Estimating the Effectiveness of Conversational Behaviors in a Reading Tutor that Listens

Gregory Aist and Jack Mostow

Project LISTEN (http://www.cs.cmu.edu/~listen)

215 Cyert Hall, Language Technologies Institute

Carnegie Mellon University

4910 Forbes Avenue

Pittsburgh, PA 15213

aist+@andrew.cmu.edu, mostow@cs.cmu.edu

Abstract

Project LISTEN's Reading Tutor listens to children read aloud, and helps them learn to read. Besides user satisfaction, a primary criterion for tutorial spoken dialogue agents should be educational effectiveness. In order to learn to be more effective, a spoken dialogue agent must be able to evaluate the effect of its own actions. When evaluating the effectiveness of individual actions, rather than comparing a conversational action to "nothing," an agent must compare it to reasonable alternative actions. We describe a methodology for analyzing the immediate effect of a conversational action, and some of the difficulties in doing so. We also describe some preliminary results on evaluating the effectiveness of conversational behaviors in a reading tutor that listens.

Introduction

The idea of getting computers to listen to children read has been around for years. However, early attempts [Bernstein and Rtischev 1991, Kantrov 1991, Phillips, McCandless, and Zue 1992] were hampered by slow hardware and the restrictions of isolated word recognizers. With the advent of affordable consumer-level machines that can recognize continuous speech in near-realtime, the time is ripe for reading tutoring as a testbed for conversational computing for children. Here we report on Project LISTEN's Reading Tutor [Mostow and Aist AAAI 1997, Mostow et al. UIST 1995, Mostow et al. AAAI 1994]. Related efforts are underway at, for example, IBM (1997) and DERA Malvern (Russell et al. 1996).

In 1996-1997, a pilot test of the Reading Tutor at an urban elementary school produced dramatic results [Mostow and Aist WPUI 1997]. The subjects were several third graders who started out reading nearly three years below grade level. They used the Reading Tutor for eight months, supervised individually by a school aide. Each student had 30-60 sessions, averaging 14 minutes per session. The six pre- and post-tested subjects advanced by an average of about two years in instructional reading level, defined as the grade level of material they could read with at least 75% accuracy and 75% comprehension. One student could read and comprehend sixth grade material, but another student showed little or no improvement.

The Reading Tutor is now deployed in classroom field studies in order to evaluate and improve it. We extended the Reading Tutor to provide a wider range of assistance, such as sounding out a word or giving a rhyming hint. We deployed this extended version of the Reading Tutor at a month-long Reading Clinic at the same school during July 1997, where 62 students used the Reading Tutor on eight Pentium Pro™ machines in a lab setting. Since August 1997, the Reading Tutor has been deployed in regular classrooms and used regularly by over 100 students. Students in these field tests have ranged from kindergarten through fourth grade, with a few fifth graders and one sixth grader. This wider range of age groups and ability levels offers new challenges for instruction, including how to make the Reading Tutor usable in ordinary classrooms with a minimum of adult assistance.

How can we improve the Reading Tutor to make it work better for all children? If the Reading Tutor could evaluate the effectiveness of its own actions, it could adjust

its behavior to work better overall, and even adapt to fit individual students.

How can a spoken dialogue agent monitor its own effectiveness? The PARADISE model for evaluating spoken dialogue agents [Walker, Litman, Kamm and Abella ACL 1997] predicts user satisfaction based on a number of factors including task success and total dialogue time. However, since the PARADISE model works over global measures of dialogue success, it is not clear how to assign credit or blame for user satisfaction to individual dialogue acts – just the agent as a whole.

Rather than user satisfaction, the ultimate measure of the Reading Tutor's success is student learning. By listening to children read, the Reading Tutor can estimate overall improvement in student performance unobtrusively and objectively [Mostow and Aist AAAI 1997]. In this paper, we estimate *local* effectiveness of tutorial interventions by looking at the differential accuracy before and after intervention opportunities – how a child reads a word before getting (or not getting) help on that word, and how the same child reads the same word just afterwards.

A Reading Tutor that Listens

Project LISTEN's Reading Tutor adapts the Sphinx-II speech recognizer [Huang et al. 1993] to listen to children read aloud. The Reading Tutor runs on a single stand-alone Pentium™. The child uses a noise-cancelling headset or handset microphone and a mouse, but not a keyboard. Roughly speaking, the Reading Tutor displays a sentence, listens to the child read it, provides help in response to requests or on its own initiative based on student performance. [Aist 1997] describes how the Reading Tutor decides when to go on to the next sentence.

The student can read a word aloud, read a sentence aloud, or read part of a sentence aloud. The student can click on a word to get help on it. The student can click on *Back* to move to the previous sentence, *Help* to request help on the sentence, or *Go* to move to the next sentence. The student can click on *Story* to pick a different story, or on *Goodbye* to log out.

