DIALOG SYSTEMS FOR SURVEYS: THE RATE-A-COURSE SYSTEM

Amanda Stent, Svetlana Stenchikova and Matthew Marge

Computer Science Department
Stony Brook University
Stony Brook, NY 11794-4400
amanda.stent@gmail.com, sveta@cs.sunysb.edu, mmarge@gmail.com

ABSTRACT

In this paper, we discuss why surveys are an interesting application of spoken dialog systems from both commercial and research perspectives. We then describe a prototype survey spoken dialog system, the Rate-A-Course system. We show how dialog epiphenomena, including the order in which questions are answered and the duration of respondents' answers, can be used to learn information beyond that covered explicitly by survey questions.

Index Terms— Speech communication

1. INTRODUCTION

Dialog systems technology is being applied to a wide range of applications. Commercial dialog systems exist for tasks including customer service [1] and travel [2]. There are also research dialog systems for tasks such as email access [3], information access [4], travel [5], tutoring [6], command-and-control [7], and planning [8].

Somewhat surprisingly, we know of no existing spoken dialog systems for conducting surveys. Surveys are a natural and commercially viable application for spoken dialog systems. Survey dialog systems also present interesting opportunities for research on spoken dialog and on survey design.

Surveys as Spoken Dialog Application The conducting of oral surveys is a highly controlled interaction with structured language use, well within the reach of current spoken language technology and with considerable commercial potential. For example, the Council of American Survey Research Organization (CASRO) estimates that the survey industry in the U.S.A. alone has over $6.7 billion in annual revenues [9].

The authors would like to thank Richard Gerrig who co-managed the design and collection of the Rate-A-Course experiment, Susan Brennan and Marie Huffman for discussions about the Rate-A-Course project, Randy Stein for helping collect subjects, and Daniel Evans for the first Rate-A-Course survey design interface. This research is based on work supported by the National Science Foundation under grant no. 0325188. 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.

Use of survey dialog systems could counteract some of the drawbacks of oral and electronic surveys while retaining the strengths of both. In particular, unlike traditional oral surveys, survey dialog systems involve no human interviewer costs, the results do not suffer from interviewer bias or other consistency issues, and data entry costs are reduced. Survey dialog systems are more interactive than web-based surveys and do not require respondents to have computers or internet access; furthermore, respondents are more likely to complete telephone surveys than web-based surveys [10]. There are potential disadvantages to survey dialog systems, including: the need to provide respondents with special instructions, respondents' attitudes towards computers affecting survey results, and the potential for system error. These can be minimized through the use of good dialog system design.

Interdisciplinary SDS/Survey Research Survey dialog systems could also lead to new research in spoken dialog systems and in survey design/analysis. Research questions include:

- How increased interactivity (e.g. clarification interactions) affects survey results [11]
- How dialog epiphenomena can be used to learn about the reliability of survey results or to improve survey design (e.g. pauses as an indicator of uncertainty, prosodic variation as an indicator of frustration/engagement)
- How best to automatically summarize and answer questions about survey results

Finally, through the collection of surveys involving open-ended questions it is possible to collect large corpora of free speech from a wide range of speakers. These corpora can be used to study spoken language, for example to improve speech recognition technology for multiple dialects.

2. THE RATE-A-COURSE SYSTEM

The Stony Brook Rate-A-Course system is a survey dialog system that permits college students to evaluate their courses

Topic	Synonyms	Answers/Ratings
Instructor	teacher, professor	very good/100, good/75, okay/50, bad/25, very bad/0
Exams	tests, midterms	too hard/0, hard/50, about right/100, easy/50, too easy/0
Class size	course size, size of the class	too packed/0, packed/50, about right/100, small/50, too small/0
Assignments	homeworks, course work	too hard/0, hard/50, about right/100, easy/50, too easy/0
Teaching assistant	t a, teachers assistant	very good/100, good/75, okay/50, bad/25, very bad/0

Fig. 1. Topics used in Rate-A-Course system experiment

System	S: We will now ask your opinion on the following aspects of your course: the instructor, the assignments and the exams. Is the instructor: very good, good, okay, bad, or very bad?
Mixed	S: Which topic was the next most important to you? Your choices are ...
User	S: Would you like to discuss another topic? U: Yes S: Which topic is the next most important to you? Your choices are ...

