Adding Spoken Dialogue to a Text-Based Tutorial Dialogue System

Diane J. Litman

Learning Research and Development Center
and
Computer Science Department
University of Pittsburgh
Pittsburgh, PA 15260 USA
litman@cs.pitt.edu

Abstract

Spoken dialogue is a natural and highly desirable form of student-computer interaction, which provides both opportunities and challenges to the area of dialogue-based tutoring systems. This paper describes a new research project wedding research in spoken language technology and instructional technology, with the goal of promoting learning gains by enhancing communication richness.

Keywords: Spoken Dialogue, Prosody, Annotation, Evaluation

1 Introduction

As dialogue systems for tutorial applications become more common (Rose and Freedman, 2000), it will become crucial to increase their communicative effectiveness by taking advantage of advances in human language technology. For example, as evidenced by other research projects such as Project LISTEN (Mostow and Aist, 2001) and DC-TRAIN (Fry et al., 2001), the time is ripe for exploring whether spoken language interactions can and/or should replace the text-based interactions found in most dialogue-based intelligent tutoring systems. This paper describes a new research project which will wed research in spoken language technology with instructional technology, with the goal of promoting learning gains by enhancing communication richness.

Moving towards spoken dialogue systems for tutorial applications has many potential benefits. First, spoken language interaction is generally considered to be the most natural and easy to use form of natural language interaction. Second, a user’s speech contains prosodic information not present in text-based interaction, which could be used to monitor both the dialogue and pedagogical effectiveness of student-computer interactions. Third, going to a fully speech-based dialogue system is a natural complement to existing talking head tutors that already focus on spoken language output, e.g., (Graesser et al., 2001; Rickel and Johnson, 2000). The hands-free aspect of spoken language will also more easily support the addition of other interface modalities, for example pointing and clicking. Finally, because tutoring is a challenging and underexplored application area for the computational linguistics community, a spoken dialogue tutoring system is likely to not only advance the state of the art in tutoring, but also in spoken dialogue systems.

2 Research Questions

My new research project will entail building a spoken language interface to an existing text-based tutoring dialogue system, and conducting experiments with users in order to address the following set of research issues:

- What are the advantages – as well as disadvantages – of using a speech-based rather than text-based tutorial dialogue system?
• Can prosody be used to infer pedagogically significant information?
• Can the tutoring system make use of such inferences?
• Can existing robustness mechanisms deal with the noise from Automatic Speech Recognition?

3 Technical Approach

The research will enhance a text-based dialogue tutoring system now being developed at the University of Pittsburgh, by first providing and then exploiting spoken language capabilities. The research will be conducted in three phases (baseline system implementation and evaluation, data analysis, and enhanced system implementation and evaluation), and will need to address the workshop focus areas (evaluation, annotation, and creation/use of public resources). The research will build on the Why2 project, a collaboration between the University of Pittsburgh and the University of Memphis in the area of dialogue-based explanation tutoring (Graesser et al., 2002). In Why2, a student first provides a natural language answer to a qualitative physics problem. The tutor then engages the student in a natural language dialogue to provide feedback and correct misconceptions, and to elicit more complete explanations.

In Phase 1, a baseline spoken dialogue version of Why2 will be developed. This system will basically replace the text-based input and output modalities of the current version of Why2 with spoken language input and output. Because the use of automatic speech recognition will increase the noise at the input level, the system will also provide a new test for Why2's current robustness mechanisms. This spoken language baseline will then be empirically evaluated, providing a new experimental condition for examining dialogue and tutoring effectiveness. This will complement the existing evaluation efforts for Why2, which will compare the Pittsburgh text-based natural language dialogue system, a human tutoring condition, a text-based reading (i.e., non-dialogue) tutoring system, and the Memphis natural language tutoring system (which among other things includes a talking head). The evaluation will also generate a corpus of human-computer dialogues, which will be used to drive the next phases of research.

