Joining-in-type Humanoid Robot Assisted Language Learning System

AlBara Khalifa, Tsuneo Kato, Seiichi Yamamoto
Graduate School of Science and Engineering, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto, 610-0394, Japan
Email: albara.khalifa@gmail.com, seyamamo@mail.doshisha.ac.jp, tsukato@mail.doshisha.ac.jp

Abstract

Dialogue robots are attractive to people, and in language learning systems, they motivate learners and let them practice conversational skills in more realistic environment. However, automatic speech recognition (ASR) of the second language (L2) learners is still a challenge, because their speech contains not just pronouncing, lexical, grammatical errors, but is sometimes totally disordered. Hence, we propose a novel robot assisted language learning (RALL) system using two robots, one as a teacher and the other as an advanced learner. The system is designed to simulate multiparty conversation, expecting implicit learning and enhancement of predictability of learners' utterance through an alignment similar to "interactive alignment", which is observed in human-human conversation. We collected a database with the prototypes, and measured how much the alignment phenomenon observed in the database with initial analysis.

Keywords: CALL, robot assisted language learning, interactive alignment, speech recognition

1. Introduction

Today's globalization produces much more opportunities of communicating in second languages (L2s) than ever. Learning second languages (L2s) is getting important for a large number of people. As a convenient and economic self-learning method, computer assisted language learning (CALL) gathers high interests. Though the CALL systems started from a fixed and passive style, interactive CALL systems, such as translation game type (Rayner et al., 2010; Wang & Seneff, 2007) and dialogue game type (Seneff & Wang & Zhang, 2004; Brusk & Wik & Hjalmarsson, 2007; Ito et al., 2008), are actively studied in accordance with the advance of automatic speech recognition (ASR). The interactive CALL systems have advantages of encouraging learners to construct utterances on their own and giving corrective feedback.

Such flexible interactive CALL systems are able to motivate learners much more than fixed and passive systems as well. To make the interactive CALL systems simulate human-human conversation more realistically, robot assisted language learning (RALL) has been proposed. Using a robot instead of a computer helps to add the dimension of using different modalities into the interaction. The modalities, such as gestures, nodding, face tracking, etc. raise the level of the experience closer to real life situation. Some studies verified the effectiveness of RALL in motivating learners (Lee et al., 2011), especially kids (Chang et al., 2010; Fridin, 2014; Keren & Fridin, 2014). If the RALL systems cover a wide range of topics, learners can train their conversational skill in L2 on the topics of their interest (Wilcock & Yamamoto, 2015).

However, ASR of L2 speech is still a challenge because the speech is often made with poor pronunciation and contains grammatical errors.

On this issue, CALL or RALL systems usually limit the range of possible learner utterances, for example, by setting the task as translation or dialogue in a predefined domain (Raux & Eskenazi, 2004). Regarding this issue, a phenomenon called "interactive alignment" occurs in human-human conversations. The "interactive alignment" is an unconscious process where both interlocutors tend to align to each other's utterances on different linguistic levels. We have collected a corpus of three-party conversations in L1 and L2, and observed similar alignment phenomenon that the subjects mimic the last utterance made by other subjects in L2 conversations (Yamamoto et al., 2015).

Hence, we propose a novel RALL system with two robots leveraging the alignment phenomenon. One robot plays a role of a teacher, and the other plays a role of an advanced learner, who makes sample answers, and even helps the human learner.

We assume the human learner to mimic the sample answer made by the robot learner, which turns out to get a high degree of predictability for ASR of L2 speech. We prototyped the RALL system with two robots, collected database with Japanese learners, and verified the alignment phenomenon occurred in the database.

The remainder is configured as follows. The design of multiparty RALL system is explained in Section 2. Database collected with the prototype system is described in Section 3. Initial analysis on the alignment phenomenon in the database is introduced and discussion is made in Section 4 and 5. Some ideas for future work is mentioned in Section 6. The paper concludes in Section 7.

2. The Design of Multiparty RALL System

2.1. Overview of the system

The design is a simulation of multiparty human-human conversation that was conducted previously (Yamamoto et al., 2015). Two robots are placed on a table in front of a learner, one playing a role of a teacher, and the other of an

advanced learner, both conducting a conversation with a human learner. The conversational learning is designed in a question and answer style, where one robot acts as the higher proficiency participant and asks all questions to both the robot learner and to the human learner.

In the multiparty human-human conversation, a phenomenon called "interactive alignment" is observed. "Interactive alignment" is an unconscious process that interlocutors tend to use the same expression as their conversation goes on.

