AI Magazine Volume 28 Number 2 (2007) (© AAAI)

Dialogue on Dialogues—Multidisciplinary Evaluation of Advanced Speech-Based Interactive Systems:

A Report on the Interspeech 2006 Satellite Event

Kristiina Jokinen, Michael McTear, and James A. Larson

■ The Dialogue on Dialogues workshop was organized as a satellite event at the Interspeech 2006 conference in Pittsburgh, Pennsylvania, and it was held on September 17, 2006, immediately before the main conference. It was planned and coordinated by Michael McTear (University of Ulster, UK), Kristiina Jokinen (University of Helsinki, Finland), and James A. Larson (Portland State University, USA). The one-day workshop involved more than 40 participants from Europe, the United States, Australia, and Japan.

In recent years there has been a growth of research focusing on speech-based interactive systems that aim to increase their communicative competence by including aspects of interaction beyond simple speech recognition and question and answer–based interaction. One of the motivations for furthering the systems' interaction capabilities is to improve the systems' naturalness and usability in practical applications. However, relatively little work has so far been devoted to defining the criteria according to which we could evaluate such systems in terms of increased naturalness and usability. It is often felt that statistical speech-based research is not fully appreciated in the dialogue community, while dialogue modeling in the speech community seems too simple in terms of the advanced architectures and functionalities under investigation in the dialogue community. From the industrial point of view, academically developed research is often too far removed from what is needed in practical applications, while academic researchers often feel that their research results do not find their way into industrial applications.

The goal of the workshop was to bring together researchers and practitioners working on the development of dialogue systems in order to clarify and discuss issues dealing with the evaluation of advanced speech-based interactive systems. The focus was on the development of practical dialogue systems that support robust, natural, and efficient spoken language interaction and on the advances in basic research areas such as speech segmentation, disfluencies, turn-taking, emotions, and adaptation.

The workshop was divided into five thematic sessions, reflecting the topics addressed in the original workshop call. Each session was opened by short, five-minute presentations on the selected topic so as to provide a starting point for discussion and was followed by ample time to work on the topics more thoroughly. We report on the highlights of the workshop discussions and the future steps discussed in the final session of the workshop.

Evaluation Criteria

The first session was devoted to discussing questions such as What metrics should be used to measure static dialogues? and What metrics should be used to measure dialogues that learn? The presenters were Nigel G. Ward ("Evaluating Real-Time Responsiveness in Dialog"), David Griol (talking about the shared work by F. Torres, L. Hurtado, S. Grau, F. Garcia, E. Sanchis, and E. Segarra, "Development and Evaluation of the DIHANA Project Dialog Systems"), and Gregory Aist (talking about the shared work by P. Michalak, G. Ferguson, and J. Allen, "Challenges in Evaluating Spoken Dialog Systems That Reason and Learn").

Measures and expectations are different for different domains and different applications. It was also pointed out that calibrated evaluation and discount usability evaluations are unreliable in real situations: there is a difference between a general user and the actual user. Real-time interaction brings forward large individual differences, and the potential value of the system is not easily estimated. Interestingly, low-quality dialogue systems still sell, so there is a need to investigate the relation between tools that enable the user to complete the task but are not sophisticated and improvements that allow sophisticated interaction with the user but are not so effective in task completion.

One of the main problems in evaluating practical dialogue systems is the fact that task completion is not a single measure, and thus a nonmodular system design does not easily lend itself to principle-based evaluation of sophisticated dialogue modeling. For instance, in VoiceXML, the current standard for many commercial applications, the dialogue design is not modular, and so system design and refinement are not possible: user satisfaction does not necessarily improve by doing a number of steps. Thus, we also should think about standards that apply to different architectures and allow different components to be assessed and compared in the system evaluation.

Another problematic issue is that dialogue behaviors change over time, and the system should thus learn new strategies. Important questions for the evaluation of such learning systems concern the ways in which the system can change, as well as the evaluation criteria to assess what the system has learned when it has learned to handle a particular task: how much better the system works now than before, how it compares to other systems, how much more easily a related task can be learned, how humanlike the new system is. Also the learning algorithm itself must be evaluated.

Finally, we also need to consider what kind of features humanlike behavior includes: What are the appropriate measures for system performance and human performance? Can they be intermingled, and if so, how can this best take place? An alternative is to use user simulations to evaluate the dialogue system, although it is also important to differentiate between prediction and evaluation: predicting appropriate actions in a given situation is part of the system design and can be approached by simulation, but evaluation of system functionality with respect to real user situations may not be possible except by engaging the users in actual usage of the system. The evaluation should also bring forward what the users expect from the system and how the use of the system affects their evaluation. For this, the users can fill in the same evaluation questionnaire twice, once before and once after the actual tasks; thus it is possible to compare what the users expect from the system and what their experience was, giving important information about how the different system properties affected the user.

