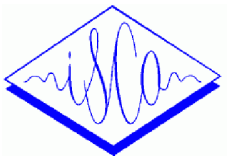

Fourth Workshop on Child
Computer Interaction (WOCCI 2014)
Singapore, September 19, 2014

Performance of a trialogue-based prototype system for English language assessment for young learners

Keelan Evanini, Youngsoon So, Jidong Tao, Diego Zapata-Rivera, Christine Luce, Laura Battistini, and Xinhao Wang

Educational Testing Service
Princeton, NJ 08541 USA
{kevanini,yso,jtao,dzapata,cluce,lbattistini,xwang002}@ets.org

Abstract

This paper describes a trialogue-based system for assessing the spoken language abilities of young learners of English. Specifically, the system employs spoken dialogue system components in interactive, conversation-based assessment tasks involving the test taker and two virtual interlocutors. The tasks are designed to be engaging for young learners of English at the elementary school level by incorporating real-life situations into the conversations and by providing immediate feedback about their spoken responses. The system was deployed in a data collection experiment with 18 young learners of English in a public school in the USA (grades 3 - 5) from a variety of language backgrounds. The system was able to produce overall task completion scores for each participant that correlated with scores based on human annotations at a rate of $r = 0.803$. In addition, the participants' responses to a user experience survey indicate that most participants felt that the virtual interlocutors understood their responses. This prototype demonstrates that this approach to using interactive, conversation-based assessments is a viable method of assessing the English language skills of young learners.

Index Terms: spoken dialogue systems, English language assessment, trialogue, young learners

1. Introduction

The number of English Language Learners (ELLs) in public schools in the USA has been on the rise in recent years, and was estimated to be 4.4 million students (9.1% of the total student population) in the 2011-2012 school year.[1] Some individual states have much higher proportions of ELL students, such as California, where approximately 23% of public school students are ELLs. With such a large number of ELL students, there is an increasing need for English Language assessments that can be used for screening and summative purposes in the K-12 domain. Most current assessments for ELL students consist of either multiple choice questions or one-on-one interviews with teachers. Multiple choice questions are less than ideal, since they do not provide a valid assessment of the conversational English skills that are required for classroom communication. Individual interviews, on the other hand, are time consuming, and require a large amount of educator resources to administer. Therefore, there is a clear need for assessments for young learners of English that provide a valid measure of their conversational speaking abilities and that can be scored efficiently. In this paper, we describe an interactive assessment based on trialogues (conversations between a test taker and two virtual characters) for young ELL students that attempts to meet these two goals.

The two main research questions in this study are as follows:

- Research Question 1: How well do standard components of spoken dialogue systems perform with input from young, non-native speakers of English?
- Research Question 2: To what extent can the interactive, trialogue-based tasks be used to assess the communicative competence of young, non-native speakers of English?

The remainder of this paper is organized as follows: first, Section 2 reviews relevant studies that have employed spoken dialogue system components for English learning and assessment as well as literature that motivated the use of the trialogue-based task design; next, Section 3 describes the design principles behind the interactive tasks that the test takers were exposed to in the study; then, Section 4 describes the methodology that was used for processing the participants' spoken responses as well as the design of the study; Section 5 presents the empirical results that address the two research questions outlined for this study; finally, Section 6 discusses the significance of the results and Section 7 provides some indications about recommended future research directions.

2. Literature Review

Most current assessments of spoken language proficiency (both automated and human-scored) are based on the stimulus-response model, in which the test taker is first presented with stimulus materials (such as an image, video, reading passage, recorded conversation, etc.) and is then prompted to provide a spoken answer in response to a question about the stimulus materials; crucially, each prompt from the system is not based on the preceding response provided by the test taker, and, thus, the assessment is not interactive. Automated systems for scoring these types of assessments involving young learners of English have been shown to achieve promising performance levels, both for tasks eliciting restricted speech, such as reading a text out loud, and those eliciting spontaneous speech, such as summarizing a lecture [1, 2]. However, none of these assessment systems have included interactive tasks that are able to evaluate a language learner's conversational skills.

