The PARLANCE Mobile Application for Interactive Search in English and Mandarin

Helen Hastie, Marie-Aude Aufaure*, Panos Alexopoulos,
Hugues Bouchard, Catherine Breslin, Heriberto Cuayáhuitl, Nina Dethlefs,
Milica Gašić, James Henderson, Oliver Lemon, Xingkun Liu, Peter Mika, Nesrine Ben Mustapha,
Tim Potter, Verena Rieser, Blaise Thomson, Pirros Tsiakoulis, Yves Vanrompay,
Boris Villazon-Terrazas, Majid Yazdani, Steve Young and Yanchao Yu

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

Abstract

We demonstrate a mobile application in English and Mandarin to test and evaluate components of the PARLANCE dialogue system for interactive search under real-world conditions.

1 Introduction

With the advent of evaluations "in the wild", emphasis is being put on converting research prototypes into mobile applications that can be used for evaluation and data collection by real users downloading the application from the market place. This is the motivation behind the work demonstrated here where we present a modular framework whereby research components from the PARLANCE project (Hastie et al., 2013) can be plugged in, tested and evaluated in a mobile environment.

The goal of PARLANCE is to perform interactive search through speech in multiple languages. The domain for the demonstration system is interactive search for restaurants in Cambridge, UK for Mandarin and San Francisco, USA for English. The scenario is that Mandarin speaking tourists would be able to download the application and use it to learn about restaurants in English speaking towns and cities.

2 System Architecture

Here, we adopt a client-server approach as illustrated in Figure 1 for Mandarin and Figure 2 for English. The front end of the demonstration system is an Android application that calls the Google Automatic Speech Recognition (ASR) API and sends the recognized user utterance to a server running the Interaction Manager (IM), Spoken Language Understanding (SLU) and Natural Language Generation (NLG) components.

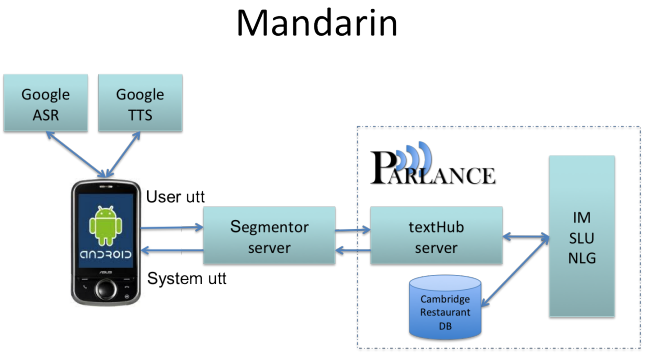

Figure 1: Overview of the PARLANCE Mandarin mobile application system architecture

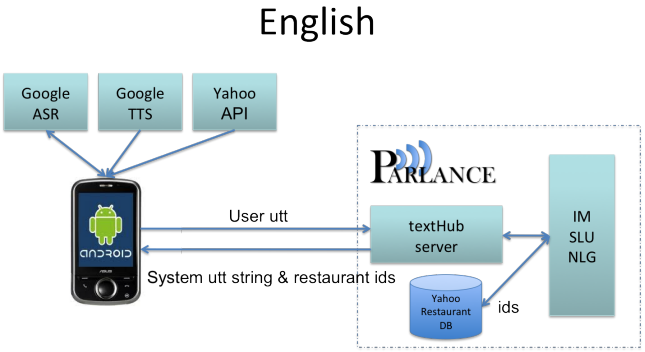

Figure 2: Overview of the PARLANCE English mobile application system architecture extended to use the Yahoo API to populate the application with additional restaurant information

When the user clicks the Start button, a dialogue session starts. The phone application first connects to the PARLANCE server (via the Java Socket Server) to get the initial system greeting which it speaks via the Google

*Authors are in alphabetical order

Proceedings of the SIGDIAL 2014 Conference, pages 260–262,
Philadelphia, U.S.A., 18-20 June 2014.

Text-To-Speech (TTS) API. After the system utterance finishes the recognizer starts to listen for user input to send to the SLU component. The SLU converts text into a semantic interpretation consisting of a set of triples of communicative function, attribute, and (optionally) value[1]. Probabilities can be associated with candidate interpretations to reflect uncertainty in either the ASR or SLU. The SLU then passes the semantic interpretation to the IM within the same server.

Chinese sentences are composed of strings of characters without any space to mark words as other languages do, for example:

```
Mandarin: 我想预订中国餐馆
English: I want to book a Chinese restaurant.
```

In order to correctly parse and understand Chinese sentences, Chinese word segmentations must be performed. To do this segmentation, we use the Stanford Chinese word segmentor[2], which relies on a linear-chain conditional random field (CRF) model and treats word segmentation as a binary decision task. The Java Socket Server then sends the segmented Chinese sentence to the SLU on the server.

The IM then selects a dialogue act, accesses the database and in the case of English passes back the list of restaurant identification numbers (ids) associated with the relevant restaurants. For the English demonstration system, these restaurants are displayed on the smart phone as seen in Figures 4 and 5. Finally, the NLG component decides how best to realise the restaurant descriptions and sends the string back to the phone application for the TTS to realise. The example output is illustrated in Figure 3 for Mandarin and Figure 4 for English.

