 2004.
[6] 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.
[7] S. Lee, H. Noh, J. Lee, K. Lee, G.G. Lee, S. Sagong, M. Kim. On the Effectiveness of Robot-Assisted Language Learning, ReCALL Journal, vol. 23(1), 2011.
[8] M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle," Readings in machine translation, 2003, p. 351.
[9] S. Lee, H. Noh, K. Lee, G.G. Lee. "Grammatical error detection for corrective feedback provision in oral conversations," Proceedings of the 25th AAAI conference on artificial intelligence, 2011.
[10] M. Ostendorf, P.J. Price, S. Shattuck-Hufnage, "The Boston University radio news corpus," Technical Report ECS-95-001, Boston University, 1995.
[11] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg, "ToBI: a standard for labeling prosody," Proceedings of International Conference on Speech and Language Processing, 1992, pp. 867-870.
[12] C. Li, J. Liu, S. Xia, "English sentence stress detection system based on HMM framework," Applied Mathematics and Computation, vol. 185, 2007, pp. 758-768.
[13] 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 Second Language Studies Workshop, Tokyo, 2010.