Fig. 2. System questions by initiative condition

over the telephone. Most courses are evaluated using written surveys. Some universities also use electronic surveys for midsemester course evaluations. Both types of survey suffer from low response rates, perhaps because students do not think that course evaluations are used by faculty or administrators [12]. Also, frequently course evaluation survey results, particularly the answers to open-ended questions which other students might find most useful, are not made available to other students for use during course selection [13]. As a result, students have created their own internet forums for exchanging comments on their courses.

The Rate-A-Course system is a prototype telephone-based spoken dialog system that could be used as a replacement for or adjunct to other course evaluation methods. The novelty factor of talking to a dialog system might increase response rates; because the survey results (including comments in response to open-ended questions) are available in electronic form, they can be distributed over the web or telephone.

Rate-A-Course System Implementation The Rate-A-Course system is implemented in VoiceXML, XML and Javascript. It runs on the BeVocal Cafe platform and uses its proprietary speech recognizer [14]. The survey questions, and potential answers, take the form of an XML document that can be automatically generated from a web-based survey design interface. This interface permits the selection of choice points (for subdialogs), question types and error-handling strategies. The XML document is used to automatically generate speech recognition grammars and to populate VoiceXML forms that act as templates for different question types. Javascript embedded in the VoiceXML forms permits automatic logging of all system and respondent interactions.

Possible answers to close-ended questions and question-related keywords taken from the XML document are used to create recognition grammars; these permit respondents to answer close-ended survey questions using full or partial sentences, using the terms specified in the question or using synonyms of question terms. In this version of the system, no attempt is made to automatically process the answers to open-ended questions during the survey.

The Rate-A-Course system permits respondents to ask for the last question to be repeated and to ask for help at any time. A request for help is interpreted as a request for clarification of the current question. The system also provides help on a recognition failure or no input; this help can be a simple repetition of the question, an explanation of the answers or an example answer, or a subdialog, depending on the XML specification for the survey. Respondents in the experiment described here were only allowed to go back or cancel for certain questions (e.g. course department and number).

The Rate-A-Course system generates structured logs in the form of question-answer pairs for all questions, as well as a complete dialog history with pointers to the audio files containing respondents' speech.

Adaptation in the Rate-A-Course System The Rate-A-Course system implements several different dialog behaviors leading to different amounts of system interactivity and adaptation:

- **Choice of question type:** In the XML document, survey designers can specify whether a question should be open-ended or close-ended and can specify valid answers to a close-ended question.
- **Question ordering:** The VoiceXML forms implement random ordering of questions when the survey designer does not specify question order.
- **Lexical adaptation:** The VoiceXML forms implement lexical adaptation, so that the system can adapt its choice of words and tense to the user's word choice.
- **Initiative:** There are VoiceXML forms for a survey with *system initiative* (the system chooses the question order), *mixed initiative* (the respondent chooses in which order to answer survey questions), or *user initiative* (the

respondent chooses which survey questions to answer, as well as the order in which to answer them).

The survey designer can give the respondent a code that prespecifies system behaviors. This means that in addition to collecting survey data, the system can be used to perform research about spoken dialog and survey design.

3. USER MODELING IN THE RATE-A-COURSE SYSTEM

One potential advantage of using survey dialog systems is that artifacts of the interaction, such as the prosody of respondent speech, can be used to learn information beyond that covered explicitly by survey questions. In other words, it is possible to *learn more than a survey asks* [15].

We recently conducted an experiment using the Rate-A-Course system. The experiment was designed to investigate issues of adaptation in conversation; specifically, how a dialog system user adapts to different degrees of system adaptation in initiative and lexical choice. It involved sixteen participants in each of the three dialog initiative conditions described earlier. Participants used the Rate-A-Course system to complete a survey about two of their courses. They were asked about five topics for each course. For each topic, they were first asked to rate that aspect of the course (e.g. "Was the instructor very good, good, okay, bad or terrible?"). Then, they were asked to explain their rating (e.g. "Why did you think the instructor was okay?"). Each participant was also asked to give their overall rating of each course. The resulting corpus contains ninety-six course evaluations. Figure 1 gives information about the course topics. Figure 2 shows how the system moves from topic to topic in each initiative condition.