In order to address this lack of interactivity, the current study investigates the use of spoken dialogue system components for assessing English language proficiency. There have

[1] http://nces.ed.gov/programs/coe/pdf/coe_cgf.pdf

been several studies that have employed components of spoken dialogue systems in the related domain of foreign language learning to develop interactive tasks for improving various aspects of a language learner's proficiency. For the most part, however, the linguistic skills evaluated through these tasks have been limited to areas such as pronunciation (e.g., [3]) and vocabulary (e.g., [4]), and have not evaluated conversational skills that are necessary for interactive communication (although see [5] for an example of an interactive language learning task involving role-playing and problem solving with an automated agent and [6] for a system that assesses cultural skills that are necessary for successful second language communication). The goal of this project is to move beyond these relatively restricted types of tasks and design an interactive system that can be used to assess a language learner's communicative competence through their ability to participate successfully in an interactive conversation.

Specifically, this study focuses on using *trialogues*, defined as interactive conversations between a test taker and two virtual characters, in spoken English proficiency assessment tasks. Trialogues, in comparison to standard two-party dialogues, provide more opportunities for the test taker to assume different conversational roles based on their level of understanding, and they facilitate more naturalistic corrective feedback and scaffolding from the virtual interlocutors. Trialogues have been used previously to create engaging, realistic scenarios in which the test taker is positioned to possess more or less information relative to each of the two virtual characters. For example, [7] utilized trialogues between a student and two virtual characters (another student and a teacher) to identify evidence of student inquiry skills. In this particular setting, the student is expected to help the less knowledgeable virtual student while receiving feedback or scaffolding from the other virtual character (i.e., a virtual teacher) who is expected to have more knowledge. A similar type of task was designed for the current study with the goal of creating conversations that are meaningful and engaging for young English learners, since these characteristics are crucial in designing tasks for teaching and assessing young learners [8, 9, 10].

3. Task Design

The tasks were designed to measure the English proficiency of students approximately 8-11 years old learning English in countries where English is not used as a first language. The language skills targeted in the tasks were identified among learning objectives specified in English curricula for elementary-aged students in several countries where English is taught as a foreign language (e.g., Brazil, China, Korea, Japan, and Mexico). The identified Listening, Reading, and Speaking skills were then designed to be measured through a series of language tasks embedded within language use contexts. The contexts of a school, specifically a classroom and the school library, were appropriate to create language use contexts which the target student population is likely to encounter in their lives. In order to better simulate language use in the real life, tasks were purposely designed for students to engage in multiple English skills (e.g., listening and speaking) in a meaningful, integrated way rather than measuring the skills discretely.

Each student participated in four trialogue-based spoken conversations during the course of the data collection session (the session also included selected response tasks as well as text-based interactions; the data from these modules will not be analyzed in this paper). The first two conversations took place in a classroom environment (Classroom1 and Classroom2) and the second two took place in the school library (Library1 and Library2). An example conversation from the Classroom1 conversation in Table 1 illustrates how the two skills (i.e., listening comprehension, and participating in a conversation by understanding and appropriately responding to short questions) were integrated within one conversation. In this example, a virtual peer (Ron) asks the test taker (Student) about instructions that were provided by the teacher while Ron was away from the classroom. In order for the student to participate in a conversation with Ron and provide a pragmatically appropriate answer to his question, the student needs to first understand the instructions that were presented by the teacher. (See [11] for a more complete description of the trialogue materials used in this study.)

Character	Utterance
Ron	*What are we learning about today?*
Student	*Weather.*
Lisa	*Yes, but it's not about any weather. You need to tell Ron more.*
Ron	*What are we learning about the weather?*
Student	*Weather around the world.*
Lisa	*That's right. We're learning about the weather around the world.*

Table 1: Sample trialogue in the Classroom1 conversation

In each of the conversations, a rule-based dialogue management strategy was employed to handle three different categories of student responses indicating how completely the student answered the original question. These three categories are defined in Table 2 along with the general type of system behavior designed for each type of response.