Similar phenomenon of the “interactive alignment” is expected to occur when the human learner try to mimic the utterance of the advanced learner role playing robot.

In order not to make the conversation fail from errors of ASR for the L2 speech, the prototype system is handled in "Wizard of Oz" method from a remote PC. That means, the actions of the robots are not automatic, but rather manually controlled by an experimenter, who is operating the control program during the experiment in the same room without letting the learners to notice that. This paper is eventually aimed to enhance ASR, however, it was not used in the initial stage.

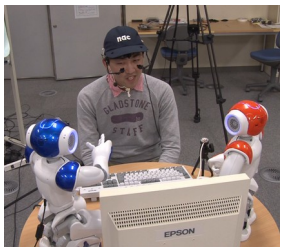

Figure 2: Experimental Setup V1 (Front Camera)

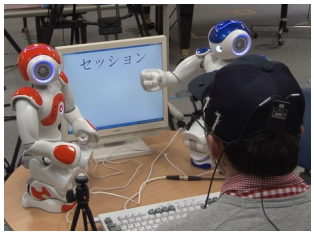

Figure 1: Experimental Setup V1 (Back Camera)

We developed two versions of the prototype system, taking the possibility of having no answer from the human learner into consideration. The first version has an additional PC monitor to show the learner a translation of a sample answer to the question in his/her mother language as a hint. Snapshots of the first version are shown in Figures 1 and 2. The second version entrust the robots more with helping the human learner, removing the PC monitor from the table. Snapshots of the second version are shown in Figures 3 and 4. We think that a more natural way to approach this issue is to let the robots repeat the question or even say a sample of the answer. The second version contains a longer scenario, and more chances for the learner to participate. More expressive hand gestures and body movements are used in this version, too. The gestures and movements are a built-in feature of the robot. The learners were asked to participate naturally to the conversation by answering the questions given to him/her.

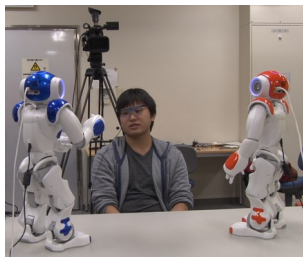

Figure 3: Experimental Setup V2 (Front Camera)

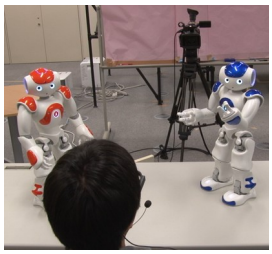

Figure 4: Experimental Setup V2 (Back Camera)

2.2. Scenario

As explained in the former section, the conversation is designed to induce the alignment phenomenon, and to leverage them to enhance the predictability of the learners' utterance for ASR. The questions are first asked to the robot playing the role of the advanced learner, then the same questions are asked to the human learner so that he/she can learn from the robot learner and answer in a similar way. Table 1 Shows an example of the conversational scenario used in the experiment.

R1:	What do you want to take when you go to a mountain?
R2:	If I have to go to a mountain, I will take a tent and a knife
R1:	How about you? What do you want to take when you go to a mountain?
Learner (example 1):	mountain! ... If I go to the mountain, I will have bring the rope
Learner (example 2):	I take to mountain... I go to mountain take Knife
R1:	Good answer

Table 1: Example of a conversational scenario. Where R1 is the robot that play the role of a teacher, and R2 is the second robot that play the role of an advanced learner.

3. Data Collection

3.1. Experimental setup

We recruited learners, and let them experience both of the versions of the experiment. We used Aldebaran NAO humanoid robots, and they were controlled using a Python program running on a remote PC. Beside using it to show translation of sample answers, the PC monitor were utilized to show instructions about the different parts of the conversation in the first version of the experiment.

3.2. Participants

The total number of the learners was 51, between the ages of 18 and 24. They are Japanese university students who had acquired Japanese as their L1 and had learned English as their L2. A part of the participants had taken the Test of English for International Communication (TOEIC), and

their score ranged from 400 to 890 (990 being the highest attainable score).

