Incremental Spoken Dialogue Systems: Tools and Data

Helen Hastie, Oliver Lemon, Nina Dethlefs
The Interaction Lab, School of Mathematics and Computer Science
Heriot-Watt University, Edinburgh, UK EH14 4AS
h.hastie, o.lemon, n.s.dethlefs@hw.ac.uk

Abstract

Strict-turn taking models of dialogue do not accurately model human incremental processing, where users can process partial input and plan partial utterances in parallel. We discuss the current state of the art in incremental systems and propose tools and data required for further advances in the field of Incremental Spoken Dialogue Systems.

1 Incremental Spoken Dialogue Systems

For Spoken Dialogue Systems (SDS) to be more frequently adopted, advances in the state-of-the-art are necessary to enable highly responsive and conversational systems. Traditionally, the unit of speech has been a whole utterance with strict, rigid turn-taking determined by a voice-activity detector. However, a large body of psycholinguistic literature indicates that human-human interaction is in fact incremental (Tanenhaus and Brown-Schmidt, 2008; Levelt, 1989). Using a whole utterance as the unit of choice makes dialogues longer, unnatural and stilted and ultimately interferes with a user's ability to focus on their goal (Allen et al., 2001).

A new generation of Incremental SDS (ISDS) are being developed that deal with 'micro-turns' (sub-utterance processing units) resulting in dialogues that are more fluid and responsive. Recent work has shown that processing smaller 'chunks' of input and output can improve the user experience (Aist et al., 2007; Skantze and Schlangen, 2009; Buss et al., 2010; Baumann et al., 2011; Selfridge et al., 2011). Incrementality enables the system designer to model several dialogue phenomena that play a vital role in human discourse (Levelt, 1989) but have so far been absent from systems. These include more natural turn-taking through rapid system responses, grounding through the generation of backchannels and feedback, and barge-ins (from both user and system). In addition, corrections and self-corrections through constant monitoring of user and system utterances play an important role, enabling the system to recover smoothly from a recognition error or a change in user's preferences. Some examples of the phenomena we are targeting are given in Figure 1.

Parlance, a FP7 EC project[1], is currently developing incremental systems for English and Mandarin. The goal of Parlance is to develop mobile, interactive, 'hyper-local' search through speech. Recent trends in Information Retrieval are towards incremental, interactive search. Spoken dialogue systems can provide a truly natural medium for this type of search, in particular for people on the move.

2 Tools and Data

The emphasis of the Parlance project is on data-driven techniques for ISDS, thereby addressing the problem of a lack of data for system development. Although incremental dialogue phenomena described in Figure 1 have been observed in human-human dialogue, more task-based data is needed. It is challenging to fabricate a situation where users produce incremental discourse phenomena as in Figure 1 frequently and in a natural manner. Wizard-

[1] http://www.parlance-project.eu

NAACL-HLT 2012 Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data, pages 15–16,
Montréal, Canada, June 7, 2012.

Backchannels (when the user pauses)
USR I want Italian food [500 ms] in the centre of town ...
SYS uh-huh
SYS OK. I found 24 Italian restaurants in the city centre. The restaurant *Roma* is in the medium price range,...

Self-correction (the system made a mistake)
USR I want Italian food in the centre of town ...
SYS OK. I found 35 Indian restaurants ...
USR No, I want Italian.
SYS oh sorry ...
SYS I have 24 Italian restaurants in the city centre ...

Holding the floor
USR I want cheap Italian food ...
SYS ok let me see
SYS I have 3 cheap Italian places ...

Figure 1: Incremental phenomena observed in human-human dialogue that systems should be able to model.

of-Oz experiments can be used to collect data from the system side, but user-initiated phenomena, such as the user changing his/her mind are more difficult to instigate. Therefore, data collections of naturally occurring incremental phenomena in human-human settings will be essential for further development of incremental systems. Such data can inform user simulations which provide means of training stochastic SDS with less initial data and can compensate for data sparsity. For example, in Dethlefs et al. (2012) the user simulation can change its mind and react to different NLG strategies such as giving information with partial input or waiting for complete input from the user. Both the academic community and industry would benefit from open access data, such as will be collected in the Parlance project and made available to the dialogue community[2]. There would also need to be a clear path from academic research on ISDS to industry standards such as VoiceXML to facilitate adoption.

Various components and techniques of ISDS are needed to handle 'micro-turns'. Challenges here include recognizing and understanding partial user input and back-channels; micro-turn dialogue management that can decide when to back-channel, self-correct and hold-the-floor; incremental NLG that can generate output while the user is still talking; and finally more flexible TTS that can handle barge-in and understand when it has been interrupted.

In summary, in order to achieve highly natural, responsive incremental systems, we propose using data-driven techniques, for which the main issue is lack of data. Carefully crafted task-based human-human data collection and WoZ studies, user simulations, shared data archives, and upgraded industry standards are required for future work in this field.

Acknowledgments

The research leading to this work has received funding from the EC's FP7 programme: (FP7/2011-14) under grant agreement no. 287615 (PARLANCE).

References

Gregory Aist, James Allen, Ellen Campana, Lucian Galescu, Carlos Gomez Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus. 2007. Software architectures for incremental understanding of human speech. In *Proceedings of SemDial / DECALOG*.

James Allen, George Ferguson, and Amanda Stent. 2001. An Architecture For More Realistic Conversational Systems. In *Proc. of Intelligent User Interfaces*.

Timo Baumann, Okko Buss, and David Schlangen. 2011. Evaluation and Optimisation of Incremental Processors. *Dialogue and Discourse*, 2(1).

Okko Buss, Timo Baumann, and David Schlangen. 2010. Collaborating on Utterances with a Spoken Dialogue Systen Using an ISU-based Approach to Incremental Dialogue Management. In *Proc. of SIGDIAL*.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. 2012. Optimising Incremental Generation for Spoken Dialogue Systems: Reducing the Need for Fillers. In *Proc of INLG*, Chicago, Illinois, USA.

Willem Levelt. 1989. *Speaking: From Intenion to Articulation*. MIT Press.

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. 2011. Stability and Accuracy in Incremental Speech Recognition. In *Proc. of SigDial*.

Gabriel Skantze and David Schlangen. 2009. Incremental Dialogue Processing in a Micro-Domain. In *Proc. of EACL*, Athens, Greece.

M.K. Tanenhaus and S. Brown-Schmidt. 2008. Language processing in the natural world. In B.C.M Moore, L.K. Tyler, and W.D. Marslen-Wilson, editors, *The perception of speech: from sound to meaning*, pages 1105–1122.

[2]As was done for CLASSiC project data at: http://www.macs.hw.ac.uk/iLabArchive/CLASSiCProject/Data/login.php