Category	System Behavior
Correct	acknowledge that the answer is correct and move on to the next task
Partially Correct	acknowledge that the answer is partially correct and ask a follow-up question to encourage the student to elaborate on the original answer
Incorrect	indicate that the answer is not correct and re-prompt

Table 2: Categories of student responses in each conversation

In the example listed in Table 1, the student's original answer (*Weather.*) was a partially correct answer, which triggered a follow-up question from the second virtual peer (Lisa) asking the student to provide additional information. After this second question, the student was able to provide a fully correct answer based on the ability to understand the request for the additional information. All of the four conversations included in this study follow this general dialogue framework, although the specific number of semantic slots that are mapped to each category in each dialogue state can vary (for example, there may be multiple slots for the Incorrect category based on the various types of possible mistakes that the test taker could make, and the specific feedback provided by the system about the incorrect response can vary accordingly). The maximum number of

student attempts at answering the question in each conversation was limited to two; if the student was still unable to provide a fully correct answer after the follow-up question, one of the virtual peers presented the correct answer to the student and the system moved on to the next task.

4. Methodology

This section presents details about the Automatic Speech Recognition (ASR) system that was used to process the spoken responses provided by the students in the trialogue-based conversations, the architecture that was used for dialogue management and Natural Language Understanding (NLU), the participants in the small-scale data collection study that deployed a prototype version of the system, and the user experience survey that the participants took after completing the session.[2]

4.1. Speech Recognition

The CMU PocketSphinx system [13] was used as the ASR component in the trialogue system. PocketSphinx was selected because it is available as open source and becasue its lightweight architecture results in fast ASR decoding times, a crucial quality in an ASR system deployed in an interactive application. The front-end of PocketSphinx features voice activity detection functionality which was used to preprocess the audio files submitted by the client application.

Language Model (LM) training was conducted using spoken responses obtained from a prototype deployment of the system with 20 participants in a separate data collection effort conducted in 2013. These 20 participants did not overlap with the 18 participants whose results will be described below, but their demographic characteristics were matched (i.e., they had similar age and native language profiles). Four separate trigram language models were trained for each of the four trialogue-based conversations included in the session. The number of responses from each conversation available for training is listed in Table 3.

Conversation	Responses
Classroom1	37
Classroom2	28
Library1	29
Library2	30
Total	124

Table 3: Amount of training data used for LM training and AM adaptation

For live data collection, we used the baseline Acoustic Model (AM) distributed with PocketSphinx, which was built using the Wall Street Journal and English Broadcast News Speech (Hub4) corpora. In order to examine the effect of improved ASR performance on the system's NLU, we also trained an adapted acoustic model using a combined MLLR and MAP adaptation approach [14] with the data shown in Table 3 (a total of approximately 62 minutes of audio). This adapted ASR system was not used in the live data collection, but its performance was evaluated in a post-hoc study.

[2]See [12] for a description of a system that used different spoken dialogue system components to implement the same tasks.

4.2. Natural Language Understanding

Dialog flows designed according to the principles described in Section 3 were created for each of the four conversations using a dialogue authoring tool based on the AutoTutor Script Authoring Tool [15]. Natural Language Understanding in each dialog state was conducted using a hybrid approach consisting of both hand-crafted regular expressions and Latent Semantic Analysis (LSA) [16]. The final semantic slot for each utterance was determined by combining scores from the regular expression and LSA matching procedures and comparing the scores to manually tuned thresholds.

4.3. Participants

Eighteen 3rd-5th grade students classified as English language learners by state criteria were recruited from a public school in a Northeastern state in the United States. Student background information on gender, grade, first language, and English proficiency, as judged by the English as a Second Language (ESL) teacher in the school, was collected at the time of recruitment. Seven male and eleven female students participated in the study, which included three 3rd graders, six 4th graders, and nine 5th graders. The majority of students, 11 of 18, were Spanish speaking students, with the remaining seven students speaking four other languages as their first languages (two Arabic, two Hindi, two Tamil, and one Telugu). The years of learning English varied substantially from 2 years to 10 years (mean = 5.80; standard deviation = 1.82), and their English proficiency was towards the higher continuum, with one student being intermediate, ten advanced-intermediate, and seven advanced. The seven students of advanced proficiency have recently exited the ESL classes.