In Phase 2, a corpus analysis will be performed, which will be used to guide the development of an enhanced spoken dialogue version of Why2 in Phase 3. The enhanced version of the system will attempt to increase pedagogical effectiveness, by taking advantage of information that is only available in speech. We will manually annotate user utterances for potential pedagogical classifications of interest (e.g., annoyance, frustration, confusion, boredom, certainty, etc., as discussed below), and automatically label prosodic and other features that can be computed at runtime. We will then use machine learning to predict the classes from the automatically obtainable features.

In Phase 3 of the research, a new version of the system based on the corpus analysis will be implemented. This enhanced system will use the learned classifier to guide dynamic adaptation of the dialogue and pedagogical behaviors of the tutoring system. We will conduct empirical evaluations to compare the original and enhanced spoken language versions of Why2, with respect to student learning gains and dialogue system performance.

4 Previous Research and Implications for Tutorial Dialogue

My focus on discovering pedagogically-driven techniques for deciding when and how to intervene during intelligent tutoring will build heavily on my previous work in using prosodic and other automatically computable features to identify and recover from communication problems. Over the last few years, I have shown that features automatically computable at system run-time can be used to predict problems that can successfully benefit from some sort of system intervention. The hypothesis of my new work is that a similar approach will prove useful for predicting and adapting to problematic pedagogical situations.

---

1 Determining what can be annotated with high reliability will likely be a significant research effort in its own right. Until our human-computer spoken dialogue data becomes available, during Phase 1 we plan to conduct preliminary annotation experiments and refine our coding manuals using existing (text-based) Why2 dialogue data. We might also conduct a Wizard of Oz study to collect human-human spoken language data.
4.1 Non-Tutorial Dialogue

My previous work in non-tutorial dialogue focused on system detection of and adaptation to user problems at both the dialogue and utterance levels. At the dialogue level, our results were based on an examination of two different corpora, and used two different definitions of problematic dialogue situations. In a corpus of human-computer dialogues obtained from experiments with the TOOT train-timetable dialogue system, we used a machine learning methodology to identify problematic dialogue situations defined as dialogues where the percentage of correctly understand utterances was below some threshold. Using a set of 23 features representing five types of knowledge sources (acoustic, dialogue efficiency, dialogue quality, experimental parameters, and lexical), we were able to predict whether a dialogue was problematic with respect to speech recognition performance with 77% accuracy (Litman et al., 1999). In human-computer data from a deployed call routing system, we used a similar approach to predict problematic dialogues defined in terms of unsuccessful task completion (Walker et al., 2000a).

Using a set of automatic, task independent features representing three knowledge sources (acoustic, semantic, and dialogue), problematic dialogues with respect to task success could be identified with 87% accuracy, and predicted from just the first exchange in the dialogue with 72% accuracy.\(^2\)

We also showed the utility of predicting problematic dialogue situations for triggering system adaptation. Using a ruleset learned from training data, we modified the TOOT system to incrementally predict whether a user was having speech recognition problems as a dialogue progressed, and to adapt to a more conservative set of dialogue strategies whenever the user’s dialogue was predicted to be problematic. Our main result was that for a population of novice users, adding adaptation to TOOT significantly increased the task success rate from 23% to 65% (Litman and Pan, 2002).

While our initial work focused on predicting and adapting to problems at the (sub)dialogue-level, our more recent research with the TOOT corpus focused on modeling user problems at the utterance level, and investigating the types of system dialogue strategy changes that might be warranted after a single problematic utterance. In particular, we focused on predicting several different turn categories potentially useful for error handling: misrecognitions, corrections, and aware sites. We also greatly expanded our feature set for representing our dialogue data, to include prosodic features (pitch, loudness, turn duration, silence between turns, silence within turns, and speaking rate), more features derived from the speech recognition component and its inputs and outputs, as well as a set of history features for the dialogue as a whole and for more local dialogue contexts. Our machine learning results showed that the use of prosodic features - in conjunction with other features that were known or automatically detectable by a spoken dialogue system - yielded the best prediction performance. For example, our combined feature set accurately predicted speaker turns that were misrecognized - with 93% accuracy for predicting the presence of transcription errors, and 90% accuracy for predicting the presence of semantic errors (Litman et al., 2000).\(^3\) We also found that our features could be used to predict user corrections of system errors with 84% accuracy (Hirschberg et al., 2001), and aware sites – turns where the user first becomes aware that the system has made an error, with 89% accuracy (Litman et al., 2001). All of these accuracy rates were significant improvements over baseline figures.