The Reading Tutor can choose from several communicative actions, involving digitized and synthesized speech, graphics, and navigation [Aist and Mostow 1997]. The Reading Tutor can provide help on a word (e.g. by speaking the word), provide help on a sentence (e.g. by reading it aloud), backchannel ("mm-hmm"), provide just-in-time help on using the system, and navigate (e.g. go on to the next sentence). With speech awareness central to its design, interaction can be natural, compelling, and effective [Mostow and Aist WPUI 1997].

Fair Alternatives to Conversational Behaviors

If a conversational agent is to evaluate the effectiveness of its own behaviors, what should it compare a given behavior against? A key observation is that one cannot compare against "doing nothing." "Doing nothing" in a conversation is in itself an action, and if done too frequently or for too long, violates the unwritten rules of conversational behavior. For example, if an agent takes a turn because it is time to take a turn, it is unreasonable to compare the turn's effectiveness to the "alternative" of saying nothing when saying nothing might be seen as a system failure. Thus, a more reasonable methodology is to compare the effectiveness of a given behavior *against the effectiveness of other equally felicitous behaviors*.

Machine learning paradigms such as active learning or reinforcement learning [Kaelbling, Littman, and Moore 1996, Sutton and Barto, forthcoming] hold promise for agents learning behavior by active exploration of the environment. However, in mixed-initiative conversation, where a "strict turn-taking" assumption (T speaks, S speaks, T speaks, S speaks) does not hold, evaluating the effectiveness of conversational behavior presents additional challenges. First, time is not discrete as in many machine learning paradigms; it is continuous, and the time that an action took place may be just as important as when it took place. For example, consider a hypothetical case of two ways of giving the same kind of help, where one way causes the user to take twice as long to respond as the other. Clearly here time is an important factor in evaluating conversational behavior. How should two outcomes that occur at different times be weighted?

Secondly, more than one Reading Tutor intervention may take place before the student responds. How should individual Reading Tutor actions be credited if more than one of them takes place between student turns? For the purposes of this paper, we defer this issue by considering only whether any of the Reading Tutor's actions between student turns were relevant to a given word, without distinguishing among alternative series of such actions.

Estimating the Effectiveness of a Conversational Behavior

How can a conversational agent evaluate the effectiveness of a conversational behavior? In this paper, we describe how we evaluated the effectiveness of the Reading Tutor's conversational behavior ("intervention") based solely on the student's reading before and after the intervention in question. Assigning credit or blame in this (or some other) way is a necessary prelude to learning which

interventions work better for which students on which words.

Consider the real example history shown below, from a 9 year old boy reading a story about cheetahs. This multimodal dialogue is rife with overlapping by both the student and the Reading Tutor.

Tue Jul 15 09:24:47.093 Tutor displays:

Years ago there were more than one hundred thousand of them in parts of Africa and Asia

Tue Jul 15 09:24:47.218 - Tue Jul 15 09:24:48.125

Tutor says: 'thousand'

Tue Jul 15 09:24:50.015 - Tue Jul 15 09:24:53.203

Student says 'thousands <breath>…'

Recognized: THOUSAND

Tue Jul 15 09:24:52.640

Student clicks on 'Years'

Tue Jul 15 09:24:52.750 - Tue Jul 15 09:24:53.531

Tutor says 'Years'

Tue Jul 15 09:24:54.390 - Tue Jul 15 09:24:58.515

Student says 'Years ago they were'

Recognized: YEARS AGO ASIA WERE

(and so forth)

For a given student attempt at a sentence, words are categorized by the Reading Tutor as correct, misread, or omitted. When the Reading Tutor intervenes on a word, we consider for the purposes of this paper that its effect can be measured by looking at the difference between the student's prior reading of that word (if any) and the student's next reading of that word in the same sentence context (if any). Thus the sequence of conversational events given above generates the following word events.

Tue Jul 15 09:24:47.093 T displays 'Years'

Tue Jul 15 09:24:47.093 T displays 'ago'

Tue Jul 15 09:24:47.093 T displays 'there'

Tue Jul 15 09:24:47.093 T displays 'were'

Tue Jul 15 09:24:47.093 T displays 'more'

Tue Jul 15 09:24:47.093 T displays 'than'

Tue Jul 15 09:24:47.093 T displays 'one'

Tue Jul 15 09:24:47.093 T displays 'hundred'

Tue Jul 15 09:24:47.093 T displays 'of'

Tue Jul 15 09:24:47.093 T displays 'them'

Tue Jul 15 09:24:47.093 T displays 'in'

Tue Jul 15 09:24:47.093 T displays 'parts'

Tue Jul 15 09:24:47.093 T displays 'of'

Tue Jul 15 09:24:47.093 T displays 'Africa'

Tue Jul 15 09:24:47.093 T displays 'and'

Tue Jul 15 09:24:47.093 T displays 'Asia'

Tue Jul 15 09:24:47.218 T says: 'thousand'

(For simplicity we look at only the beginning of speech)

Tue Jul 15 09:24:50.015 S omits 'YEARS'

Tue Jul 15 09:24:50.015 S omits 'AGO'

… (S omits THERE WERE TEN)

Tue Jul 15 09:24:50.015 S reads correctly 'THOUSAND'

… (S omits OF THEM IN PARTS OF AFRICA AND ASIA)

Tue Jul 15 09:24:52.750 T says 'Years'

(We treat student-initiative and tutor-initiative help identically for the purposes of this paper.)