Semiautomatic Design of Dialogues

This next session focused on two different semiautomatic design methods for dialogue design: example-based learning and reinforcement learning. The initial presentations were by Gary Geunbae Lee (discussing shared work by Sangkeun Jung, “Dialog Studio: An Example-Based Spoken Dialog System Development Workbench”) and Tim Paek (“Reinforcement Learning for Spoken Dialogue Systems: Comparing Strengths and Weaknesses for Practical Deployment”).

The design of dialogue systems still seems to be more art than engineering, and there is no clear methodology of how to build spoken dialogue systems. The systems are more or less structured software programs, and design principles are heuristics obtained by trial-and-error experimentation. Extensive iterative design is necessary as it is difficult to predict all usage situations, but there is no principled way to guide how to develop a dialogue manager for problems in new dialogues.

One approach to dialogue design and evaluation is to automate system evaluation by checking which strategies work on the basis of corpus data. For instance, new system responses can be learned from the old system responses using example-based learning methods. A problem with this approach is how to get new instances, that is, how to dynamically extend the example database. One solution might be to integrate a speech-recognition and dialogue model into a system so that they inform each other back and forth.

Dialogue managers and user parameters can also be optimized using machine-learning techniques such as reinforcement learning, which allows an optimal path to be found in the state space. The dialogue strategy specifies for each state what the next action to be invoked is, and the number of strategies increase exponentially as the number of states and actions increases. The learning problem is to automatically find the optimal strategy that minimizes the objective function. In the Markov decision process (MDP) that is used to describe dialogue systems, the quantity to be optimized is formalized as a weighted sum of dialogue costs (such as duration, errors, distance to task completion). Dialogue design thus boils down to finding the optimal strategy in an MDP, that is, learning an optimal policy or mapping between actions and states. The optimal value of a state is the expected sum of costs incurred from the state and following the optimal strategy until the final state is reached.

However, it seems as if the reinforcement style is good for certain types of domains only where the task is well-formed, such as providing information about day and time, hotel booking, or tourist information. In question-answering or negotiation dialogues where the content is important, the shortest interaction is not always the best, so the objective function that we try to minimize may not be appropriate. The question also arises whether we can get reliable data about the goodness of the system by automatic design: the method focuses on the evaluation of the development of the dialogue system rather than on the usability or the user experience.

The benefit of automatic design from the corpus can also be measured in terms of work load and resources: the method requires a large amount of annotated data, the production of which is costly and time-consuming. In fact, it seems as if manually crafted rules are easier to produce, and they work equally well. Moreover, they have the advantage that the rules can be explained to the user. The dialogue manager is not a black box, but the user should have control over what features and aspects to add in the dialogue management.

On the other hand, machine-learning models have often been applied statically so that the policy, once learned, is used as a fixed policy, and further learning or adaptation is not possible. The problems with new users and different dialogue situations can be tackled by online learning and also by allowing the users to change parameters later. In this way it is possible to model dynamic systems that adapt to novel creative behavior. However, online learning can suffer from the lack of reliable teaching: it is difficult to determine what is noise and what is proper use of infrequent strategies.

Automatic dialogue design also prompts the question about the best practices for defining the dialogue management states: what the system should do and what kind of states it should have. Applications of reinforcement learning technique take it for granted that a set of dialogue states and actions is given, but they do not consider how well the sets describe the actual dialogue situations. Most applications concern only system confirmation and repairs as components that can be reused in dialogue management best practices. However, best practices for industry and research are different, and although new better practices are brought forward by research, it is hard to change established industrial best practices afterwards. In order to bridge the gap, it would be useful to find the best practices used in industry and help researchers in industry to develop better objective functions to evaluate the dialogue systems. Thus research can influence industrial practices by pointing to those different aspects that should be added in the industrial dialogue design.

Methodologies for Improving Dialogue Design

The next session dealt with evaluating alternative methodologies for improving dialog design. The presentations were given by Rebecca Passonneau (talking about shared work with Ester Levin, "A WOZ Variant with Contrastive Conditions") and Zoraida Callejas (discussing shared work with Ramón López-Cózar, "Human-centered Development of Interactive Systems: Improving Usability in Early Lifecycles Stages").