Other research has shown the utility of adapting to user problems at the utterance level. A spoken dialogue system that automatically adapted initiative based on participant roles, features of the current utterance and dialogue history (Chu-Carroll, 2000) outperformed a non-adaptive version in terms of usability and efficiency (Chu-Carroll and Nickerson, 2000). Dynamically deciding whether to confirm each user utterance during a task-oriented dialogue was also shown to improve system performance (Smith, 1998).

4.2 Tutorial Dialogue

For my current work I plan to take a similar approach, but now predict and adapt to problematic pedagogical situations in tutorial dialogues. Perhaps the most direct application of our previous work would be to attempt to detect and correct “tutorial-level misrecognitions.” Even if there is perfect speech recognition, other components of the system can still fail. In the text-based version of Why2,
for example, there are cases where the student answers the system correctly, but the system believes that the student’s response was in fact incorrect (e.g., because the natural language component couldn’t handle the particular wording used by the student).

I also plan to pursue detecting and responding to certain affective states, motivated by recent work in both the tutoring and prosodic literatures. Evens has proposed detecting problematic affective student states (e.g., confusion, boredom, anger, and frustration) in the context of the Circsim-Tutor text-based dialogue system, with the goal of having the system then take steps to rectify the problematic situation (Evens, 2002). We hope to also pursue the detection and exploitation of such states. Initially we will focus on using the prosodic features examined in our earlier work for this type of prediction, building on current work exploring the use of prosody, sometimes in conjunction with other types of information, to detect emotional speaker states from speech (e.g., happy, afraid, angry, sad, or neutral (Polzin and Waibel, 1998); annoyance/frustration versus other (Shriberg et al., 2001); and emotional utterances such as anger and irritation that can signal “trouble in communication” and trigger a system action to prevent further breakdown (Batliner et al., 2001)). We will also look at related tasks such as detecting off-talk (Siepmann et al., 2001) and subjective language (e.g., expressing emotions, opinions, etc.) (Wiebe et al., 1999), which might also be useful for triggering system intervention.

Our longer term goal is to identify and rectify other types of potentially problematic pedagogical situations in tutorial dialogue (e.g., particular levels of student initiative (Jordan and Siler, 2002; Moore, 2002; Roque, 2002)), using our previous prosodic and other predictors.

5 Expected Contributions

Spoken dialogue is a natural and highly desirable form of student-computer interaction, which provides both opportunities and challenges to the area of dialogue-based tutoring systems. The opportunity arises from the presence of speech features such as prosody that serve as data for comprehending the nature of the student-computer interaction. The challenge is first in interpreting the speech itself and then in making productive use of those features.

The research and resulting technology from this proposal will hopefully lead to more natural and effective dialogue-based systems for tutoring applications. The major contributions will be 1) the development of a spoken dialogue tutoring system that can be empirically compared with text-based dialogue systems, and 2) an understanding of whether and how information from the speech signal can be used to predict pedagogically significant information that can be usefully exploited for learning gains.

Acknowledgements

Many thanks to Pam Jordan and to the reviewers of this paper for extremely helpful comments, and to Kurt VanLehn for his encouragement and help in undertaking this project. This research is supported by the National Science Foundation under Grant No. 9720359 to the Center for Interdisciplinary Research on Constructive Learning Environments (CIRCLE) at the University of Pittsburgh and Carnegie-Mellon University.

References


Jack Mostow and Gregory Aist. 2001. Evaluating tutors that listen: An overview of Project LiSTEN. In K. Forbus and P. Feltonvich, editors, Smart Machines in Education.


