Demonstration of the PARLANCE system: a data-driven, incremental, spoken dialogue system for interactive search

Helen Hastie, Marie-Aude Aufaure[*], Panos Alexopoulos, Heriberto Cuayáhuitl, Nina Dethlefs,
Milica Gasic, James Henderson, Oliver Lemon, Xingkun Liu, Peter Mika, Nesrine Ben Mustapha,
Verena Rieser, Blaise Thomson, Pirros Tsiakoulis, Yves Vanrompay, Boris Villazon-Terrazas, Steve Young

email: h.hastie@hw.ac.uk. See `http://parlance-project.eu` for full list of affiliations

Abstract

The PARLANCE system for interactive search processes dialogue at a micro-turn level, displaying dialogue phenomena that play a vital role in human spoken conversation. These dialogue phenomena include more natural turn-taking through rapid system responses, generation of backchannels, and user barge-ins. The PARLANCE demonstration system differentiates from other incremental systems in that it is data-driven with an infrastructure that scales well.

1 Introduction

The PARLANCE system provides interactive search through a Spoken Dialogue System (SDS). This SDS aims to be *incremental* to allow for more natural spoken interaction. Traditionally, the smallest unit of speech processing for interactive systems has been a full utterance with strict, rigid turn-taking. The PARLANCE architecture, however, is an incremental framework that allows for processing of smaller 'chunks' of user input, which enables one to model dialogue phenomena such as barge-ins and backchannels. This work is carried out under the FP7 EC project PARLANCE [1], the goal of which is to develop interactive search through speech in multiple languages. The domain for the demonstration system is interactive search for restaurants in San Francisco. An example dialogue is given in Table 1.

SYS	Thank you for calling the Parlance Restaurant system. You may ask for information by cuisine type, price range or area. How may I help you?
USR	I want to find an Afghan restaurant.........which is in the cheap price range.
SYS	....[uhuhh]........ The Helmand Palace is a cheerful setting for authentic Afghan cuisine.
USR	What is the address and phone number?
SYS	The address 2424 Van Ness Ave ....

Table 1: Example dialogue excerpt for restaurant information in San Francisco

2 Background

Previous work includes systems that can deal with 'micro-turns' (i.e. sub-utterance processing units), resulting in dialogues that are more fluid and responsive. This has been backed up by a large body of psycholinguistic literature that indicates that human-human interaction is in fact incremental (Levelt, 1989).

It has been shown that incremental dialogue behaviour can improve the user experience (Skantze and Schlangen, 2009; Baumann et al., 2011; Selfridge et al., 2011) and enable 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 dialogue phenomena that will be demonstrated by the PARLANCE system include more natural turn-taking through rapid system responses, generation of backchannels and user barge-ins. The system differentiates from other incremental systems in that it is entirely data-driven with an infrastructure that potentially scales well.

3 System Architecture

Figure 1 gives an overview of the PARLANCE system architecture, which maintains

[*] Authors are in alphabetical order

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

Proceedings of the SIGDIAL 2013 Conference, pages 154–156,
Metz, France, 22-24 August 2013. ©2013 Association for Computational Linguistics

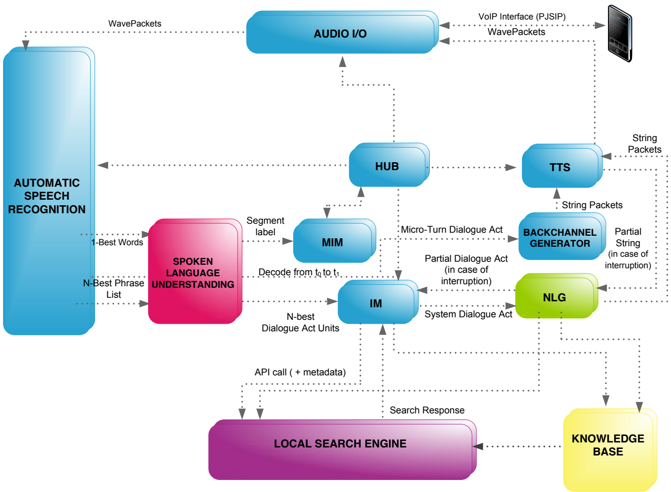

Figure 1: Overview of the Parlance system architecture

the modularity of a traditional SDS while at the same time allowing for complex interaction at the micro-turn level between components.

Each component described below makes use of the PINC (Parlance INCremental) dialogue act schema. In this scheme, a *complete* dialogue act is made up of a set of *primitive* dialogue acts which are defined as *acttype-item* pairs. The PINC dialogue act scheme supports incrementality by allowing SLU to incrementally output primitive dialogue acts whenever a complete *acttype-item* pair is recognised with sufficient confidence. The complete dialogue act is then the set of these primitive acts output during the utterance.

