The Spoken Dialogue Challenge

Alan W Black and Maxine Eskenazi
Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA, USA
{awb,max}@cs.cmu.edu [name]

Abstract

In the field of speech and language processing the introduction of "Challenges" has helped focus the focus a field, allowing detailed comparisons of systems and techniques, bringing new members into the field, and facilitating advancements in core research. Although the idea of a spoken dialogue challenge has been discussed for some time, this is the first attempt to bring these discussions together and take concrete action.

1 Motivation for a centralized Challenge

The idea of a centralized challenge in the fields of speech and language where different systems and techniques are applied to the same data has been with us for some time. Probably the most lengthy and successful challenge is the DARPA-funded speech recognition set of challenges [3] which started in the 1980s and came to a peak in the 1990s. They still continue through different programs, and in spite of being criticized for focusing Automatic Speech Processing (ASR) on simple measures of success, it is clear that they have helped make ASR fundamentally better.

The important points in a successful challenge are: a well-defined task that the community considers to be challenging for the current state of the art; sufficient number of participants with varied systems so that the performance of different techniques can be evaluated; generally accepted evaluation/success criteria; and a sense of collaboration between participants so that details of their entries may be reasonably shared. It is important that the challenge does not become an outright competition, but striving to produce the "best" system encourages exploring new ideas.

2 Introduction

Within the field of spoken dialogue systems perhaps the most similar coordinated evaluation of multiple systems was part of the DARPA-funded Communicator program. This program gave spoken dialogue system access to a flight information and booking system. In 2000-2001 a number of systems developed as part of the program were centrally evaluated by NIST. A number of external callers accessed each system and the results were published in [9]. Since spoken dialogue systems do not have as clear a measure of success as ASRs, a number of different measures were designed, both objective and subjective. And other analyses of the evaluation were made by a number of groups including [1].

Other competitions have been developed including the Loebner Prize [5], which addresses the issues of (at least originally) text-based conversations that attempt to pass the Turing Test. This year, however, the Loebner Prize is being held in conjunction with Interspeech 2009 and will have a speech component.

3 Proposed Spoken Dialogue Challenge

Although it is clear there are many possible choices for this challenge we feel the first round of the challenge should be simple and clear, allowing for the types of tasks to grow in variety over the years. Our initial proposal builds on our own experience in building deployed spoken dialogue systems. Our system has already collected a large number of example dialogues that can be used for training and made available in the first year to participants.

The target domain is bus information. This was chosen because it is a popular domain, used by several different dialogue systems. It is useful to the public and doesn't require dealing with sensitive information. The system is simple enough that it can be re-implemented without

extensive additional work. We have already collected over 75,000 deidentified dialogues in the domain that we can distribute for training purposes.

The second important detail to define is what we expect of the participants (the task). There is a wide variety of research interests in the current spoken dialogue field. It is impossible for a single task to provide suitable challenges for everyone in its first iteration, so it is important to find a task in which a substantial subset can compete.

Given the bus information task as the most wide-reaching for the first year, we see three levels of participation:

1. Build a bus information system with your spoken dialogue architecture
2. Build a bus information system with free software tools such as Olympus II.
3. Take the existing Let's Go Bus Information system and adapt it with your components.

These three levels offer a certain amount of freedom to the participants. If participants are only interested in one small part of the dialogue task they can added new components to the Let's Go Bus Information system [7], such as a new recognizer, a new synthesizer, or address core dialog components like error recovery etc. The initial language would be (US) English but we would like to extend this in later years.

3.1 Evaluation

Evaluation of spoken dialogue systems is still very much a research issue and we see the Spoken Dialogue Challenge as a mechanism to aid that area of research. Following structures in the Speech Synthesis Blizzard Challenge [2], we propose to have a number of different groups of users call the submitted systems.

Group 1 - Dialog Researchers: each participant group will provided a number of callers who will be given scenarios and asked to call some of the participating systems. Dialogue Researchers are probably the best users of a system and are also the most personally interested in completing the task.

Group 2 – Native Speaker Undergraduates: this group will be paid and intended to be the most homogeneous group of callers.

Group 3 - Volunteers: by requesting for volunteers through mailing lists and the web we will collect the third set of callers.

Having three sets of callers will enable us to perform correlation between the groups and therefore try to find reliable statistics.

We also propose to run the Challenge on two levels. All participating systems will take part in these initial evaluations. The best (or most stable) systems can then be deployed on the Let's Go Live system which provides bus information to the people of Pittsburgh. Thus real users, who are interested in the time of the next bus rather than the success of the dialogue system, will provide an additional set of caller statistics.

The first groups of users will be given scenarios (they are unlikely to be familiar with Pittsburgh busses), and will fill in a web questionnaire after each call. For the live test with real users no questionnaire will be possible. Real users are uninterested in answering any further questions, and a design that gets some subset to fill in questionnaires would probably yield a different result than standard real users.

