Interspeech2006 - ICSLP Satellite Workshop

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

September 17, 2006, Pittsburgh, PA, Westmoreland Room, The Westin Convention Center Pittsburgh

9:00-9:30 Session 1: Welcome

The goal of the workshop was to identify and enumerate criteria for evaluating robust and efficient interaction in spoken dialogue systems, and to define the criteria according to which we could evaluate speech-based interaction systems.

9:30-10:45 Session 2: Evaluation criteria

- *Evaluating Real-time Responsiveness in Dialog* by Nigel G. Ward
- *Development and evaluation of the DIHANA project dialog systems* by D. Griol, F. Torres, L. Hurtado, S. Grau, F. Garcia, E. Sanchis, and E. Segarra
- *Challenges in evaluating spoken dialog systems that reason and learn* by Gregory Aist, Phil Michalak, George Ferguson, James Allen
- Discussion:
    1. What metrics should be used to measure static dialogs?
    2. What metrics should be used to measure dialogs that learn?

Calibrated evaluation and discount usability evaluations are unreliable. Real-time interaction brings forward large individual differences, and the potential value of the system is un-estimated. Low-quality dialogue systems still sell, and there is a difference between a general user and the actual user. Also it is not clear how to present the improvements: we need standards for visualization of evaluation measures, and also measures for corrected errors. We should also take into consideration influence of physiological info like reaction times.

Considering VoiceXML, the problem is that dialogue design is not modular, and user satisfaction does not necessarily improve by doing a number of steps: there is no formative evaluation, i.e. evaluation for forming and refining system design is not possible. Task completion is not a single measure, and we also should think about standards that apply to different architectures.

Also, the system should learn since the dialogue behaviours change over time. However, in what ways can the system change, and what are the evaluation criteria after the system has learnt to handle X: can the system do X, how human-like is X, how much better the system works, how it compares to other systems, how much more easily a related task can be learnt. We also need to evaluate the learning algorithm itself.

What about different measures for system performance and human performance, and for learning, system learning and user learning? One alternative is to use user

simulations to evaluate the dialogue system. However, the goal for dialogue systems is not to behave like human. What do humans want when talking to other humans? And do we know that?

We also need to make a difference between prediction and evaluation. Measures and expectations are different for different domains and applications. The evaluation could also bring forward what the users expect from the system, and how the use of the system affects their evaluation. For instance, the users can fill in the same evaluation questionnaire twice, one before and one after the actual tasks (pretest and posttest), and thus we get to know what the users expect from the system and what their experience was, gaining interesting and important information about how the different system properties affected the on the user.

10:45-11:00 Break

11:00-12:00 Session 3: Semi-automatic design of dialogs

- *Dialog Studio: An Example-Based Spoken Dialog System Development Workbench* by Sangkeun Jung, Cheongjae Lee, Gary Geunbae Lee
- *Reinforcement Learning for Spoken Dialogue Systems: Comparing Strengths and Weaknesses for Practical Deployment* by Tim Paek
- Discussion
    1. What techniques can apply to semi-automatic design of dialogs?
    2. How to evaluate these techniques?

This section focussed on two different automatic design methods for dialogue management: example-based learning and reinforcement learning.

The design of dialogue systems still seems to be more art than engineering since we lack methodology of how to build spoken dialogue systems. The systems are more or less structured software programs based on application knowledge and experimentation with users, while design principles are heuristics obtained by trial and error-type experimentation. It is hard to predict all usage situations and thus extensive iterative design is necessary, but with a new application, new design work is needed again: there is no principled way to guide how to develop a dialogue manager for problems in new dialogues. We need new ASR and NLU, but also new dialogue patterns.

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 learnt from the old system responses using example-based learning methods. A problem with this approach is how to get new instances, i.e. dynamically extend the example database. One solution might be to integrate ASR and DM into a system so that they inform each other back and forward.

Dialogue managers and user parameters can also be optimized using machine learning techniques such as reinforcement learning which allows optimal path to be found in the state space. The dialogue strategy specifies for each state what is the next action to be invoked, and the number of strategies increases exponentially as the number of states and actions increases. The learning problem is to automatically find the optimal

strategy which minimizes the objective function. In Markov Decision Process (MDP) that is used to describe dialogue systems, the quantity to be optimized is formalised as a weighted sum of dialogue costs (such as duration, errors, distance to task completion). Dialogue design thus boils down to finding optimal strategy in a MDP, i.e. learning an optimal policy or the mapping between actions and states. The optimal value of a state is the expected sum of costs incurred from state and following optimal strategy until the final state is reached.

However, it seems like reinforcement style method is good only for “simple” domains where the task is well-formed, like 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 if 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. There is a difference in subjective and objective evaluations, and usability evaluations show that people prefer different dialogue styles and also that different strategies are appropriate in different situations.

We can also ask the benefit of automatic design from the corpus 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 like 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 can 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 once learnt policy 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 on-line learning, and also by allowing the users to change parameters later. In this way it is possible to model dynamic systems which adapt to novel creative behaviour. 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 prompted 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 practises for industry and research are different, and although new better practises are brought forward by research, it is hard to change established industrial best practises afterwards. In order to bridge the gap, it would be useful to find the best practices used in industry and help them to develop better objective functions to evaluate the dialogue systems. Thus research can influence the industrial practises by pointing to those different aspects that should be added in the industrial dialogue design.