Tue Jul 15 09:24:54.390 S reads correctly 'YEARS'

Tue Jul 15 09:24:50.015 S reads correctly 'AGO'

Tue Jul 15 09:24:50.015 S misreads 'ASIA'

Tue Jul 15 09:24:50.015 S reads correctly 'WERE'

Tue Jul 15 09:24:50.015 S omits 'TEN'

… (S omits THOUSAND OF THEM IN AFRICA AND)

Tue Jul 15 09:24:50.015 S omits 'ASIA'

For each word, we define its *transition* with respect to an intervention as the tuple (first attempt, second attempt). We select those word events that relate to the word 'years', and show below the transitions, and the intervention(s) to which they are credited:

New sentence =No intervention on 'years'=> Omit word

Omit word =T said 'years'=>Read correctly

Experiment: 1997 Summer Reading Clinic

227,693 word transitions	…to new sentence		…to word correct		…to word misread		…to word omitted	
165,048 without word intervention		80,549		56,003		9,973		18,523
62,645 with word intervention		19,571		32,836		5,911		4,327
99,176 **from new sentence**								
72,845 without	40.41%	29,434	49.31%	35,918	5.66%	4,125	4.62%	3,368
26,331 with	29.01%	7,638	55.56%	14,630	9.82%	2,585	5.61%	1,478
87,945 **from word correct**								
59,787 without	63.28%	37,834	28.28%	16,910	3.11%	1,861	5.32%	3,182
28,158 with	36.59%	10,302	53.63%	15,100	6.03%	1,699	3.75%	1,057
15,715 **from word misread**								
10,637 without	48.12%	5,118	15.20%	1,617	31.58%	3,359	5.10%	543
5,078 with	26.68%	1,355	41.10%	2,087	29.93%	1,520	2.28%	116
22,762 **from word omitted**								
19,691 without	31.67%	6,237	7.88%	1,552	2.61%	514	57.83%	11,388
3,071 with	8.79%	270	33.18%	1,019	3.45%	106	54.58%	1,676
2,095 **from other event**								
2,088 without	92.24%	1,926	0.29%	6	5.46%	114	2.01%	42
7 with	85.71%	6	0.00%	0	14.29%	1	0.00%	0

Did the Reading Tutor help students correct errors? To assess the effect of the Reading Tutor's interventions, we analyzed the 227,693 word transitions from the summer 1997 reading clinic data, summarized in the table above.

Each word transition enumerated describes two successive events involving a word: seeing a word in a new sentence, reading a word correctly, misreading a word, omitting a word, or going on to a new sentence. The "other" category may include starting a new story.

The table categorizes word transitions based on whether the Reading Tutor did or did not intervene *on that word* between these two events. For the version of the Reading Tutor used for the 1997 Summer Reading Clinic, "word intervention" included a rich set of responses to student mistakes or missed words [Aist and Mostow CALL 1997]: reading the word, reading the sentence, sounding out the word, breaking down the word into syllables, giving a rhyming hint, and providing a word that starts the same. Conversely, "without intervention" included cases where the Reading Tutor said nothing, backchannelled, helped with a different word, or went on to the next sentence.

Several caveats are in order. First, the word transitions were computed based on automated speech recognition. We were able to analyze much more data than would be feasible to transcribe by hand, but at the cost of imperfect accuracy. Second, the "word intervention" class lumps all interventions together. We defer to future analysis a comparison of alternative interventions (and sequences thereof). Third, the table shows only immediate effects of intervention on reading a word in the context of the current sentence; we defer analysis of how intervention affects subsequent encounters of the same or related words. Finally, decisions to intervene were not random. Consequently the set of cases where the Reading Tutor intervened differed statistically from where it did not. In particular, one would expect cases where the Reading Tutor intervened to involve poorer reading, on average. (We plan randomized experiments to eliminate such bias.)

The table suggests a number of interesting observations. First, fewer than half the word transitions involve interventions on that word, according to our classification. For example, suppose the student reads a 10-word sentence, the Reading Tutor gives feedback on one word, and the student rereads the sentence. This scenario yields only 1 transition with a word intervention, but 9 without.