4.4. User Experience Survey

A survey consisting of nine questions was administered at the end of the data collection session to investigate student perceptions about the tasks, and to later examine any relationship between student perceptions and their performance on the tasks. Out of the nine questions, the following two questions are particularly relevant to the focus of the present study.

- Question #8: *You felt that the people on the computer understood what you said.*
- Question #9: *The people on the computer said something that did not make sense to you.*

The students were asked to indicate their degree of agreement for each of the statements on a 4-point Likert scale, representing Strongly Agree, Agree, Disagree, and Strongly Disagree. The responses were coded from 1 (Strongly Disagree) to 4 (Strongly Agree) for statistical analyses.

5. Results

5.1. System Performance

In this section, we present the performance results for the ASR and NLU components of the system. ASR performance is measured in terms of the standard metrics, Word Error Rate (WER) and Sentence Error Rate (SER); NLU performance is measured by first mapping the semantic slots produced by the NLU system to the three categories defined in Section 3 (*Correct*, *Partially Correct*, *Incorrect*) and then comparing this category to

the category assigned to each response by human annotators[3] (all responses were included in the analysis, even if a preceding response in the conversation was misclassified by the system). This approach to evaluating the NLU performance is more meaningful than measuring the accuracy at the slot level, since the type of dialog act performed by the system in response to the test taker's utterance depends on which of these three categories the utterance is mapped to (see Table 2).

First, Table 4 presents the ASR results for each of the four conversations, as well as overall. As the table shows, the responses in the Classroom2 and Library2 conversations were easier for the system to process correctly than the responses from the Classroom1 and Library1 conversations; this is because the students produced more variation in the content of the responses in the Classroom1 and Library1 conversations (the perplexity values for the responses in the four scenarios are as follows: Classroom1 = 4.55; Classroom2 = 2.33; Library1 = 10.71; Library2 = 2.64).

Conversation	Responses	WER	SER
Classroom1	33	0.656	0.879
Classroom2	25	0.433	0.520
Library1	27	0.648	0.963
Library2	32	0.385	0.781
Overall	117	0.557	0.795

Table 4: ASR performance for the deployed system

The ASR performance varied widely for the 18 individual students who participated in the study. While the overall mean WER was 0.557, the minimum WER for an individual participant was 0.193, the maximum WER was 0.931, and the standard deviation was 0.220. After performing combined MLLR+MAP adaptation on the AM adaptation set described in Section 4.1, the overall WER was reduced to 0.496.

Next, Table 5 presents the NLU accuracy for three different text versions of each response: the manual transcription, the ASR using the baseline system that was deployed for the live data collection, and the ASR system with the adapted acoustic model. The results based on the transcriptions and the adapted ASR system were obtained by sending the responses to the NLU module after the live data collection.

Conversation	Responses	Trans.	ASR	ASR Adapted
Classroom1	33	90.9%	72.7%	72.7%
Classroom2	25	96.0%	88.0%	84.0%
Library1	27	100%	88.9%	88.9%
Library2	32	96.9%	81.3%	84.4%
Overall	117	95.7%	82.1%	82.1%

Table 5: NLU accuracy based on transcriptions, ASR from the deployed system, and ASR from the adapted system

As Table 5 shows, the NLU module was able to assign the correct category to nearly all of the responses when the transcriptions were used, with an overall accuracy rate of 95.7%. As expected, the overall NLU performance dropped when the ASR output was used, with an overall accuracy of 82.1% using both the live ASR output and the adapted AM. The ASR WER for each participant correlates with the NLU accuracy for that participant (using the baseline ASR output) at a rate of $r = -0.610$ ($N = 18$, $p < 0.01$).