Evaluation will be through objective and subjective measures. Although evaluation of dialogue systems is still a research issue, we will offer the conventional techniques including task completion and number of turns as well as questionnaires about user satisfaction. As we do have some support for labeling we would also consider calculating word error and other dialogue state level labeling to find accuracy of systems.

3.2 Where to base the first iteration of the Challenge

Why should this challenge be based at Carnegie Mellon? First CMU has been prominent in spoken dialogue research for many years and importantly has provided free software systems to jumpstart others' system building efforts. Olympus II [6] contains everything needed from recognition, synthesis, parses, generators and a dialogue system that allows others to build complete systems. We have provided tutorials describing how to use each of the components and the whole systems.

We also have the experience of running Let's Go Lab [8], which offers a real-time platform with real users that spoken dialogue studies from other institutions have used. We have already gained experience in dialogue evaluation and in providing support for other dialogue teams. This makes the first challenge a clear extension of our existing spoken dialogue evaluation services. After this first iteration, other institutions and applications will be chosen under similar criteria.

3.3 Challenges for the Challenge

Challenges must be supported by the community. The committee governing this challenge therefore consists of leading experts representing a variety of areas in spoken dialogue research: Dr Jason Williams, (AT&T), Dr Dan Bohus, (Microsoft), Prof David Traum (USC), Prof Kristiina Jokinen (Helsinki) and Prof Helen Meng (CU Hong Kong).

Evaluation of spoken dialogue systems is still very much a research issue. Thus the Spoken Dialogue Challenge will become a mechanism that serves that area of research too. The initial metrics that will be produced will include objective measures such as task completion, number of turns etc, and subjective measures (from questionaires) such as user satisfaction. The dialogue data collected during the Challenges will be made available to researchers to further investigate evaluation metrics.

There are a number of non-trivial issues in bringing a successful challenge like this together. Although we do not yet have all of the solutions we are aware of many of the issues.

To be successful we need to have participants. Since there is no funding for participants it is hard for them to devote resources to a task for which they are not being funded. Building a dialogue system is a considerable amount of effort and it may be that only a few groups have the additional resources to participate. We will make every effort to make participation as easy as possible but we know from experience that getting, especially initial, participants is hard. As the Challenge matures over time, we expect to see more participants as people see it as a worthwhile endeavor, and modify their research agenda to include it.

A second issue is computing infrastructure, for an anticipated international set of participants we have to address how callers access systems. International telephone charges notwithstanding, international lines are often more prone to noise than national lines (and often have a delay). Alternatives such as voice over IP (including Skype) have been suggested but they bring in their own issues, ASR of VoIP encoded calls is not the same as ASR over cell phones or landlines. We may have versions of the systems made available at centers on each continent where there are participants.

3.4 Future extensions

We see The Spoken Dialogue Challenge as a long term progressive event. There are a number of current research areas that will be hard to include in the initial instantiation but are certainly worth including in future years. A couple of specific areas are worth considering. It would be good if the systems could be automatically tested. This implies some sort of user simulator, such research has been considered by a number of research groups and it is certainly an interesting direction to take. The second area that is currently active in dialogue research is reinforcement learning or any technique that requires significant numbers of interactions and dynamically learns from them. Again we would like to see this added to future Challenges.

4 Timeline

After consultation with participants at SigDial 2009, we will put out a call for participation in mid-October 2009 in a Challenge that will start in early 2010. We will allow at least 6 months for groups to develop their systems, and allow 3 months for evaluation. The intention is to then present the results and descriptions of the participating systems at a dedicated workshop or special session of a conference.

References

[1] Bennett, C., "A Comparative Analysis of DARPA Communicator Systems," Presentation at DARPA Communicator PI Meeting, New Orleans, Louisiana, July 2001

[2] Black, A., and Tokuda, K., (2005) Blizzard Challenge -- 2005: Evaluating corpus-based speech synthesis on common datasets Interspeech 2005, Lisbon, Portugal.

[3] DARPA, "The DARPA speech recognition evaluation workshops," http://www.nist.gov/speech/publications/index.htm.

[4] Eskenazi, M., Black, A., Raux, A. and Langner, B. "Let's Go Lab: a platform for evaluation of spoken dialog systems with real world users", Interspeech 2008, Brisbane, Australia.

[5] Loebner Prize

http://www.loebner.net/Prizef/loebner-prize.html

[6] Olympus II Dialog System Framework http://wiki.speech.cs.cmu.edu/olympus/index.php/Olympus

[7] Raux, A., Bohus, D., Langner, B., Black, A., and Eskenazi, M. (2006) Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience, Interspeech 2006 - ICSLP, Pittsburgh, PA.

[8] Raux, A., Langner, B., Black, A. and Eskenazi, M. "Building Practical Spoken Dialog Systems" ACL/HLT 2008 Tutorial, Columbus, Ohio.

[9] Walker, M., Passonneau, R., and Boland, J. "Quantitative and qualitative evaluation of Darpa Communicator spoken dialogue systems" 39th ACL, pp 515-222, Toulouse, France, 2001.