Here, we use the corpus of course evaluations to look at how dialog epiphenomena can be used to build user models of survey respondents. We look at two aspects of the interaction: the order in which respondents chose to discuss course topics; and the duration (in seconds) of their responses to the open-ended question about each course topic. To build user models, we adapt the SMARTER procedure described in [16]. We use course topics as domain attributes and course topic ratings as attribute values. We assume independence of course topics. We map topic ratings to numeric values in [0,100] for each topic as shown in Figure 1.

Models We built six multiattribute decision models [17] for each course for each participant. To evaluate the fit of our models, we computed the Pearson correlation between participants' actual course ratings and the course rating estimates produced by each model. Each model is a weighted sum of the ratings of the topics for the course:

$$course\_rating = \sum_{i=1}^{k} w_i topic\_rating_i$$

The first model (**Equal**) weights each topic rating equally.

The second and third models use the order in which topics are discussed in the dialog to rank order course topics. The second model uses rank-order centroid (ROC) weights [18]:

ROC: $w_i = 1/k \sum_{j=i}^{k} 1/j$

The third model uses rank-sum (RS) weights [17]:

RS: $w_i = 2(k+1-i)/k(k+1)$

The fourth model weights each topic by the duration of the participant's answer to the open-ended question for that topic, divided by the total amount of time the participant spent answering open-ended questions for that course:

Duration: $w_i = duration_i / \sum_{j=1}^{k} duration_j$

The fifth and sixth models rank order topics by the duration of the participant's answer to the open-ended question for that topic, and then use rank-order centroid weights (**ROCD**) and rank-sum weights (**RSD**).

Results For the system initiative condition, there was no significant relationship between actual participant course ratings and any automatic estimates of participant course ratings. By contrast, for the mixed initiative condition, there was a moderate significant relationship between actual course ratings and: ROC (r = .51, t(30)=3.24, $p < .05$); RS (r = .50, t(30)=3.15, $p < .05$); Duration (r=.50, t(30)=3.18, $p < .05$); ROCD (r=.59, t(30)=3.95, $p < .05$); and RSD (r=.55, t(30)=3.63, $p < .05$)[1]. For the user initiative condition, there was a marked significant relationship between actual course ratings and: Equal (r=.65, t(30)=4.63, $p < .001$); ROC (r=.62, t(30)=4.37, $p < .001$); RS (r=.63, t(30)=4.49, $p < .001$); Duration (r=.67, t(30)=4.96, $p < .001$); ROCD (r=.69, t(30)=5.15, $p < .001$); and RSD (r=.67, t(30)=4.98, $p < .001$).

We conclude that when a survey dialog gives the respondent at least some initiative, (1) we can learn the factors that contributed most to a respondent's overall course rating, by ranking them either in the order they were discussed or by response duration (for open-ended responses); (2) we can use the ranked factors to estimate the respondent's overall course rating. However, while user initiative leads to shorter dialogs for respondents without cost to the user model, other criteria (such as a desire to obtain values for all aspects of a course) may mitigate against user initiative in survey dialog systems.

There are several possible reasons why our models are not more highly correlated with participants' overall course ratings. For some participants there were factors (e.g. the textbook) that were not among the course topics in the survey. Also, some participants commented that their course did not have a TA so they could not discuss that topic.

The topic order data from this experiment contains other trends. Participants in the user initiative condition usually only discussed one or two aspects of a course (mean: 1.8 topics), most often the instructor and/or exams. Furthermore, what matters to participants most about a course is somewhat course-specific: only 9 of 32 participants in the mixed or user initiative conditions chose the same topic first for both

[1] A Bonferroni adjustment has been applied to all per-comparison p values.

courses they evaluated, although 23 had the same topic in the first two for both courses. Thus, the models described above are both user- *and* course-specific models, rather than simply user-specific models.