Finally, an overall task completion score was assigned to each of the participants based on how many of the responses that they provided during their entire session were mapped to each of the three categories listed in Table 2. For each speaker, the overall task completion score was calculated by first converting the categories to numeric scores as follows: *Correct* = 3, *Partially Correct* = 2, *Incorrect* = 1. Then, the average score was computed across all responses provided by each participant in the four trialogue-based conversations contained in the session. This overall score thus provides a global indication of each participant's ability to participate successfully in the conversational tasks and provide correct answers to the questions. Then, these overall scores were correlated with the corresponding overall scores computed based on the manual annotations for the three categories to evaluate the performance of the system at predicting a participant's conversational ability. The correlation of these overall task completion scores was $r = 0.915$ based on the transcriptions and $r = 0.803$ based on the ASR output.

5.2. Survey Results

Student responses to the survey questions were generally very positive. Table 6 summarizes the responses to the two questions that gauged the student's perception of the system's level of understanding: Q8 (*You felt that the people on the computer understood what you said.*) and Q9 (*The people on the computer said something that did not make sense to you.*)

Question	Strongly Agree	Agree	Disagree	Strongly Disagree
Q8	2	12	3	1
Q9	0	2	8	8

Table 6: Results from the user experience survey for the two questions concerning the participants' perceptions of the system's level of understanding

The results in Table 6 indicate that 14 out of 18 students felt that the characters understood what they said, and 16 out of 18 students disagreed that the characters said something that did not make sense. These findings are encouraging in that students were given the impression that the characters understood them and reacted appropriately to what they had said. This is a necessary condition for the students to be engaged in the tasks in a meaningful way.

6. Discussion

With respect to Research Question 1, the current study shows that the performance of the open-source PocketSphinx system has a relatively high overall WER (0.557) on this dataset including spoken responses from young, non-native learners of English. This is not surprising, given that the Language Model training set only included approximately 30 responses for each conversation, and the fact that no Acoustic Model adaptation was conducted for the deployed system. However, the NLU results with an overall accuracy of 82.1% indicate that the dialogue engine is relatively robust to ASR errors. This is likely

[3]The annotation was conducted by authors of the paper who have an in-depth knowledge of the design of the system.

because many of the ASR errors occur on less important words in the conversation (such as function words instead of key content words) that are not relevant for the regular expression and LSA matching techniques employed by the NLU module. The post-hoc ASR adaptation experiment demonstrated that a small set of in-domain spoken responses (totalling approximately one hour of audio) was able to reduce the overall WER from 0.557 to 0.496. Somewhat surprisingly, this reduction in WER did not lead to a corresponding increase in NLU accuracy for the system with the adapted AM. Again, this is likely due to the fact that the improvements to the ASR performance were focused on less important words in the participants' utterances.

In order to investigate the relationship between actual system performance and the student perceptions of performance, correlations were calculated between the system's NLU accuracy for each speaker and their responses to the two user experience survey questions described in Section 4.4; neither of these correlations were significant. This somewhat surprising result could be explained by the fact that the dialogue flows and system responses were designed to be relatively robust to ASR and NLU errors (by limiting the number of turns per dialogue and by providing system responses that can be pragmatically appropriate for a wide variety of input utterances), so that participant responses with incorrect NLU classifications may not have had a substantial effect on the participant's experience. Secondly, the low correlations may merely be the reflection of the narrow distribution of the survey responses and the small number of participants in the study. However, it should be mentioned that the two students sharing the lowest overall NLU accuracy score among the 18 participants (57.1%) responded with Strongly Disagree and Disagree to Q8, which suggests that the poor NLU performance for those participants did have a negative effect on their perception of the system's understanding.

7. Conclusion

In this paper, we described a novel system for using trialogue-based, interactive tasks to assess conversational skills of young learners of English enrolled in public schools in the USA. While the results demonstrate that the approach has promise for assessing a young learner's communicative competence in English through overall task completion rates across the conversations, there is still much more research and development that needs to be done to demonstrate the feasibility of using this approach on a larger scale. First a follow-up with a larger number of students is required to demonstrate the validity of the results in a population with additional first language backgrounds and a wide range of proficiency profiles. In particular, it will be important to inlcude a larger number of students with limited English proficiency in a subsequent study, in order to evaluate the validity of diagnostic information about an ELL student's proficiency provided by the system. This experiment is currently being planned, with the aim of improving the ASR and NLU performance by incorporating the responses from the 18 participants included in the current study as additional training data.