Feature	*Value*
Total number of experiments	51 experiments (30 for V1 21 for V2)
Number of omitted experiments	6 experiments
Number of video recording for every experiment	4 for V1 3 for V2
Approximate average length of every video recording	3 minutes for V1 10 minutes for V2
Number of sentences expected to be uttered by the learner	2 for V1 21 for V2
Number of questionnaire questions	42

Table 2: Corpus features

3.3. Corpus creation

The corpus created in this experiment benefited from the previously created corpus of multiparty human-human conversations (Yamamoto et al., 2015) in the experimental setting and the conversational scenario. This experiment considered human-robot conversations in L2 only. Video recordings were collected from every camera (three cameras in the first version of the experiment, and two only in the second version). The cameras were taking front view of the learner, and a back view which faced the robots. The first version of the experiment had a side view camera too. The front view camera was connected to a microphone, which was attached to the head of the learner and positioned near his/her mouth. An eye tracking system was used to capture the gaze of the learner throughout the experiments. In the first version of the experiment, NAC EMR-9 was used and was set on a cap worn by the learner. In the second version we used Tobii Pro Glasses 2 system which was more advanced and it was eye glasses worn by the learner. There are 24 recordings (six experiments' data were omitted due to technical problems) of about 3 minutes each from the first version of the experiment, while the second version has 21 recordings, each has a length of about 10 minutes. All learners signed an agreement of collecting their data to be used for research purpose in the lab. A questionnaire of 42 questions were answered by every participant and stored in an electronic format (i.e., a Google Sheet). Table 2 shows some features of the corpus.

The utterances of the learners needed to be transcribed into text format, in order to analyze them. The annotation software EUDICO Linguistic Annotator (ELAN) was used to transcribe the utterances of the learners, and to annotate the gazing activities. Figure 5 shows a snapshot of the annotation in ELAN.

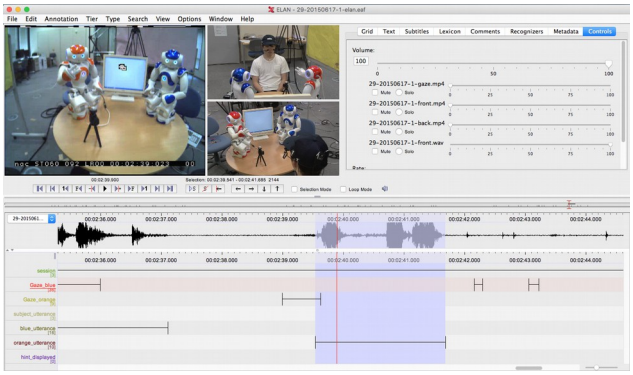

Figure 5: Snapshot of ELAN software

The different video and audio recordings were annotated on a frame-by-frame manner to find the start and end time of every utterance and every gaze action. The utterances of the human learner were manually transcribed to text format so it can be compared with the text that was synthesized by the robot during the conversation.

Average Word Similarity	*MED*	*M-MED*
Overall	41%	58%
In version 1	57%	59%
In version 2	36%	56%

Table 3: Word level alignment ratio of utterances data using the minimum edit distance algorithm (MED) and the modified version of minimum edit distance algorithm (M-MED).

4. Analysis

We used dynamic programming to find the similarity between learner's transcribed utterances (in text format) and robot's transcribed utterances on a word level. However, in many cases, the human learners were skipping the conditional phrase of the utterance of the robot and gave a response by mimicking the part of the sentence that gives a short answer. For example, the answer “If I have to go to a mountain I will take a rope” was answered by the human learner as “I will take a rope”, skipping the conditional phrase at the beginning of the answer, which is still a correct response but would not have a high similarity value using the minimum edit distance algorithm. To consider such cases, a modification of the algorithm was designed by freeing the start point of the robot's utterance during the comparison in a way to give better analysis. Other variations of the minimum edit distance algorithm have similar kind of modification and can also be used in this case, like the minimum edit distance with block operations (Shapira & Storer, 2003). Table 3 shows the average similarity of utterances between the human learner and the robot.

5. Discussion

Alignment of utterances between robots and humans can be noticed in the data, where the similarity between the

utterance of the human learner and the robot learner indicates a tendency in the human learner to mimic its utterance, or the important part of it. This may mean that the system could convey hints to the human learner in a natural and indirect way, since the human learner seemed to learn from the robot learner without external instruction.

Calculating the similarity of utterances using minimum edit distance may not be the best method in this case, since it is concerning only the syntax of the utterance. In real life, having 100% similarity all the time is not natural, and if it occurs, it should not be an indication of learning the language. In addition to that, the modified version of minimum edit distance algorithm considered the occurrence of the conditional phrase at the beginning of the sentence, while it may not be the case all the time. Handling syntactic similarity may not be the best choice in this case and the semantic comparison of the utterances should be considered.

We also made preliminary experiments of what features affect word similarity between utterances by learners and more advanced learners to obtain knowledge for designing language model of ASR (Tanizoe et al., 2016). The experimental results suggested that some learners tend to give an answer in more familiar syntactic structure to the non-native learner (e.g., "I will give a scarf to a girlfriend.") when the advanced learner gives an answer in some syntactic structure unfamiliar to the non-native learner (e.g., "I will give a girlfriend a scarf."). In designing language model of ASR such modification adopting similar syntactic structure should be considered.

