Chapter 14

EVALUATING DIALOGUE STRATEGIES IN MULTIMODAL DIALOGUE SYSTEMS

Steve Whittaker, Marilyn Walker

University of Sheffield, Sheffield, United Kingdom
{s.whittaker,m.a.walker}@sheffield.ac.uk

Abstract

Previous research suggests that multimodal dialogue systems providing both speech and pen input, and outputting a combination of spoken language and graphics, are more robust than unimodal systems based on speech or graphics alone (André, 2002; Oviatt, 1999). Such systems are complex to build and significant research and evaluation effort must typically be expended to generate well-tuned modules for each system component. This chapter describes experiments utilising two complementary evaluation methods that can expedite the design process: (1) a Wizard-of-Oz data collection and evaluation using a novel Wizard tool we developed; and (2) an Overhearer evaluation experiment utilising logged interactions with the real system. We discuss the advantages and disadvantages of both methods and summarise how these two experiments have informed our research on dialogue management and response generation for the multimodal dialogue system MATCH.

Keywords: User modelling; Natural language generation; Wizard-of-Oz experiments; Overhearer method; User-adaptive generation; Multiattribute decision theory.

1. Introduction

Multimodal dialogue systems promise users mobile access to a complex and constantly changing body of information. However, mobile information access devices such as PDAs, tablet PCs, and next-generation phones offer limited screen real-estate and no keyboard or mouse. Previous research suggests that spoken language interaction is highly desirable for such systems, and that systems that provide both speech and pen input, and that output a combination of spoken language and graphics, are more robust than unimodal systems (André, 2002; Oviatt, 1999). However, such systems are complex to build and typically significant research and evaluation effort must be expended.
to generate well-tuned modules for each system component. Furthermore, during the development process, it is necessary to evaluate individual components to inform the design process before the whole system is robust enough for data collection with real users. This chapter describes experiments utilising two complementary evaluation methods that can be applied to collect information useful for design during the design process itself. We summarise how we have used these methods to inform our research on improved algorithms for (a) dialogue management and (b) generation for information presentation in multimodal dialogue.

Our testbed application is MATCH (Multimodal Access To City Help), a dialogue system providing information for New York City (Johnston and Bangalore, 2000; Bangalore and Johnston, 2000; Johnston and Bangalore, 2001; Johnston et al., 2002). MATCH runs standalone on a Fujitsu PDA, as shown in Figure 1, yet can also run in client-server mode across a wireless network. MATCH provides users with mobile access to restaurant, entertainment and transportation information for New York City (NYC). Figure 2 depicts the multimodal architecture supporting MATCH, which consists of a series of agents which communicate through a facilitator MCUBE. Figure 2 shows modules that support users in specifying inputs via speech, gesture, handwriting or by a combination of these. Other modules support output generated in speech, using a graphical display, or a combination of both these modes. Automatic Speech Recognition (ASR) is provided by AT&T's Watson engine (Sharp et al., 1997), and the Text-To-Speech (TTS) is based on AT&T's Natural Voices (Beutnagel et al., 1999). MATCH uses a finite-state approach (MMFST) to parse, integrate, and understand multimodal and unimodal inputs (Johnston and Bangalore, 2000). See (Johnston et al., 2002) for more architectural detail.

Our primary research focus has been to provide MATCH with improved capabilities for dialogue management and generation during the information presentation portion of the dialogue, i.e., research on the Multimodal Dialogue Manager (MDM), Text Planner and Multimodal Generator in Figure 2. During this dialogue phase, the system retrieves from its database a set of options that match the user’s constraints. The user must then evaluate the various options before selecting one. Even in a multimodal system such as MATCH, that displays some information graphically, this is a complex and time-consuming process: the user must browse a list or graphical representation of the options and access information about each one.

For example, consider a user’s request to Show Italian Restaurants in the West Village. Figure 3 shows the large number of highlighted options generated as a graphical response. To make an informed choice, the user has to access more detailed information about each individual restaurant either with speech or by graphical browsing. In addition to the tedium of sequentially