The Map Task Dialogue System: A Test-bed for Modelling Human-Like Dialogue

Raveesh Meena **Gabriel Skantze** **Joakim Gustafson**
KTH Speech, Music and Hearing
Stockholm, Sweden
raveesh@csc.kth.se, gabriel@speech.kth.se, jocke@speech.kth.se

Abstract

The demonstrator presents a test-bed for collecting data on human–computer dialogue: a fully automated dialogue system that can perform Map Task with a user. In a first step, we have used the test-bed to collect human–computer Map Task dialogue data, and have trained various data-driven models on it for detecting feedback response locations in the user's speech. One of the trained models has been tested in user interactions and was perceived better in comparison to a system using a random model. The demonstrator will exhibit three versions of the Map Task dialogue system—each using a different trained data-driven model of *Response Location Detection*.

1 Introduction

A common procedure in modelling human-like dialogue systems is to collect data on human–human dialogue and then train models that predict the behaviour of the interlocutors. However, we think that it might be problematic to use a corpus of human–human dialogue as a basis for implementing dialogue system components. One problem is the interactive nature of the task. If the system produces a slightly different behaviour than what was found in the original data, this would likely result in a different behaviour in the interlocutor. Another problem is that humans are likely to behave differently towards a system as compared to another human (even if a more human-like behaviour is being modelled). Yet another problem is that much dialogue behaviour is optional and therefore makes the actual behaviour hard to use as a gold standard.

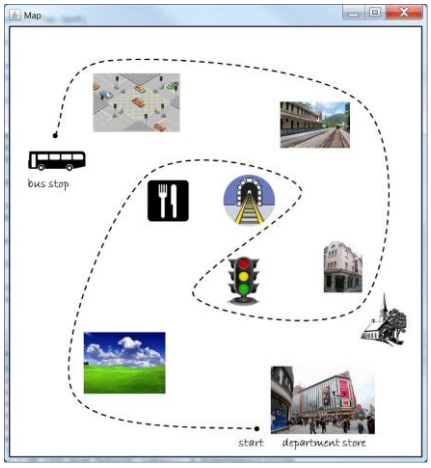

Figure 1: The Map Task system user interface

To improve current systems, we need both a better understanding of the phenomena of human interaction, better computational models and better data to build these models. An alternative approach that has proven to be useful is to train models on human–computer dialogue data collected through *Wizard-of-Oz* studies (Dahlbäck et al., 1993). However, the methodology might be hard to use when the issue under investigation is time-critical behaviour such as back-channels.

A third alternative is to use a *boot-strapping* procedure, where more and more advanced (or human-like) versions of the system are built iteratively. After each iteration, users interact with the system and data is collected. This data is then used to train/improve data-driven models of interaction in the system. A problem here, however, is how to build the first iteration of the system, since many components, e.g., Automatic Speech Recognition (ASR), need some data to be useful at all.

In this demonstration we present a test-bed for collecting data on time-critical human–computer dialogue phenomena: a fully automated dialogue system that can perform the Map Task with a

Proceedings of the SIGDIAL 2013 Conference, pages 366–368,
Metz, France, 22-24 August 2013.

user (Skantze, 2012). In a first step, following the boot-strapping procedure, we collected human–computer Map Task dialogue data using this test-bed and then trained various data-driven models on this data for detecting feedback response locations in user's speech. A trained model has been implemented and evaluated in interaction with users—in the same environment used for collecting the data (Meena et al., in press). The demonstrator will exhibit three versions of the Map Task dialogue system—each using a different trained data-driven model of *Response Location Detection* (RLD).

2 The Map Task Dialogue System

Map Task is a common experimental paradigm for studying human–human dialogue. In our set-up, the user (the information *giver*) is given the task of describing a route on a map to the system (the information *follower*). The choice of Map Task is motivated partly because the system may allow the user to keep the initiative during the whole dialogue, and thus only produce responses that are not intended to take the initiative, most often some kind of feedback. Thus, the system might be described as an *attentive listener*.

The basic components of the system can be seen in Figure 2. Dashed lines indicate components that were not part of the first iteration of the system (used for data collection), but which have been used in the second iteration of the system that uses a model trained on the collected data. To make the human–computer Map Task dialogue feasible without any full speech understanding we have implemented a trick: the user is presented with a map on a screen (see Figure 1) and instructed to move the mouse cursor along the route as it is being described. The user is told that this is for logging purposes, but the real reason for this is that the system tracks the mouse position and thus knows what the user is currently talking about. It is thereby possible to produce a coherent system behaviour without any speech recognition at all, only basic speech detection. This often results in a very realistic interaction[1].

The system uses a simple energy-based speech detector to chunk the user's speech into inter-pausal units (IPUs), that is, periods of speech that contain no sequence of silence longer than 200 ms. Such a short threshold allows the system to give backchannels (seemingly) while the user is speaking or take the turn with barely any gap. The end of an IPU is a candidate for the RLD model to identify whether it is a Response Location (RL).

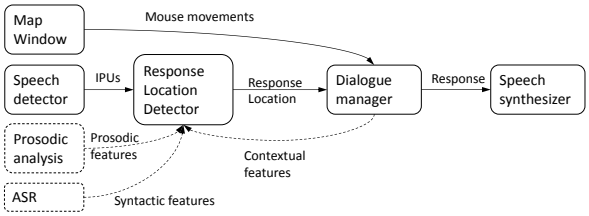

Figure 2: The basic components of the system.

