INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Spoken Dialogs With a Virtual Science Tutor

Wayne Ward, Daniel Bolanos, Ronald Cole

Boulder Language Technologies, Boulder, Colorado, USA
{wward,dani,rcole}@bltek.com

Abstract

My Science Tutor (MyST) is an intelligent tutoring system designed to improve science learning by elementary school students through conversational dialogs with a virtual science tutor in an interactive multimedia environment. Marni, a lifelike 3-D character, attempts to elicit self-expression from students, process their spoken explanations to assess understanding, and scaffold learning by asking open-ended questions accompanied by illustrations, animations or interactive simulations. MyST uses automatic speech recognition, natural language processing and dialog modeling technologies to interpret student responses and manage the dialog.

Index Terms: spoken dialog, virtual tutors

1. Introduction

My Science Tutor (MyST) is an intelligent tutoring system intended to provide an intervention for 3rd, 4th and 5th grade children who are struggling with science [1]. In our study, MyST was used as a supplement to normal classroom instruction using the Full Option Science System (FOSS), an inquiry-based science program that is widely used in US elementary schools. Students learn science in MyST through natural spoken dialogs with Marni, a lifelike 3-D computer character that is on screen at all times. Marni asks students open-ended questions related to illustrations, animations or interactive simulations displayed on the computer screen. We call these conversations with Marni *multimedia dialogs*, since students simultaneously listen to and think about Marni's questions while viewing illustrations and animations or interacting with a simulation. The media facilitate dialogs with Marni by focusing the student's attention, by helping them visualize the science they are discussing, and by enabling them to talk about and try to explain the science they were seeing. During tutoring sessions, Marni engages students in natural spoken dialogs based on principles of Questioning the Author (QtA) [2]. Each 20-minute tutorial dialog is oriented around a set of key concepts the student is expected to have learned from classroom instructional activities. MyST utilizes automatic speech recognition, character animation, robust semantic parsing, and dialog modeling to support conversations with Marni, as well as the integration of multimedia content into the dialogs. The student's computer shows a full screen window that contains Marni, a display area for presenting media and a display button that indicates the listening status of the system. Marni produces accurate visual speech, head and face movements that are synchronized with her speech.

2. QTA dialog management strategy

QTA was designed to improve comprehension of texts that are discussed as they are read aloud in the classroom. The focus is to have students reflect on what an author is trying to say in order to build a representation from the text. The approach uses open-ended questions to initiate discussion (What is the author trying to say?) to help students focus on the author's message (That's what she says, but what does she mean?) and to help students link information (How does that fit with what the author already told us?). The focus is on getting students to explain concepts in their own words and then to reflect on their explanation. Static illustrations, simple animations and interactive animations are used to focus student attention and provide a frame of reference for discussion with questions like *So, what's going on here?* Tutorial dialogs are designed to get students to articulate concepts and be able to explain processes underlying their thinking. The goal of a tutorial session is to elicit responses from students that show their understanding of a specific set of points. Two QtA strategies used in MyST are *marking* and *revoicing*. They require that the system identify the student's dialog content (marking it) followed by repeating (revoicing) a paraphrase of the information back to the student as a part of the next question; e.g., *You mentioned that electricity flows in a closed path. What else can you tell me about how electricity flows?*

3. Implementing tutorial dialogs

The Phoenix spoken dialog system [3] is used to implement the dialogs. Each tutorial session in MyST is designed to cover a few main points (2-4) in a 15 to 20-minute session with a student. For the system (Marni), the goal of a tutorial session is to elicit responses from students that show their understanding of a specific set of points, or more specifically, to entail a set of propositions. Marni's behavior in a dialog with a student, including the presentation of media within dialogs, is controlled by a *task* file. The *task* file contains the definition of the semantic frames to be used by the application. A task frame is a data object that contains all of the information available to interact about the frame:

- Frame Elements – the extracted information
- Templates for generating responses
- Pattern-Action pairs, called Rules, for generating responses contingent on certain conditions in the context.

By default, Marni will attempt to elicit speech to fill the Frame Elements representing the propositions of a frame. A sequence of interface actions is generated to elicit a response. The set of interface actions used are: flash(), movie(), show(), clear(), speak() and synth(). An example action sequence would be *flash(Components); speak(Tell me about that.)*. This sequence would run the Flash file *Components* and would have Marni speak the contents of a recorded audio file. In order to elicit speech to fill a frame element, the system developer specifies a list of action sequences for the element. During a session, the

Dialog Manager (DM) keeps count of how many times each element has been prompted for and uses the next action sequence in the list. Once it has exhausted the list, it gives the element the value FAIL, and will move on.