But what is a good objective function? What are quality answers? It has been shown that task completion and user satisfaction do not necessarily make the best objective function. If the system is considered as a tool, they are fine, but the objective function should be flexible so as to allow different types of dialogue systems to be evaluated and compared. For instance, what makes human-human dialogues to work well is their flexibility, i.e. the tool is not the system but the dialogue that the system enables us to conduct with the back-end application.

Evaluation can also be understood as being similar to evaluating teaching in a class: how effective and satisfactory the teaching has been. Class room evaluation can be affected by other things than actual teaching as well: e.g. the teacher might have given you bad marks earlier or turn-taking failed (the teacher did not answer your question), etc. The features that predict user satisfaction are not in a linear relation, but we need to look at them in a holistic way, combining prediction and methodology. Users don't usually understand what "explicit evaluation" means, and we need to be careful when comparing and projecting laboratory evaluations towards real user evaluations: the criteria are different when evaluating a military application and an educational application.

12:00-13:00 Session 4: Methodologies for improving dialog design

- *A WOz Variant with Contrastive Conditions* by Ester Levin and Rebecca Passonneau
- *Human-centered Development of Interactive Systems: Improving Usability in Early Lifecycles Stages* by Zoraida Callejas, Ramón López-Cózar
- Discussion
    1. How to evaluate alternative methododologies for improving dialog design?
    2. What metrics to use?

WOZ-paradigm has been used to collect data, and the question is how to update WOZ technique to better resemble human-machine dialogues. This is related to enhancing MDP approach to learn optimal dialogue strategies: now we do not pretend collecting human-human dialogues, but simply collect and improve human-machine dialogues. This is done via wizard ablation: by removing functionality and studying the difference, e.g. 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 need not be understood in a similar way by two different wizards. In order to determine what the wizard is meant to understand from the user contributions also presupposes that the system is already designed, e.g. the repertoire of dialogue acts and the strategy to choose between dialogue acts is already fixed. Thus the method does not really address the problem of dialogue design but is pre-imposed system enhancement.

Concerning different users, elderly users' requirements for multimodal dialogue system are difficult to meet. The research shows that people don't like animated agents at home as they intrude their privacy, and elderly people also speak differently. However, one way to go on is to cluster users and pre-select dialogue strategies

according to the users' abilities, focussing on adaptability modelling in the dialogue system.

13:00-14:30 Lunch

14:30-15:30 Session 5: Modeling dialogs

- *Activity-based dialogue analysis as evaluation method* by Bilyana Martinovska, Ashish Vaswani
- *Unifying language modeling capabilities for flexible interaction* by Deryle Lonsdale and Rebecca Madsen
- *Two faces of spoken dialogue systems* by Jens Edlund, Mattias Heldner & Joakim Gustafson
- Discussion
    1. How to evaluate alternative modeling techniques
    2. What metrics to use?

Different users as well as different activities can trigger similar behaviours, 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, e.g. by measuring back-channelling. However, interactive vs transactive metaphors plus mixed metaphors for backchannel communication are needed: measuring task completion rate or efficiency are not sufficient to determine interactivity, since *hh* and *hm* could be equally efficient, but differ in style, in the competence demanded, and in the learning strategies desired. As for practical application, interactive recipes can be used as scripts of cognitive behaviour, which can then be made more concrete in the particular application. For instance, in a recent MIT work, a dating system was developed by monitoring how the people interacted and gave feedback to each other, and extracting those features that could be used to define if the person is interested in someone or not.

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 set-up a evaluation contest based on shared resources like MapTask.

15:30-16:30 Session 6: Multimodal dialogs and visual input

- *DS-UCAT: A new Multimodal Dialogue System for an Academic Application* by Ramón López-Cózar, Zoraida Callejas, Germán Montoro
- *Computer vision, eye tracking, spoken dialog systems, and evaluation: Challenges and opportunities* by Gregory Aist
- Discussion
    1. What are the advantages and disadvantages of the X+V approach to implementing multimodal dialogs?
    2. What are alternative architectures to X+V?
    3. How does visual input affect dialogs?
    4. Can X+V be extended to support visual input?

Multimodal application is usually advantageous as it allows getting best benefits by combining different modalities, e.g. selecting suitable modality, vision vs. 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, detection of the level of noise to choose access to visual aid is also problematic, and visualization of info rather than talking as reaction is needed. Is visual learning possible? As the user is engaged in the conversation, it is necessary to think what kind of information is delivered to different people in the same space. Also turn-taking needs to be considered, as well as tokenizing: what constitutes a single visual event, and what is a sequence of events. Finally, it is not yet clear how vision and speech align.

Multimodal systems are also said to add to more human-like naturalness, but how to measure the impact of different modalities, or user satisfaction of such systems? Extra problems are also encountered concerning control strategies in the system: 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 force the user get back to the system.

16:30-16:45 Break

16:45 - 18:00 Session 7: Next steps

- How to disseminate the results of this workshop to the speech dialog community?
    - Edit minutes and post on workshop web site
    - Paper targeted to speech dialogue practitioners for publication in Speech Technology Magazine, AAAI Review or similar magazine
    - Paper targeted to the academics community for publication in a referred journal (Speech Communication)
    - CD for SpeechTech conference
    - Wikipedia discussion forum “Talkdialogue”