Each time the RLD detects a RL, the dialogue manager produces a response, depending on the current state of the dialogue and the position of the mouse cursor. Table 1 shows an example Map Task dialogue illustrating the different types of responses the system could produce. The dialogue manager always starts with an Introduction and ends with an Ending, once the mouse cursor has reached the destination. Between these, it selects from the other responses (Reprise Fragment, Clarification and Backchannel), partly randomly, but also depending on the length of the last user turn (longer turn triggers Repeat or Restart) and the current mouse location (e.g., a Guess if mouse hovers over the same area for some time).

Table 1: An example Map Task dialogue between a user (U) and system (S) [*User utterances with ASR error(s) are indicated with a **]

S	Introduction	Could you help me to find my way to the bus stop?
U		start from the department store
S	Backchannel	yeah
U		and eh
S	Guess	Should I start by going west?
U		yeah do that
U		then you will get to a meadow and when you get to the meadow **you get to a meadow until you get to the meadow*
S	Repeat	Eh, could you repeat that?
U		you go straight and you see a meadow on your right side
S	Clarification Request	A green field?
U		ehm yeah a field
S	Backchannel	mhm
U		pass the meadow and turn right so you are going north **the meadow and turn right to your going north*
S	Backchannel	okay
U		you get to a restaurant

[1] An example video can be seen at http://www.youtube.com/watch?v=MzL-B9pVbOE.

S	Reprise Fragment	A restaurant
U		then you have to turn right and that will take you to **you have turned right and at will take you turn*
S	Restart	Eh, I think I lost you at the restaurant, could you take that again?
…		
U		at the junction go south and then you will get to the bus stop
S	Ending	Okay, thanks a lot.

3 Data-driven models of RLD

Using the system described in the previous section a corpus of 50 human–computer Map Task dialogue was collected and used to train a data-driven model of RLD. Since we didn't have a sophisticated model of RLD during the first iteration a naïve model was used. This model would wait for a random period between 0 and 800 ms after an IPU ended. If no new IPUs were initiated during this period, a RL was detected. Each IPU in the corpus was then manually labelled as either Hold (a response would be inappropriate) or Respond (a response is expected) type. On this data various models were trained on online extractable features—covering syntax, context and prosody. Table 2 illustrates the performance of the various models. Going a step further, model #6 was deployed in the Map Task dialogue system (with an ASR component) and evaluated in user interactions. The result suggests that the trained model provide for smooth turn-transitions in contrast to the Random model (Meena et al., in press).

Table 2: Performance of various models of RLD [*NB: Naïve Bayes; SVM: Support Vector Machine; Models with * will be exhibited in the demonstration*]

#	*RLD model*	*% accuracy (on ASR results)*
1*	Random	50.79% majority class baseline
2	Prosody	64.5% (SVM learner)
3	Context	64.8% (SVM learner)
4*	Prosody + Context	69.1% (SVM learner)
5	Syntax	81.1% (NB learner)
6*	Syntax + Prosody + Context	82.0 % (NB learner)

4 Future applications

The Map Task test-bed presented here has the potential for modelling other human-like conversational behaviour in dialogue systems:

***Clarification strategies*:** by deploying explicit (*did you mean turn right?*) and implicit (a reprise such as *turn right*) or elliptical ('*right?*') clarification forms in the *grounding* process one could investigate the efficiency and effectively of these human-like clarification strategies.

User utterance completion: It has been suggested that completion of user utterances by a dialogue system would result in human-like conversational interactions. However, completing user's utterance at every opportunity may not be the best strategy (DeVault et al., 2009). The presented system could be used to explore when it is appropriate to do so. We have observed in our data that the system dialogue acts Guess (cf. Table 1) and Reprise often helped the dialogue proceed further – by completing user utterances – when the user had difficulty describing a landmark on a route.

Visual cues: the system could be integrated in a robotic head, such as Furhat (Al Moubayed et al., 2013), and visual cues from the user could be used for improving the current model of RLD. This could be used further to explore the use of extra-linguistic system behaviours, such as head nods and facial gestures, as feedback responses.

Acknowledgement

This work is supported by the Swedish research council (VR) project I*ncremental processing in multimodal conversational systems* (2011-6237)

References

Al Moubayed, S., Skantze, G., & Beskow, J. (2013). The Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multiparty Interaction. *International Journal of Humanoid Robotics, 10*(1).

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies – why and how. In *Proceedings from the 1993 International Workshop on Intelligent User Interfaces* (pp. 193-200).

DeVault, D., Sagae, K., & Traum, D. (2009). Can I Finish? Learning When to Respond to Incremental Interpretation Results in Interactive Dialogue. In *Proceedings of SIGdial* (pp. 11-20). London, UK.

Meena, R., Skantze, G., & Gustafson, J. (in press). A Data-driven Model for Timing Feedback in a Map Task Dialogue System. To be published in *14th Annual Meeting of the Special Interest Group on Discourse and Dialogue - SIGdial*. Metz, France.

Skantze, G. (2012). A Testbed for Examining the Timing of Feedback using a Map Task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Portland, OR.