The tutorial developer may also specify a set of Rules for the frame. Rules are pattern-action pairs that can be used to generate action sequences conditioned on features of the context. Rule pattern definitions are Boolean expressions based on element values in the context. If the rule evaluates to true, one of the action sequences following it are sent to the interface manager. Like when prompting for an element, the system keeps count of the number of times a rule has been used and uses the next sequence each time. Figure 1 shows an example frame with a rule. The tutor would initially try to elicit information about flow direction by showing an animated Flash file named *Flow* and having the agent say *Tell me about what's going on here*. If the student responded with *it goes from plus to minus,* where the direction of electrical flow reversed, the parse would be

```
Frame: FlowDirection
 [DirFlow]
   Action: flash(Flow); speak(What's going on here.)
   Action: speak(What do you notice about the flow?)
[DirFlow].[Origin]
   Action: flash(Flow); speak(which side of the battery is the electricity coming from?)
[DirFlow].[Destination]
   Action: flash(Flow); speak(which side of the battery is the electricity going to?)
Rules:
# Got direction backward
([DirFlow].[Origin] == "pos") || ([DirFlow].[Destination] == "neg")
Action: flash(Flow); synth(Tell me again about the flow?)
Action: flash(Flow); synth(What direction is it going?)
```

Figure 1: *Example task frame*

[DirFlow].[Origin]: pos   [DirFlow].[Destination]: neg

The mapping of *plus* and *minus* to the canonical forms *pos* and *neg* is done by the parser. When the parse is integrated into context, the rule would fire and the tutor would continue to show the flash animation *Flow*, and the avatar would say "*Tell me again about the flow*".

4. Wizard-Of-Oz interface

. In order to gather and model data from effective multimedia dialogs of the sort we would like to create, we developed an interface to MyST that allows a remote human tutor to be inserted into the interaction loop. At each point in a dialog when the system is about to take an action, the action is first shown to the human wizard who may accept or change the action. The WOZ interface is a pluggable MyST component. If the Wizard is not connected, MyST sends the output straight to the user. If the Wizard connects to the session, MyST automatically sends actions to the Wizard for approval or revision. If the Wizard disconnects from the session, the system switches automatically to independent mode.

5. Speech recognizer

The speech recognizer is a large vocabulary continuous speech recognition system based on a Viterbi Beam Search [4]. Its trigram Language Models were trained using the CMU-Cambridge LM Toolkit [5]. Feature extraction from the audio was carried out using Mel Frequency Cepstral Coefficients (MFCC) plus the logarithm of the signal energy. Cepstral coefficients were extracted using a 20 ms window size and a 10 ms shift, which produces a feature vector each 10 ms that is composed of 12 MFCCs plus the log energy and the first and second order derivatives. Acoustic Models are clustered triphones based on Hidden Markov Models using Gaussian Mixtures to estimate the probabilities of the acoustic observation vectors. The system uses filler models to match the types of disfluencies found in applications. The recognizer can output word graphs, but MyST currently uses only the single best scoring hypothesis. The recognizer is configured to run in approximately real time so the delay after the student quits speaking and before the system is ready to respond is kept short. This is necessary to promote a fluent and engaging dialog.

6. Use of spoken responses

MyST does not use the information extracted from students' responses to grade students, and the system never tells the student that a response is wrong. Thus, the interaction style used in QtA is especially well suited to ASR-based systems. After each spoken response produced by a student, the system decides whether the current point should be discussed further, whether to present an illustration, animation or investigation accompanied by a prompt, or to move on to another point. In sessions where the system is able to accurately recognize and parse student responses, it is able to adapt the tutorial dialog to the individual student. It may move on as soon a student expresses an understanding of a point, or delve more deeply into a discussion of concepts that are not correctly expressed by the student. It may present more background material if the student doesn't seem to grasp the basic elements under discussion. If the system is unable to elicit student responses that fill any of the semantic roles related to the science concepts in a dialog, it will proceed through the session with a default tutorial presentation as specified in the *task* file.

In cases where the system understands the student, it is also able to apply *marking* and other techniques that use information from the student's response to generate a follow-on question. These dialog techniques are designed to assure the student that Marni is listening to and understands what the student is saying. Marni does not simply recognize and parrot back keywords spoken by the students. It represents the events and entities in the student's response, and it also represents the relations expressed between them, and communicates this understanding back to the student. The extracted representation is compared to the desired propositions to decide what action to take next.

Using spoken responses in this way provides a robust system interaction. False Negative errors by the system, in which the system misses correct information provided by the student, account for the bulk of Concept errors. In this case, the system simply continues to talk about the same point in a different way rather than moving on. When a False Accept error occurs, where the system fills in an element because of a recognition error, the system may move on from a point before it is sufficiently covered. Recapitulations by the system or errors by the student in later frames catch most of these. Thus, dialogs are designed to use speech understanding to increase efficiency and naturalness of the interaction while minimizing the impact of system errors.