3.1 Recognition and Understanding

The Automatic Speech Recogniser (ASR) and Spoken Language Understanding (SLU) components operate in two passes. The audio input is segmented by a Voice Activity Detector and then coded into feature vectors. For the first pass of the ASR[2], a fast bigram decoder performs continuous traceback generating word by word output. During this pass, while the user is speaking, an SLU module called the "segment decoder" is called incrementally as words or phrases are recognised. This module incrementally outputs the set of primitive dialogue acts that can be detected based on each utterance prefix. Here, the ASR only provides the single best hypothesis, and SLU only outputs a single set of primitive dialogue acts, without an associated probability.

On request from the Micro-turn Interaction Manager (MIM), a second pass can be performed to restore the current utterance using a trigram language model, and return a full distribution over the complete phrase as a confusion network. This is then passed to the SLU module which outputs the set of alternative complete interpretations, each with its associated probability, thus reflecting the uncertainty in the ASR-SLU understanding process.

3.2 Interaction Management

Figure 1 illustrates the role of the Micro-turn Interaction Manager (MIM) component in the overall Parlance architecture. In order to allow for natural interaction, the MIM is responsible for taking actions such as listening to the user, taking the floor, and generating back-channels at the micro-turn level. Given various features from different components, the MIM selects a micro-turn action and sends it to the IM and back-channel generator component to generate a system response.

Micro-turn Interaction Manager A baseline hand-crafted MIM was developed using predefined rules. It receives turn-taking information from the TTS, the audio-output component, the ASR and a timer, and updates turn-taking features. Based on the current features and predefined rules, it generates control signals and sends them to the TTS, ASR, timer and HUB. In terms of micro-turn taking, for example, if the user interrupts the system utterance, the system will stop speaking and listen to the user. The system also outputs a short back-channel and stays in user turn state if the user utterance provides limited information.

Interaction Manager Once the MIM has decided when the system should take the floor, it is the task of the IM to decide what to say. The IM is based on the partially observable

[2] http://mi.eng.cam.ac.uk/research/dialogue/ATK_Manual.pdf

Markov decision process (POMDP) framework, where the system's decisions can be optimised via reinforcement learning. The model adopted for PARLANCE is the Bayesian Update of Dialogue State (BUDS) manager (Thomson and Young, 2010). This POMDP-based IM factors the dialogue state into conditionally dependent elements. Dependencies between these elements can be derived directly from the dialogue ontology. These elements are arranged into a dynamic Bayesian network which allows for their marginal probabilities to be updated during the dialogue, comprising the *belief state*. The belief state is then mapped into a smaller-scale summary space and the decisions are optimised using the natural actor critic algorithm.

HUB The HUB manages the high level flow of information. It receives turn change information from the MIM and sends commands to the SLU/IM/NLG to 'take the floor' in the conversation and generate a response.

3.3 Generation and TTS

We aim to automatically generate language, trained from data, that is (1) grammatically well formed, (2) natural, (3) cohesive and (4) rapidly produced at runtime. Whilst the first two requirements are important in any dialogue system, the latter two are key requirements for systems with incremental processing, in order to be more responsive. This includes generating back-channels, dynamic content reordering (Dethlefs et al., 2012), and surface generation that models coherent discourse phenomena, such as pronominalisation and coreference (Dethlefs et al., 2013). Incremental surfacce generation requires rich context awareness in order to keep track of all that has been generated so far. We therefore treat surface realisation as a sequence labelling task and use Conditional Random Fields (CRFs), which take semantically annotated phrase structure trees as input, in order to represent long distance linguistic dependencies. This approach has been compared with a number of competitive state-of-the art surface realisers (Dethlefs et al., 2013), and can be trained from minimally labelled data to reduce development time and facilitate its application to new domains.

The TTS component uses a trainable HMM-based speech synthesizer. As it is a parametric model, HMM-TTS has more flexibility than traditional unit-selection approaches and is especially useful for producing expressive speech.

3.4 Local Search and Knowledge Base

The domain ontology is populated by the local search component and contains restaurants in 5 regional areas of San Francisco. Restaurant search results are returned based on their longitude and latitude for 3 price ranges and 52 cuisine types.

4 Future Work

We intend to perform a task-based evaluation using crowd-sourced users. Future versions will use a dynamic Knowledge Base and User Model for adapting to evolving domains and personalised interaction respectively.

Acknowledgements

The research leading to this work was funded by the EC FP7 programme FP7/2011-14 under grant agreement no. 287615 (PARLANCE).

References

T. Baumann, O. Buss, and D. Schlangen. 2011. Evaluation and Optimisation of Incremental Processors. *Dialogue and Discourse*, 2(1).

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

N. Dethlefs, H. Hastie, H. Cuayáhuitl, and O. Lemon. 2013. Conditional Random Fields for Responsive Surface Realisation Using Global Features. In *Proceedings of ACL*, Sofia, Bulgaria.

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

E. Selfridge, I. Arizmendi, P. Heeman, and J. Williams. 2011. Stability and Accuracy in Incremental Speech Recognition. In *Proceedings of SIGDIAL*, Portland, Oregon.

G. Skantze and D. Schlangen. 2009. Incremental Dialogue Processing in a Micro-Domain. In *Proceedings of EACL*, Athens, Greece.

B Thomson and S Young. 2010. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. *Computer Speech and Language*, 24(4):562–588.