One potential application of this work would be in designing course evaluation summarization and question-answering functionalities for the Rate-A-Course system. For example, if a particular student demonstrates a strong preference for a particular aspect of their courses over time, then summaries targeted to that student could focus on that aspect [19]. Or if multiple students chose one aspect of a particular course first, that aspect could be emphasized in summaries for that course.

Finally, other dialog epiphenomena could be useful for this or related tasks. For example, we could look at the emotional strength of a respondent's answer to an open-ended question, measured using prosody or using word-based ratings of emotional strength.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the Stony Brook Rate-A-Course survey dialog system, a survey dialog system that can supplement or replace existing written and electronic course evaluation surveys. We showed how the dialog epiphenomena, specifically the order in which questions are answered and the durations of responses to open-ended questions, can be used to learn more than a survey asks.

We are currently analyzing the experimental data used in this work to test the hypotheses for which the data was initially collected, and to build richer user models based on more features. In future work, we plan to extend the Rate-A-Course system so that callers can hear summaries of evaluations from the system and from web-based course surveys; to extend the system's range of dialog behaviors for further dialog adaptation studies; and to explore methods of automatically transcribing, indexing and summarizing respondents' answers to open-ended questions.

5. REFERENCES

[1] A. Gorin et al., "How may I help you?," *Speech Communication*, vol. 23, pp. 113–127, 1997.

[2] I. Urbina, "Your train will be late, she says cheerily," *The New York Times*, November 24 2004.

[3] M. A. Walker et al., "Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email," in *Proceedings of COLING/ACL*, 1998, pp. 1345–1352.

[4] S. Singh et al., "Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system," *Journal of Artificial Intelligence Research*, vol. 16, pp. 104–133, 2002.

[5] M. Walker et al., "DARPA Communicator: Cross-system results for the 2001 evaluation," in *Proceedings of ICSLP*, 2002.

[6] D. Litman and S. Silliman, "ITSPOKE: An intelligent tutoring spoken dialogue system," in *Companion Proceedings of HLT/NAACL*, 2004.

[7] G. Aist et al., "Talking through procedures: An intelligent space station procedure assistant," in *Proceedings of EACL*, 2003, pp. 187–190.

[8] J. Allen et al., "Towards conversational human-computer interaction," *AI Magazine*, vol. 22, no. 4, pp. 27–37, 2001.

[9] Council of American Survey Research Organizations, "Media facts – commercial regulations and survey research," 2006, http://www.casro.org/media/.

[10] S. Fricker et al., "An experimental comparison of web and telephone surveys," *Public Opinion Quarterly*, vol. 69, no. 3, pp. 370–392, 2005.

[11] M. Schober et al., "Misunderstanding standardized language," *Applied Cognitive Psychology*, vol. 18, pp. 169–188, 2004.

[12] K. Spencer and L. Schmelkin, "Student perspectives on teaching and its evaluation," *Assessment & Evaluation in Higher Education*, vol. 27, no. 5, pp. 397–409, 2002.

[13] W. Wilhelm and C. Comegys, "Course selection decisions by students on campuses with and without published teaching evaluations," *Practical Assessment, Research & Evaluation*, vol. 9, no. 16, 2004.

[14] BeVocal, "BeVocal Cafe," http://cafe.bevocal.com.

[15] M. Bosnjak and T. Tuten, "Classifying response behaviors in web-based surveys," *Journal of Computer-Mediated Communication*, vol. 6, no. 3, 2001.

[16] W. Edwards and F. Barron, "SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement," *Organizational Behavior and Human Decision Processes*, vol. 60, pp. 306–325, 1994.

[17] W. Stillwell et al., "A comparison of weight approximation techniques in multiattribute utility decision making," *Organizational Behavior and Human Performance*, vol. 28, pp. 62–77, 1981.

[18] F. Barron and B. Barrett, "Decision quality using ranked attribute weights," *Management Science*, vol. 42, pp. 1515–1525, 1996.

[19] A. Stent et al., "Trainable sentence planning for complex information presentations in spoken dialog systems," in *Proceedings of ACL*, 2004, pp. 79–86.