To assess the effects of word intervention compared to none, we must therefore compare transition frequencies rather than counts. For example, what happened after a word was misread? The next attempt was correct 41% of the time with a word intervention – but only 15% without. Likewise, attempts after omitted words succeeded 33% of the time with intervention, versus only 7.9% without. Some transitions to missed words are *more* frequent with intervention, ostensibly due to bias or going on less often.

Overall, the results seem to indicate that intervention on a misread or omitted word was indeed more effective than no intervention on that word. While the result itself merely confirms our strong suspicion, the methodology by which we obtained it is important: by analyzing (noisy) transcripts of students' oral reading automatically, we have been able to obtain preliminary results on local effectiveness of the Reading Tutor's behavior.

Conclusion

What does this paper contribute? We have discussed how educational effectiveness should be a key component of evaluation for tutorial spoken dialogue agents. We have argued that rather than comparing a conversational action to "nothing", an agent must compare it to reasonable alternative actions. We have described a methodology for inferring the immediate tutorial effect of a conversational action, and some of the difficulties in doing so. Finally, we have described some preliminary results on evaluating the effectiveness of conversational behaviors in a reading tutor that listens.

Acknowledgments

This material is based upon work supported by NSF under Grants IRI-9505156 and CDA-9616546, by DARPA under Grant F336159311330, and by the first author's National Science Foundation Graduate Fellowship and Harvey Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency or the official policies, either expressed or implied, of the sponsors or of the United States Government. We thank Fort Pitt Elementary School, our colleagues, and countless others who have helped Project LISTEN. We especially thank Dan Barritt, a CMU Human-Computer Interaction Institute Master's student, who ran the Project LISTEN portion of the 1997 Summer Reading Clinic at Fort Pitt.

References (see also www.cs.cmu.edu/~listen)

Aist, G. S. March 1997. Challenges for a mixed initiative spoken dialog system for oral reading tutoring. AAAI Spring Symposium on Computational Models for Mixed Initiative Interaction. AAAI Tech. Rept. SS-97-04.

Aist, G. S., and Mostow, J. September 1997. Adapting human tutorial interventions for a reading tutor that listens: Using continuous speech recognition in interactive educational multimedia. CALL 97: Theory and Practice of Multimedia in Computer Assisted Language Learning. Exeter, UK.

Bernstein, J., and Rtischev, D. 1991. A voice interactive language instruction system. *Proceedings of the Second European Conference on Speech Communication and Technology (EUROSPEECH91).* Genova, Italy.

Huang, X. D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F., and Rosenfeld, R. 1993. The Sphinx-II speech recognition system: An overview. *Computer Speech and Language* 7(2):137-148.

IBM. 1997. IBM Lead Story: Watch Me Read. http://www.ibm.com/Stories/1997/06/voice5.html.

Kaelbling, L.P., Littman, M.L., and Moore, A.W. 1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*. Volume 4, pages 237-285.

Kantrov, I. 1991. Talking to the Computer: A Prototype Speech Recognition System for Early Reading Instruction, Technical Report 91-3, Center for Learning, Teaching, and Technology, Education Development Center, 55 Chapel Street, Newton, MA 02160.

Mostow, J., and Aist, G. S. July 1997. The sounds of silence: Towards automatic evaluation of student learning in a reading tutor that listens. *Proceedings of Fourteenth National Conference on Artificial Intelligence*, 355-361.

Mostow, J., and Aist, G. S. October 1997. When Speech Input is Not an Afterthought: A Reading Tutor that Listens. Workshop on Perceptual User Interfaces, Banff, Alberta, Canada.

Mostow, J., Hauptmann, A., and Roth, S. November 1995. Demonstration of a reading coach that listens. *Proceedings of the Eighth Annual Symposium on User Interface Software and Technology (UIST 95)*, pp. 77-78. ACM SIGGRAPH, SIGCHI, & SIGSOFT, Pittsburgh, PA.

Mostow, J., Roth, S., Hauptmann, A. G., and Kane, M. August 1994. A prototype reading coach that listens. *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, American Association for Artificial Intelligence, Seattle, WA, pp. 785-792. Recipient of AAAI-94 Outstanding Paper Award.

Phillips, M., McCandless, M., and V. Zue. September 1992. Literacy Tutor: An Interactive Reading Aid. Technical Report, Spoken Language Systems Group, MIT Laboratory for Computer Science, MIT.

Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bohnam, B., and Barker, P. 1996. Applications of automatic speech recognition to speech and language development in young children. *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia PA.

Sutton, R. S. and Barto, A. G. Forthcoming. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press/Bradford Books. http://www.cs.umass.edu/~rich.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. 1997. PARADISE: A framework for evaluating spoken dialogue agents. ACL 1997, Madrid, Spain.