The Future of Spoken Dialogue Systems is in their Past: Long-Term Adaptive, Conversational Assistants

David Schlangen
Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
david.schlangen@uni-bielefeld.de

Abstract

A sketch of dialogue systems as long-term adaptive, conversational agents.

1 Introduction

"*Show me the lecture notes from last year*", you say to your bow-tied virtual assistant. It does, but unfortunately, "*this will not do. Pull up all the new articles I haven't read yet*". Your assistant obliges, pointing your attention to a "*new article from your friend, Jill Gilbert*". A video call later, your lecture preparation is done—Jill will actually give it, via video link—and you go on with your day.

This of course describes the first scene from Apple's "Knowledge Navigator" concept video (Apple Computer Inc., 1987; Colligan, 2011). Not much of what that video showed was actually technically possible at the time, but it captured the promise of personalized natural language interfaces that many people saw and hoped would be realised soon. Having to deal with the constraints of reality, however, research and development of spoken dialogue interfaces had to set itself the more modest aim of replacing, in certain settings, mouse and keyboard, rather than personal assistants.

Recent years have seen two developments that bring that more ambitious goal back into focus. First, the required basic technologies such as speech recognition and speech synthesis have matured to a state where they begin to allow the necessary flexibility of spoken in- and output. Second, it has become not only possible but completely unremarkable for large portion of the population to carry with them sensor-rich, networked computing devices—their smartphones—during large parts of their day.

In this position paper, I'd like to sketch what the opportunities are that this situation offers, for the creation of dialogue systems that are *long-term adaptive* and *conversational*, and act as *assistants*, not interfaces.

2 Long-Term Adaptive ...

The fact that users carry with them the same device (or class of devices; it only matters that access is constant), provides the chance of repeated interactions with what is understood to be the same system. To make use of this, the system must

- learn from errors / miscommunications, by improving internal models (acoustic model, language model, semantic models: how are tasks structured for particular user); and it must
- build up personal common ground:

– What has been refered to previously, and how? Which tasks have been done together, and how?
– Which situations have been shared? (Where a multi-sensor device can have detailed situational information.)

While the first point mostly describes current practice (user adaptation of speech resources), there is much to be explored in the building up of common ground with a technical device.

3 ... Conversational ...

Interaction with these systems must be less driven by fixed system-intiative, and be more conversational:

- User and system must be able to mean more than they say, by making use of context, both from

NAACL-HLT 2012 Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data, pages 11–12,
Montréal, Canada, June 7, 2012.

the ongoing conversation as well as from the common ground that was built up over previous interaction.

• Systems should be responsive, incremental, providing feedback where required; realising a tight interaction loop, not strict turn-based exchanges.

• Things will go wrong, so error handling needs to be graceful and natural, using the full range of conversational repair devices (Schlangen, 2004; Purver, 2004); including handing off tasks to other modalities if expected success rate is low.

• Conversations express and project personality, emotionality, sociality; systems need to model the dynamics of this as part of their modelling of the conversation.

Again, these are active areas of research (for responsive systems, see e.g. (Skantze and Schlangen, 2009; Buß et al., 2010; Schlangen et al., 2010); for error handling / acting under uncertainty, see e.g. (Williams and Young, 2007); for social aspects of dialogue, see e.g. (Kopp, 2010)); pulling them together in this kind of application will likely provide new challenges and insights for all of them.

4 ... Assistants

Of course, the systems will need to provide actual services, for it at all to come to repeated conversations. While providing the services lies outside the domain of speech research, there are some unique requirements that conversational access poses:

• To be usefully embeddable into conversational systems, back-end applications are needed that are interaction-ready; e.g., by providing confindence information about their results, and, building on this, by suggesting ways to improve quality through additional information.

• Not all back-end services are under the control of the application developer or provide APIs, and the semantic web is not going to happen. The reach of a virtual assistant can be increased if it can be *taught* to do tasks like use a website to book a train. Some promising first work in this direction exists (Allen et al., 2007).

5 Resources

Building dialogue systems is always hard, as many different components need to be integrated. Systems as sketched above bring the additional challenge of requiring work on mobile platforms; a framework that provides the required interfaces and infrastructure would be very helpful.

References

James F. Allen, Nathanael Chambers, George Ferguson, Lucian Galescu, Hyuckchul Jung, Mary Swift, and William Taysom. 2007. PLOW: A collaborative task learning agent. In *Proceedings of the National Conference on Artificial Intelligens (AAAI)*, Vancouver, BC, Canada.

Apple Computer Inc. 1987. The knowledge navigator concept video. http://youtu.be/HGYFEI6uLy0.

Okko Buß, Timo Baumann, and David Schlangen. 2010. Collaborating on utterances with a spoken dialogue system using an isu-based approach to incremental dialogue management. In *Proceedings of the SIGdial 2010 Conference*, pages 233–236, Tokyo, Japan, September.

Bud Colligan. 2011. How the knowledge navigator video came about, Nov. http://www.dubberly.com/articles/how-the-knowledge-navigator-video-came-about.html.

Stefan Kopp. 2010. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, 52(6):587–597.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, Unversity of London, London, UK, August.

David Schlangen, Timo Baumann, Hendrik Buschmeier, Okko Buß, Stefan Kopp, Gabriel Skantze, and Ramin Yaghoubzadeh. 2010. Middleware for incremental processing in conversational agents. In *Proceedings of the SIGdial 2010 Conference*, pages 51–54, Tokyo, Japan, September.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th Workshop of the ACL SIG on Discourse and Dialogue*, Boston, USA, April.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 745–753, Athens, Greece, March.

Jason Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):231–422.