In addition, future studies of the trialogue-based assessment should experiment with allowing the participants to provide more attempts at answering the questions in each conversation. In the current study, the total number of attempts was limited to two in order to have more control over the conversation flow and to reduce the chance that the students would become frustrated after multiple incorrect responses. However, given the relatively high NLU performance of the current system, future studies should employ additional feedback strategies to determine whether some participants would be able to eventually provide correct responses with sufficient scaffolding. Finally, the conversation-based approach should be extended to include more open-ended task types. The current system was designed for young learners of English, since tasks with relatively constrained responses are more appropriate for speakers with lower proficiency; however, more complex trialogues requiring additional language skills should be developed in order to be able to demonstrate that ths approach to assessing communicative compentance can also be valid for more proficient speakers.

8. References

[1] K. Evanini and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types," in *Proceedings of Interspeech*, 2013.

[2] J. Cheng, Y. Z. D'Antillio, X. Chen, and J. Bernstein, "Automatic assessment of the speech of young English learners," in *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA9)*, 2014.

[3] P.-H. Su, T.-H. Yu, Y.-Y. Su, and L.-S. Lee, "NTU Chinese 2.0: A personalized recursive dialogue game for computer assisted learning of Mandarin Chinese," in *Proceedings of the Interspeech Workshop on Speech and Language Technology in Education*, Grenoble, France, 2013.

[4] C. J. Cai, R. C. Miller, and S. Seneff, "Enhancing speech recognition in fast-paced educational games using contextual cues," in *Proceedings of the Interspeech Workshop on Speech and Language Technology in Education*, Grenoble, France, 2013.

[5] S. Seneff, C. Wang, and C.-Y. Chao, "Spoken dialogue systems for language learning," in *Proceedings of NAACL-HLT*, 2007.

[6] W. L. Johnson and A. Valente, "Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures," in *Proceedings of the 20th National Conference on Innovative Applications of Artificial Intelligence*. New York, NY: AAAI Press, 2008.

[7] A. C. Graesser, M. A. Britt, K. K. Millis, P. Wallace, D. F. Halpern, Z. Cai, K. Kopp, and C. Forsyth, "Critiquing media reports with flawed scientific findings: Operation ARIES! A game with animated agents and natural language trialogues," in *Intelligent Tutoring Systems (2)*, ser. Lecture Notes in Computer Science, V. Aleven, J. Kay, and J. Mostow, Eds., vol. 6095. Springer, 2010, pp. 327–329.

[8] L. Cameron, "Challenges for ELT from the expansion in teaching children," *ELT Journal*, vol. 57, no. 2, pp. 105–112, 2013.

[9] A. Hasselgren, "Assessing the language of young learners," *Language Testing*, vol. 17, no. 2, pp. 337–354, 2005.

[10] P. McKay, *Assessing Young Language Learners*. Cambridge, UK: Cambridge University Press, 2006.

[11] Y. So, D. Zapata-Rivera, Y. Cho, C. Luce, and L. Battistini, "Using trialogues to measure English language skills," *Educational Technology & Society (Special Issue on Technology Supported Assessment in Formal and Informal Learning)*, forthcoming, March/April 2015.

[12] C. M. Mitchell, K. Evanini, and K. Zechner, "A trialogue-based spoken dialogue system for assessment of English language learners," in *Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS), Napa, USA*, 2014.

[13] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "PocketSphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proceedings of ICASSP*, 2006.

[14] S. Goronzy and R. Kompe, "A combined MAP + MLLR approach for speaker adaptation," in *Proceedings of the Sony Research Forum*, vol. 99, 1999.

[15] S. Susarla, A. Adcock, R. V. Eck, K. Moreno, and A. Graesser, "Development and evaluation of a lesson authoring tool for AutoTutor," in *AIED 2003 Supplemental Proceedings*, 2003, pp. 378–387.

[16] S. D'Mello and A. Graesser, "Design of dialog-based intelligent tutoring systems to simulate human-to-human tutoring," in *Where Humans Meet Machines*, A. Neustein and J. Markowitz, Eds. Springer, 2013, pp. 233–270.