Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 903–922

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Data-driven models for timing feedback responses in a Map Task dialogue system☆

Raveesh Meena*, Gabriel Skantze, Joakim Gustafson

KTH Royal Institute of Technology, Department of Speech, Music and Hearing, Lindstedtsvägen 24, 10044 Stockholm, Sweden

Received 8 October 2013; received in revised form 17 December 2013; accepted 3 February 2014
Available online 14 February 2014

Abstract

Traditional dialogue systems use a fixed silence threshold to detect the end of users' turns. Such a simplistic model can result in system behaviour that is both interruptive and unresponsive, which in turn affects user experience. Various studies have observed that human interlocutors take cues from speaker behaviour, such as prosody, syntax, and gestures, to coordinate smooth exchange of speaking turns. However, little effort has been made towards implementing these models in dialogue systems and verifying how well they model the turn-taking behaviour in human–computer interactions. We present a data-driven approach to building models for online detection of suitable feedback response locations in the user's speech. We first collected human–computer interaction data using a spoken dialogue system that can perform the Map Task with users (albeit using a trick). On this data, we trained various models that use automatically extractable prosodic, contextual and lexico-syntactic features for detecting response locations. Next, we implemented a trained model in the same dialogue system and evaluated it in interactions with users. The subjective and objective measures from the user evaluation confirm that a model trained on speaker behavioural cues offers both smoother turn-transitions and more responsive system behaviour.

Keywords: Spoken dialogue systems; Timing feedback; Turn-taking; User evaluation

1. Introduction

Traditionally, dialogue systems have rested on a very simple model for turn-taking: the system uses a fixed silence threshold to detect the end of the user's utterance, after which the system responds. However, this model does not capture human–human dialogue very accurately. Sometimes a speaker simply hesitates and no turn-change is intended; sometimes the turn changes after barely any silence (Sacks et al., 1974). Therefore, such a simplistic model can result in systems that frequently produce responses at inappropriate occasions, or produce delayed responses or no response at all when expected, thereby causing the system to be perceived as interruptive or unresponsive. Related to the problem of turn-taking is that of *backchannels* (Yngve, 1970). Backchannel feedback – short acknowledgements such as *uh-huh* or *mm-hmm* – are used by human interlocutors to signal continued attention to the speaker, without claiming

☆ This paper has been recommended for acceptance by A. Potamianos.

* Corresponding author. Tel.: +46 08 790 7872.
E-mail addresses: raveesh@csc.kth.se (R. Meena), gabriel@speech.kth.se (G. Skantze), jocke@speech.kth.se (J. Gustafson).

http://dx.doi.org/10.1016/j.csl.2014.02.002
0885-2308/

the conversational floor. If a dialogue system is to manage smooth exchange of speaking turns and provide backchannel feedback without being interruptive, it must be able to first identify suitable locations in the user's speech to do so.

Human conversational partners are skilled at managing smooth turn-transitions. Duncan (1972) observed that human interlocutors continuously monitor cues, such as content, syntax, intonation, paralanguage, and body motion, in parallel to manage turn-taking. Similar observations have been made in various other studies investigating the turn-taking and back-channelling phenomena in human conversations. Ward (1996) has suggested that a low pitch region is a good cue that backchannel feedback is appropriate. On the other hand, Koiso et al. (1998) have argued that both syntactic and prosodic features make significant contributions in identifying turn-taking and backchannel relevant places. Cathcart et al. (2003) have shown that syntax in combination with pause duration is a strong predictor for backchannel continuers. Gravano and Hirschberg (2011) identified seven turn-yielding and six backchannel-inviting cues spanning over prosodic, acoustic, and lexico-syntactic feature that could be used for recognition and generation of turns and backchannel.

However, there is a general lack of studies on how such models could be used online in dialogue systems and to what extent that would improve the interaction. There are two problems in doing so. First, the data used in the studies mentioned above are from human–human dialogue and it is not obvious to what extent the models derived from such data transfers to human–machine dialogue. Second, many of the features used in the proposed models were manually extracted. This is especially true for the transcription of utterances, but several studies also rely on manually annotated prosodic features.

This article builds upon our earlier work on automatic detection of relevant feedback response locations in the user's speech (Meena et al., 2013). In our earlier work, we presented a data-driven model of what we call *Response Location Detection* (RLD). The model is fully online and only relies on automatically extractable features – comprising syntax, prosody and context. The model has been trained on human–computer dialogue data and has been implemented in a dialogue system that is in turn evaluated by users. The setting is that of a Map Task, in which the user describes a route and the system may respond with, for example, acknowledgements or clarification requests. The presented approach exemplifies a boot-strapping procedure where more and more advanced versions of the system are built iteratively. After each iteration, users interact with the system and data is collected, which is then used to improve the data-driven models in the system. In this article, we extend the analysis by exploring the use of other classifiers, testing the robustness of a model against ASR errors, and further analysing the objective and subjective performance of the trained model used in user evaluation.

In Section 2, we discuss previous studies on cues that human interlocutors use to manage turn-taking and backchannels. We will also discuss some of the proposed computational models. In Section 3, we describe the test bed that we used for boot-strapping a Map Task dialogue system to collect data and develop an improved incremental version of the system. In Section 4, we will discuss the various data-driven models that we have trained in this work. We describe the various features we have explored, and discuss their performance using various learning algorithms on our data for online use. In Section 5, we discuss the subjective and objective evaluation schemes used for verifying the contributions of a trained model in user interactions. Finally, in Section 6, we discuss the key contributions and limitations of the models presented in this paper, and conclude with some ideas for future extensions of this work.

2. Background

Two influential theories that have examined the turn-taking mechanism in human conversations are the signal-based mechanism of Duncan (1972) and the rule-based mechanism proposed by Sacks et al. (1974). According to Duncan, "the turn-taking mechanism is mediated through signals composed of clear-cut behavioural cues, considered to be perceived as discrete". Duncan identified six discrete behavioural cues that a speaker may use to signal the intent to yield the turn, involving prosody, syntax and gestures. Speakers may display these behavioural cues either singly or together, and when displayed together the likelihood that the listener attempts to take the turn increases. According to the rule-based mechanism of Sacks et al. (1974) turn-taking is regulated by applying rules (e.g. "one party at a time") at Transition-Relevance Places (TRPs) – possible completion points of basic units of turns, in order to minimize gaps and overlaps. The basic units of turns (or turn-constructional units) include sentential, clausal, phrasal, and lexical constructions.