7. Data collection

During the development process, a total of 1156 sessions were collected from 347 students. During the final year of the project, an assessment was conducted in which the system was run in stand-alone (fully automatic) mode. In the assessment, 988 sessions were collected from 118 students. Each MyST dialog session produces a set of speech files (from the student) and a log file that contains time-stamped entries for the events that occurred during the dialog. Manual transcription of the speech files is performed off-line and is introduced into the log file later. Some additional pieces of information stored in the log file are: : the output of the automatic speech recognition (ASR) system, the extracted frame elements, the current context, the frame name and frame element that is generating the system response, the number of times this frame element has been used, and the action sequence generated for the response. When operating in WOZ mode, the MyST system log includes all wizard-generated actions. The log records whether the wizard accepts each proposed system action, or how they changed it. The data collected during the WOZ phase of system development were used to re-train speech recognition and natural language processing models. Analysis of the log files gave insight into problems with tutorials and the need for additional multi-media resources or modifications to cause the system to behave more like the wizards.

8. System evaluation

During the 2010-2011 school year, the MyST system was evaluated by comparing learning gains of students who received tutoring sessions with either the virtual tutor Marni (MyST) or with human tutors in small groups. Students were randomly assigned within classrooms to the tutoring condition (virtual or human), and these groups were compared with students from intact control classrooms. Students completed one of four FOSS modules-- *Variables, Magnetism & Electricity, Measurement, and Water.* All students received similar classroom instruction. The hypotheses for the study were: 1) students in MyST and human-tutored groups would have roughly similar gains from pre to post test, 2) tutored students would have significantly greater gains than students in the control condition. The FOSS Assessing Science Knowledge (ASK) instruments were used to measure learning gains for each of the four modules in the study. The ASK assessments consist of identical pre and post versions with open-ended, short answer, multiple choice and graphing items administered before the beginning of the FOSS lessons, and immediately after classroom instruction and tutoring ended. Pairs of raters from Boulder Language Technology scored assessments from tutored students and control students. All scoring was blind to tutoring group. Inter-rater reliabilities for two raters were high (counting only the open-ended items), ranging from 0.89 to 0.98. Internal reliabilities were lower, ranging from 0.60 to 0.89 for both pre and post versions of the assessments. Scores used for outcome analysis were the averages across both raters.

Research was conducted at schools with students from a large range of socioeconomic and ethnic backgrounds. Eighty-three (83) students received MyST tutoring, 69 were human tutored (both in 12 classrooms) and 1015 students in 50 classrooms in 20 schools received only classroom instruction and no tutoring. Sixty-two (62) classrooms were included in the analysis. To make comparisons, outcome scores were converted to *Residual Gain Scores*, which compared groups on the average differences between their observed and expected scores. Additionally, residual gain scores were estimated and evaluated assuming and not assuming equal variances.

Direct comparisons of residual gain for the randomly assigned groups (MyST and Human Tutored) showed no significant differences between groups with $t = -1.14$, d$f = 150$, $p = 0.25$. The effect size, however, favored the human-tutored group. In the three-way comparison with the control group, MyST and human tutored groups had insignificantly different residual pre/post gains; the control students, on the other hand, had significantly less residual pre/post gains. A Univariate ANOVA (using scores standardized by module test) showed a main effect for tutoring condition with $F = 26.2$, d$f = (2, 1164)$, $p < 0.01$. Post-hoc tests showed no significant differences between MyST and human tutored groups; significant differences were found between MyST and the control group ($d = .53$), and human tutored students and the control group ($d = .68$). Differences in residual gain scores were also tested using hierarchical models with classroom used as a grouping variable. MyST students showed significantly higher scores than the controls ($t = 2.5$, d$f = 60$, $p = 0.014$), as did the human-tutored group when compared with controls ($t = 3$, d$f = 60$, $p < 0.01$).

9. Component evaluations

The data collected in the WOZ condition were used to train the models used by the speech and natural language processing modules in the assessment. The same sets of models were used throughout the assessment. In order to characterize how well the system understood student input, speech recognition and parser outputs were extracted from the assessment log files and evaluated against transcripts of the corresponding student speech. Transcripts were available for 587 sessions, which totaled 17,113 student utterances. These were used for the analysis of ASR and concept extraction performance.