The WOZ-paradigm has been used to collect data, and the question is how to update the WOZ technique to better resemble human-machine dialogues. This is also related to enhancing the MDP approach to learn optimal dialogue strategies: we collect and improve human-machine dialogues. This is done through wizard ablation: by removing functionality and studying the difference, for example, in how speech understanding errors by the system can be handled more naturally. There are several points to consider in this approach, however. First, the training of the wizard takes time and does not guarantee consistent behaviour: the instructions may not be understood in a similar way by two different wizards. Determining what the wizard is meant to understand from the user contributions also presupposes that the system is already designed, for example, that the repertoire of dialogue acts and the strategy to choose between dialogue acts are already fixed. Thus the method does not really address the problem of dialogue design but is a preimposed system enhancement.

Modeling Dialogues, Multimodality, and Visual Input

Alternative modeling techniques were discussed in the two afternoon sessions. The speakers were Bilyana Martinovska (presenting shared work with Ashish Vaswani, "Activity-Based Dialogue Analysis as Evaluation Method"), Deryle Lonsdale (talking about shared work with Rebecca Madsen, "Unifying Language Modeling Capabilities for Flexible Interaction"), and Jens Edlund (discussing shared work with Mattias Heldner and Joakim Gustafson, "Two Faces of Spoken Dialogue Systems").

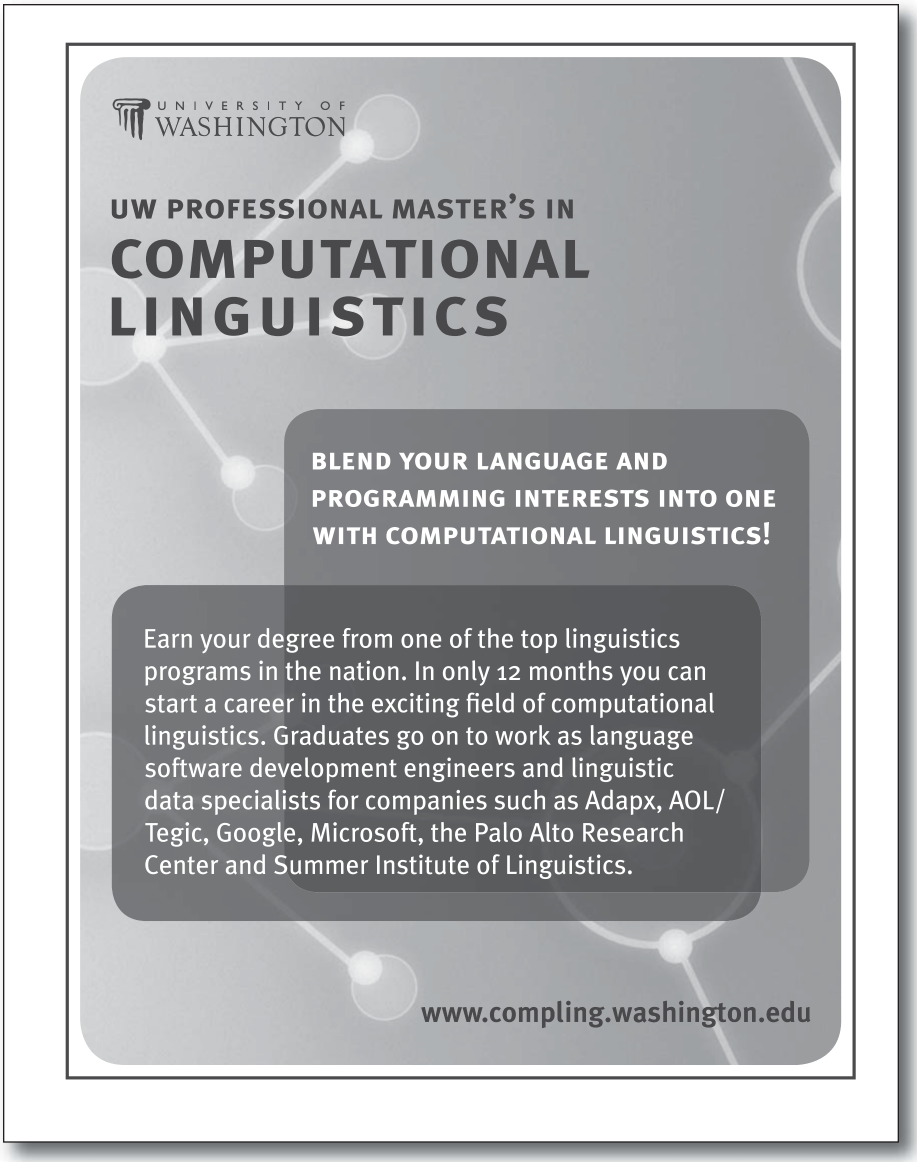
Different users as well as different activities can trigger similar behaviors, but the dialogues are still different concerning the liveliness of the activity. We can try to measure dynamics of interaction in dialogues, as exemplified by the difference between an Italian dinner and a sermon, for example, by measuring backchanneling. As for practical application, interactive recipes can be used as scripts of cognitive behavior, which can then be made more concrete in the particular application.