6. Future Work

The implicit learning used in the experiment can be applied to cover different conversational scenarios with a wide variety of topics which can be extended easily from the system. Learners can train their conversational skill in L2 on the topics of their interest (Wilcock & Yamamoto, 2015).

Additionally, it is important to investigate how the learner obtain lexical and syntactic knowledge when he/she hears utterances in the similar syntactic structure from the perspective of pedagogy. We will, therefore, create a corpus of collecting speech data of the learners in a conversational scenario in which the advanced learner repeats the utterances in the similar syntactic structure in a series of answers to questions from the robot playing the role of the teacher.

We are also considering to conduct a semantic similarity calculation for the utterances in order to have a better autonomous responses by the system. We are planning to build a language model for the system to be fully automated and for the ASR to work properly in such difficult case that involve non-native language detection.

Finally, more dynamic adaptation can be added to the system in accordance with the learner responses and interaction, like the speech rate of the robots, or the choice of the level of linguistic difficulty in the scenario.

7. Conclusion

We proposed a novel language learning model based on observations that an interlocutor tends to produce speech by mimicking utterances from interlocutors of higher L2 proficiency. This was noticed in the calculated similarity between the utterance of the human learner and robot learner. The system consists of two humanoid robots that are manually controlled to chat with each other and with a human learner. The results of this experiment can help to enhance the predictability of the ASR used in the system in future work and can allow the use of wide variety of topics in the conversation.

8. Bibliographical References

Brusk, J., Wik, P., and Hjalmarsson, A. (2007). DEAL: A Serious Game for CALL Practicing Conversational Skill in Trade Domain. *The proceedings of SLaTE-Workshop on Speech and Language Technology in Education*. Pennsylvania, USA.

Chang, C. W., Lee, J. H., Chao, P. Y., Wang, C. Y., & Chen, G. D. (2010). Exploring the Possibility of Using Humanoid Robots as Instructional Tools for Teaching a Second Language in Primary School. *Educational Technology & Society*, 13(2), 13-24.

Fridin, M. (2014). Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Computers & Education.* 70, 53-64.

Ito, A., Tsutsui, R., Makino, S., & Suzuki, M. (2008, September). Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system. *In Ninth Annual Conference of the International Speech Communication Association*. Brisbane, Australia. 2819-2822.

Keren, G., & Fridin, M. (2014). Kindergarten Social Assistive Robot (KindSAR) for children's geometric thinking and metacognitive development in preschool education: A pilot study. *Computers in Human Behavior.* 35, 400–412.

Lee, S., Noh, H., Lee, J., Lee, K., Lee, G. G., Sagong, S., & Kim, M. (2011). On the effectiveness of robot-assisted language learning. *ReCALL*, 23(01), 25-58.

Raux, A. & Eskenazi, M. (2004). Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges. *STIL/ICALL Symposium.* Venice, Italy.

Rayner, E., Bouillon, P., Tsourakis, N., Gerlach, J., Nakao, Y., and Baur, C. (2010). A multilingual CALL game based on speech translation. *Proc. Proceeding of LREC*. Valetta, Malta. http://archive-ouverte.unige.ch/unige:14926.

Seneff, S., Wang, C., & Zhang, J. (2004). Spoken conversational interaction for language learning. *STIL/ICALL Symposium.* Venice, Italy.

Shapira, D., & Storer, J. A. (2003, October). Large edit distance with multiple block operations. *String processing and information retrieval* (pp. 369-377). Springer Berlin Heidelberg.

Tanizoe, T., Ishida, M., Okonogi, K., Khalifa, A., Umata, I., Kato, T., Yamamoto, S. (2016). Features Analysis of Interactive Alignment in Multiparty Conversations, *Proc. IEICE Japan*, (in Japanese).

Wang, C., & Seneff, S. (2007, April). Automatic Assessment of Student Translations for Foreign Language Tutoring. *Proc. Proceedings of NAACL/HLT* (pp. 468-475). Rochester, NY.

Wilcock, G., & Yamamoto, S. (2015). Towards Computer-Assisted Language Learning with Robots. Wikipedia and CogInfoCom. *Accepted by CogInfoCom*.

Yamamoto, S., Taguchi, K., Ijuin, K., Umata, I., & Nishida, M. (2015). Multimodal corpus of multiparty conversations in L1 and L2 languages and findings obtained from it. *Language Resources & Evaluation*, 49(4), 857-882. DOI 10.1007/s10579-015-9299-2.