As discussed above, the PARLANCE mobile application can be used as a test-bed for comparing alternative techniques for various components. Here we discuss two such components: IM and NLG.

[1]This has been implemented for English; Mandarin uses the rule-based Phoenix parser.

[2]http://nlp.stanford.edu/projects/chinese-nlp.shtml

Figure 3: Screenshot and translation of the Mandarin system

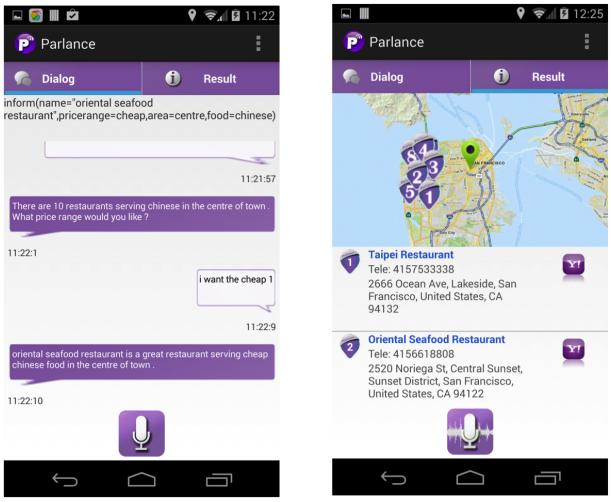

Figure 4: Screenshot of dialogue and the list of recommended restaurants shown on a map and in a list for English

2.1 Interaction Management

The PARLANCE Interaction Manager is based on the partially observable 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. In the PARLANCE application, hand-crafted policies

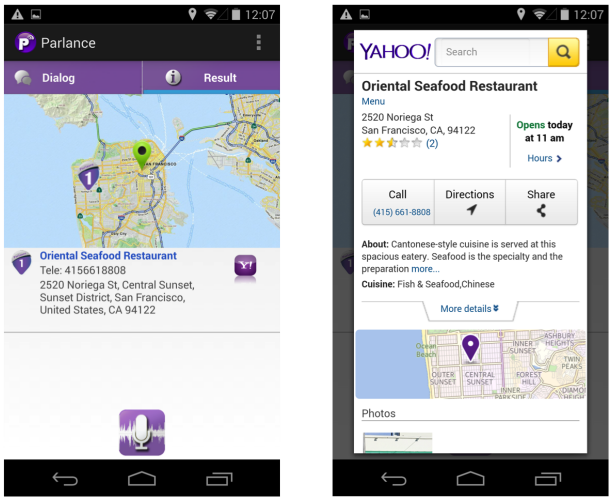

Figure 5: Screenshot of the recommended restaurant for the English application

can be compared to learned ones.

2.2 Natural Language Generation

As mentioned above, the server returns the string to be synthesised by the Google TTS API. This mobile framework allows for testing of alternative approaches to NLG. In particular, we are interested in comparing a surface realiser that uses CRFs against a template-based baseline. The CRFs take semantically annotated phrase structure trees as input, which it uses to keep track of rich linguistic contexts. Our approach has been compared with a number of competitive state-of-the art surface realizers (Dethlefs et al., 2013), and can be trained from example sentences with annotations of semantic slots.

2.3 Local Search and Knowledge Base

For the English system, the domain database is populated by the search Yahoo API (Bouchard and Mika, 2013) with restaurants in San Fransisco. These restaurant search results are returned based on their longitude and latitude within San Francisco for 5 main areas, 3 price categories and 52 cuisine types containing around 1,600 individual restaurants.

The Chinese database has been partially translated from an English database for restaurants in Cambridge, UK and search is based on 3 price categories, 5 areas and 35 cuisine types having a total of 157 restaurants. Due to the language-agnostic nature of the PARLANCE system, only the name and address fields needed to be translated.

3 Future Work

Investigating application side audio compression and audio streaming over a mobile internet connection would enable further assessment of the ASR and TTS components used in the original PARLANCE system (Hastie et al., 2013). This would allow for entire research systems to be plugged directly into the mobile interface without the use of third party ASR and TTS.

Future work also involves developing a feedback mechanism for evaluation purposes that does not put undue effort on the user and put them off using the application. In addition, this framework can be extended to leverage hyperlocal and social information of the user when displaying items of interest.

Acknowledgements

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

References

H. Bouchard and P. Mika. 2013. Interactive hyperlocal search API. Technical report, Yahoo Iberia, August.

N. Dethlefs, H. Hastie, H. Cuayáhuitl, and O. Lemon. 2013. Conditional Random Fields for Responsive Surface Realisation Using Global Features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

H. Hastie, M.A. Aufaure, P. Alexopoulos, H. Cuayáhuitl, N. Dethlefs, M. Gasic, J. Henderson, O. Lemon, X. Liu, P. Mika, N. Ben Mustapha, V. Rieser, B. Thomson, P. Tsiakoulis, Y. Vanrompay, B. Villazon-Terrazas, and S. Young. 2013. Demonstration of the PARLANCE system: a data-driven incremental, spoken dialogue system for interactive search. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, Metz, France, August.

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.