Another evaluation method introduced in the session was screening, which is widely used in game evaluations. However, dialogue system evaluation is usually different since the evaluators are participating in the dialogue themselves, and there is usually a huge difference between participating and observing an activity. Besides screening, it may be possible to have a test suite or to set up an evaluation contest based on shared resources like MapTask.

The last session focused on multimodal applications. The speakers were Ramón López-Cózar (discussing shared work by Zoraida Callejas and Germán Montoro, "DS-UCAT: A New Multimodal Dialogue System for an Academic Application"), and Gregory Aist ("Computer Vision, Eye Tracking, Spoken Dialog Systems, and Evaluation: Challenges and Opportunities").

Multimodal applications are usually considered advantageous as they allow getting the best benefits by combining different modalities, for example, selecting a suitable modality, vision or speech, for educative purposes. Also applications for special users can be built, and thus universal usability is possible. From the evaluation point of view, the question is how to evaluate multimodal systems, since the modalities add extra complexity to the evaluation process. For instance, using a noise-level detector to decide whether to switch to a visual mode is problematic, and often it is necessary (or at least more useful) to provide feedback by visualizing information rather than talking.

Multimodal systems are also said to add to more humanlike naturalness, but how do researchers measure the impact of different modalities or user satisfaction of such systems? Extra problems are also encountered concerning control strategies in the system: it is not easy to evaluate mixed-initiative dialogue strategies. Emotional dialogue management cues are also to be taken into account. Moreover, the user may not be paying attention to the system and the task, and the problem is how to get the user back to the system.



Dissemination of the Results

The lively discussions brought up many problematic and unresolved issues related to dialogue system evaluation but also pointed out research activities and expertise for solving fundamental problems and for developing shared understanding further. The multidisciplinary discussion revealed the complexity of the issues and made it clear that it is important to continue dialogue on these issues. All the participants shared this view.

As the start for the next steps, a summary of the discussions was posted on the workshop website, and a special issue of the journal *Speech Communication* on "Evaluating new methods and models for advanced speech-based interactive systems" is on the way. Journal submissions will be open to those who did not attend the workshop as well as those who did. In addition, a Wikipedia discussion forum, *Dialoguetalk,* is available and open to everyone who registers.

We hope that the workshop has provided a forum for further studies and has offered inspiration and insight into the evaluation of advanced speech-based systems. For more information, contact the organizers at interspeech06-dod@helsinki.fi. The project website is located at www.ling.helsinki.fi/~kjokinen/ICSLP06-DoD. Dialoguetalk is located at www.dialoguetalk.net.

Kristiina Jokinen has played a leading role in many dialogue-related technology projects at University of Helsinki. Her research concerns spoken dialogue systems, cooperation, multimodal communication, and evaluation of interactive systems—topics that she is also pursuing in her forthcoming book *Constructive Dialogue Management – Speech Interaction and Rational Agents* (John Wiley & Sons). She is an adjunct professor of interaction technology at the University of Tampere, and she was a Nokia Foundation visiting fellow at Stanford University in 2006. She is secretary of SIGDial, the ACL/ISCA special interest group for discourse and dialogue.

James A. Larson specializes in training and consulting for user interfaces and voice-enabled applications. As cochair of the W3C Voice Browser Working Group, Larson is a leader for the standarization of languages for developing speech applications, including VoiceXML 2.0, Speech Recognition Grammar, and Speech Synthesis Markup Languages. Larson is a member of the adjunct faculty at Portland State University and Oregon Health Sciences University/Oregon Graduate Institute, where he designs and teaches courses for user interfaces and speech applications. He is a contributing author for *Speech Technology* magazine and has written many articles and books on user interfaces.

Michael McTear is a professor of computer science at University of Ulster. His main research interest is spoken dialogue technology. Other interests include user modeling, natural language processing, and language acquisition. He is the author of the widely used textbook *Spoken Dialogue Technology: Toward the Conversational User Interface* (Springer Verlag, 2004) and coinvestigator of the Queen's Communicator dialogue system. He has given several invited talks and was the recipient of the Distinguished Senior Research Fellowship of the University of Ulster in 2006. He is a member of the Scientific Advisory Group of SIGDial and the ACL/ISCA Special Interest Group for Discourse and Dialogue.