Taking Turns Talking About Text in a Reading Tutor that Listens

Greg Aist

4215 Newell-Simon Hall
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, Pennsylvania 15213 USA

Abstract

In this paper we report on ongoing work on turn-taking in Project LISTEN's Reading Tutor (Mostow & Aist CALICO 1999). Project LISTEN's Reading Tutor listens to children read aloud and helps them learn to read. The Reading Tutor's repertoire of turn-taking behaviors includes not only alternating turns, but also backchanneling, interrupting, and prompting.

Turn-taking in conversation

Human-computer spoken dialog relies on taking turns: knowing not just what to say, but when to say it. Humans take turns in many different ways: alternating with another speaker, backchanneling, taking another turn if there is a pause in the conversation, interrupting, and pausing. Spoken dialog systems began by assuming a very strict model for turn-taking: alternating turns. Telephone-based dialog systems were among the first to allow the user to interrupt, or "barge-in". As human-computer spoken conversation extends into new domains and new social roles, computers should be able to take turns using the full, rich set of turn-taking behaviors that humans engage in.

A Reading Tutor that Listens

Project LISTEN's Reading Tutor represents a special case of spoken dialog: A single task – teaching reading – but on text from any domain. The Reading Tutor allows users to type in and narrate their own stories, which can then be used to teach children to read (Mostow & Aist USPTO 1999). The Reading Tutor uses such recorded human speech for its expressiveness and warmth. Turn-taking behaviors are generated using on a combination of user-added resources such as story narrations, and lexical resources such as word recordings, recordings of syllables, recordings of letters, and other recorded prompts and phrases. Turn-taking behaviors are selected using a combination of reading-specific heuristics – such as "sound out a word only if it has four or fewer phonemes" – and task-independent heuristics such as "take a turn if the user falls silent for an extended period of time."

Turn-taking in the Reading Tutor

Following Sacks et al. (1974), Cassell et al. (1994), Thorisson (1996), and Ward

(1996), we describe turn-taking actions as heuristic rules with left-hand preconditions based on dialog state and right-hand actions consisting of multimodal output.

We divide the turn-taking state using two binary variables:

USER-TURN Is the user taking a turn?
SYSTEM-TURN Is the system taking a turn?

We use two variables instead of a single "whose turn is it?" variable for two reasons.
1. Separate variables allows us to represent our uncertainty about whether the user is taking a turn separately from whether the system is taking a turn.
2. In human-human conversation, turns *typically* overlap. The assumption that one participant is speaking at a time "reflects ideology more than practice" (Tannen 1994, page 62). Thus machine interlocutors should model overlap, including cases where both participants take extended turns cooperatively or competitively.

This two-variable framework lets us represent a wide variety of turn-taking behaviors. Of particular note are prompting, backchanneling, and interrupting. The Reading Tutor prompts in response to extended student inactivity. The Reading Tutor backchannels in response to a slight pause. Finally, the Reading Tutor can interrupt to correct a student mistake. We operationalize the concept "interrupt to correct a mistake" as follows:

If USER-TURN=true and SYSTEM-TURN=false and the user has misread an important word and continued to the next word, then cough and underline a word.

We introduced these turn-taking behaviors and the architecture underlying them into the Reading Tutor beginning in 1996-97. We have used a variety of methodologies to evaluate these turn-taking actions, and improve them.

To evaluate and improve the robustness of the turn-taking architecture, we tested the revised Reading Tutor in progressively more difficult environments. We started with adult users in the laboratory and ended up with children using the Reading Tutor on their own in a classroom without the researchers present, supervised by a teacher who was teaching the rest of the class.

To evaluate and improve the quality of the interactions, we invited critique of videotaped interaction by an expert in reading instruction. The expert was asked to view videotape of the Reading Tutor interacting with a student and judge the Reading Tutor's responses.

To explore the local effects of on the dialog, we introduced systematic variation in the Reading Tutor's responses as an "invisible experiment" (Mostow & Aist 2000), similar to varying dialog choices when using reinforcement learning to learn dialog policies (Singh et al. 1999). For example, to look for the local effect of interrupting, we modified the Reading Tutor to sometimes interrupt when its heuristic for interrupting applied, and sometimes remain silent. Analyzing what happens next in the dialog – for example, whether the student speaks again, clicks *Go* to move on, or does something else – allows us to quantify the local effects of conversational actions with respect to what would have happened if another action had been taken instead.

Conclusion

We have described some of the progress that spoken dialog systems have made towards taking turns like humans, enumerated some contributions made towards that goal within Project LISTEN's Reading Tutor, and described some of the evaluation methodologies appropriate to conversational behaviors.

What remains to be done? To expand turn-taking capabilities into novel areas, the field of dialog research needs to include user-system dialogs with varied social roles. Many systems focus on some form of information seeking (train timetables, movie times) or assistance in constructing a plan (travel agent, military logistics). We have looked at a less explored area in human-computer spoken dialog: tutoring. Other social roles – an adversary or a partner in a game, or a computerized salesperson that tries to maximize company profit rather than end-user satisfaction – may offer new insights. Much interesting work remains to build dialog systems that take turns as well as humans.

Acknowledgements

Jack Mostow, Alex Rudnicky, and Nancy Green supervised the master's project on which this work is based. Project LISTEN is a collaborative effort whose continued success depends on many people's hard work.

This material is based upon work supported in part by the National Science Foundation under Grant Nos. CDA-9616546, REC-9720348, and REC-9979894, and by the author's NSF Graduate Fellowship and Harvey Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government.

References

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M. 1994. Animated Converstaion: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. Proceedings of SIGGRAPH '94. (ACM Special Interest Group on Graphics). http://justine.www.media.mit.edu/people/justine/siggraph94.ps

Mostow, J., & Aist, G. 2000. Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbes & P. Feltovich (Eds.), forthcoming book on AI and education. AAAI Press.

Mostow, J. and Aist, G. CALICO 1999. Giving Help and Praise in a Reading Tutor with Imperfect Listening: Because Automated Speech Recognition Means Never Being Able to Say You're Certain. CALICO Journal 16:3, 407-424. Special issue (M. Holland, Ed.), Tutors that Listen: Speech recognition for Language Learning.

Mostow, J. and Aist, G. USPTO 1999. Reading and Pronunciation Tutor. United States Patent No. 5,920,838. Filed June 2, 1997; issued July 6, 1999. US Patent and Trademark Office.

Sacks, H., Schegloff, E. A., and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50(4): 696-735.

Singh, S., Kearns, M. S., Litman, D. J., & Walker, M. A. 1999. Reinforcement learning for spoken dialogue systems. In Proceedings of NIPS*99, to appear as S. A. Solla, T. K. Leen, and K.-R. Müller, (Editors), Advances in Neural Information Processing Systems 12. Cambridge, MA: MIT Press.

Tannen, D. 1984. Conversational style: Analyzing talk among friends. Norwood NJ: Ablex.

Thorisson, K. R. 1996. Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. Ph.D. thesis, MIT Media Lab. http://kris.www.media.mit.edu/people/kris/thesis.html

Ward, N. 1996. Using Prosodic Clues to Decide When to Produce Back-channel Utterances. In Proceedings of the 1996 International Symposium on Spoken Dialogue, pages 1728-1731, Philadelphia PA.