The speech recognizer vocabulary size was 6235 words. There is a large variance in WER across students, but the rate is very stable across modules with an overall average of 38.7%. From previous studies using the WOZ data, we had expected a lower WER (of around 31%). In an effort to understand this difference, the test set perplexity and WER for the WOZ test set and assessment test set were compared for different training conditions. The results are shown in Table 1. The woz training condition used only the woz training data. The mix training condition used a training set that combined the woz and assessment training sets. For language models trained on woz data, the perplexity of woz test data was 45.9 compared to 87.8 for assessment test data. Even though the same material was covered, language usage was substantially different in the two conditions. Re-training the lm using the mixed training data reduced perplexity on assessment data to 68.4 and WER to 34.7. Training acoustic models on the mixed data and using a woz-trained lm gave a WER of 35.1 on the assessment data. Training both acoustic and language models on the mixed training set reduces WER on the assessment data to 27.7, indicating that both language usage and acoustic match are different between the woz and assessment data. Using the mixed-trained models, the WER for the WOZ data was 19.0 compared to 27.7 for assessment, indicating that the assessment data is substantially more difficult.

The behavior of the virtual tutor is more dependent on Concept Accuracy than on Word Error Rate. One way to measure the effect of speech recognition errors on the system is to look at the accuracy of extraction of frame elements, i.e., the accuracy with which concepts are extracted from speech. Grammars are created for each investigation using the training data. The tutorials (i.e., the individual tutoring sessions aligned to each classroom science investigations) have an average of 8 frames with an average of 5 frame elements per frame, thus there are about 40 frame element classes on average in an investigation. Reference parses were created for each hand transcribed utterance by parsing the transcripts. The speech recognizer output for the utterances was also parsed and Recall and Precision of frame elements were calculated compared to the reference parses. Recall is the percentage of the reference elements that were correctly extracted from the recognizer output. Precision is the percentage of the elements extracted from the recognizer output that were correct. These results are also fairly consistent across modules with an overall average of Recall = 76% and Precision = 80%.

There is considerable redundancy in some of the student speech. If a frame element is repeated, it doesn't matter that the system extracts all identical repetitions, as long as it gets at least one (and they are the same). An even better measure of understanding that affects the system's actions would be to estimate the extraction rate for unique frame elements. Redundant frame elements were deleted from both reference and hypothesis parses and the concept accuracy was computed using only unique elements, resulting in an overall Recall= 79% and Precision= 82%. So 79% of the relevant information in the reference parses was correctly extracted from the ASR output. Of the information extracted, 82% of the elements were correct. Despite a Word Error Rate of around 39%, the system understood an average of about 80% of the relevant information. Given the nature of QtA dialogs and the way speech input is used by the system, as described above, this level of concept identification accuracy was sufficient to produce both engaging and effective dialogs. This was evidenced by very positive survey results of students and teachers.

Train		Test			
Acoustic Models	Language Models	WOZ		Assessment	
		PP	WER	PP	WER
woz	woz	45.9	30.9	87.8	38.7
mix	mix	60.1	19.0	68.4	27.7
mix	woz	45.9	21.7	87.8	35.1
woz	mix	60.1	25.3	68.4	34.7

Table 1. *WERs for different training/test conditions*

10. Conclusions

This paper described the design of a conversational virtual tutor for elementary school science. Data were presented that show the system is significantly better at achieving learning gains that the business-as-usual control condition, and not significantly worse than trained human tutors.

The system performed well and was accepted enthusiastically by students, even though the WER was considerably larger than expected. Analysis showed the automatic (assessment) condition to be significantly more difficult than a woz condition. One reason for the robust system-level results is the way in which speech is used by the system. As concept accuracy improves, the system is able to adapt the presentation to the user and become more efficient and engaging, but good learning results do not critically depend on the word error rate. Using the Phoenix robust parsing techniques, about 80% of the relevant information was correctly extracted, even though the WER was 39%.

11. Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305B070008, and the National Science Foundation through Grant DRL 0733323. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education or the National Science Foundation.

12. References

[1] Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S.V., Becker, L., "My science tutor: a conversational multimedia virtual tutor for elementary school science", ACM Trans. Speech and Lang. Proc., 7(4), 2011.

[2] Beck, I., McKeown, M., Sandora, C., Kucan, L., and Worthy, J., "Questioning the author: A yearlong classroom implementation to engage students with text", *The Elementary School Journal, 96*(4):385-414, 1996.

[3] Ward, W., "Extracting information from spontaneous speech*", In Proceedings of the International Conference on Spoken Language Processing, 1994.*

[4] Bolanos, D., Cole, R., Ward, W., Borts, E., and Svirsky, E., "FLORA: Fluent Oral Reading Assessment of children's speech", *ACM Trans. Speech and Lang. Proc., 7(4), 2011.*

[5] Clarkson, P.R., and Rosenfeld, R., "Statistical language modeling using the CMU-Cambridge toolkit", In Proceedings ESCA Eurospeech 1997.