Interacting with Computers 24 (2012) 482–498

Contents lists available at SciVerse ScienceDirect

Interacting with Computers

journal homepage: www.elsevier.com/locate/intcom

Towards building intelligent speech interfaces through the use of more flexible, robust and natural dialogue management solutions ☆

Fernando Fernández-Martínez *, J. Ferreiros, J.M. Lucas-Cuesta, J.M. Montero-Martínez, R. San-Segundo, R. Córdoba

Grupo de Tecnología del Habla, GTH (Speech Technology Group), Universidad Politécnica de Madrid, UPM, Ciudad Universitaria s/n, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 22 October 2010
Received in revised form 26 June 2012
Accepted 20 September 2012
Available online 22 October 2012

Keywords:
Spoken dialogue systems
Mixed initiative
Bayesian Networks
Contextual information
Usability
Electronic devices control

ABSTRACT

In this paper a Bayesian Networks-based solution for dialogue modelling is presented. This solution is combined with carefully designed contextual information handling strategies. With the purpose of validating these solutions, and introducing a spoken dialogue system for controlling a Hi-Fi audio system as the selected prototype, a real-user evaluation has been conducted. Two different versions of the prototype are compared. Each version corresponds to a different implementation of the algorithm for the management of the actuation order, the algorithm for deciding the proper order to carry out the actions required by the user. The evaluation is carried out in terms of a battery of both subjective and objective metrics collected from speakers interacting with the Hi-Fi audio box through predefined scenarios. Defined metrics have been specifically adapted to measure: first, the usefulness and the actual relevance of the proposed solutions, and, secondly, their joint performance through their intelligent combination mainly measured as the level achieved with regard to the user satisfaction. A thorough and comprehensive study of the main differences between both approaches is presented. Two-way analysis of variance (ANOVA) tests are also included to measure the effects of both: the system used and the type of scenario factors, simultaneously. Finally, the effect of bringing this flexibility, robustness and naturalness into our home dialogue system is also analyzed through the results obtained. These results show that the intelligence of our speech interface has been well perceived, highlighting its excellent ease of use and its good acceptance by users, therefore validating the approached dialogue management solutions and demonstrating that a more natural, flexible and robust dialogue is possible thanks to them.

1. Introduction

Speech is the most widely used natural means of communication between people. Speech also is of increasing importance as a user–machine interface. As a result of the knowledge and the experience accumulated during almost half a century of research in the field of speech technology, the time has now come to design automated dialogue systems that make use of the communicative aspects of speech. In particular, it is essential to incorporate into the design of these systems some ideas related to the concept of *ambient intelligence* (AmI) (Augusto, 2007; Aarts and de, 2009), for providing intelligent interfaces that are able to conduct a natural dialogue, including negotiations in order to achieve the goals required by users.

A dialogue system can be seen as a computer application that enables interaction and communication between users and machines as naturally as possible. Besides the typical recognition and text-to-speech conversion modules and other components, dialogue systems usually contain a module called dialogue manager (DM). This module is responsible for a dual task: to interpret the intention of the user and to decide how to continue the dialogue.

To provide users successfully with answers resembling a human–human interaction as much as possible, we believe that the design of a dialogue system should be approached from both a theoretical and practical point of view. Thus, we must pay attention not only to dialogue management and modelling, but also to the enhancement of these models with knowledge about the specific tasks of the dialogue and the application domain (i.e. task and domain models). Thus, it is feasible to develop procedures that support the user–machine interaction with useful elements of

☆ This paper has been recommended for acceptance by D. Murray.

* Corresponding author. Tel.: +34 91 549 57 00x4228; fax: +34 91 336 73 23.
E-mail addresses: ffm@die.upm.es (F. Fernández-Martínez), jfl@die.upm.es (J. Ferreiros), juanmak@die.upm.es (J.M. Lucas-Cuesta), juancho@die.upm.es (J.M. Montero-Martínez), lapiz@die.upm.es (R. San-Segundo), cordoba@die.upm.es (R. Córdoba).

0953-5438/$ - see front matter © 2012 British Informatics Society Limited. All rights reserved.
http://dx.doi.org/10.1016/j.intcom.2012.09.003

communication for carrying out a collaborative and cooperative dialogue.

Although the interest in ambient intelligence in the domain of home dialogue systems is growing significantly (Berton et al., 2006), the benefits that this intelligence might bring are not often demonstrated or clearly identified (de Ruyter et al., 2005).

In this work we are presenting the evaluations that we have conducted to examine the effects of our dialogue management solutions, but more specifically to address the following research questions:

- Will the level of flexibility (i.e. absence of rules or restrictions that might restrict the dialogue in any way), robustness (i.e. ability to recover missing information and to handle errors when the user input has ASR and SLU errors occurred by noises or unexpected inputs) and naturalness (i.e. ability to negotiate with the user in achieving the dialogue goals similarly to the way a human would help) achieved in the home dialogue system be perceived (e.g. by means of a good user satisfaction rate)?
- What is the effect of bringing this intelligence into a home dialogue system on the perception of the ease of use of the interactive systems in the environment?
- Will the acceptance of home dialogue systems increase if the proposed solutions are implemented in these systems?

Finding performance figures from real-world applications that can be extrapolated to other systems or be accepted worldwide is a really complicated task, as all of them are directly related to a specific dialogue system. Nonetheless, there is a general agreement on *usability* as the most important performance figure (Schulz and Donker, 2006; Turunen et al., 2006; Raux et al., 2005; Walker et al., 2000), even more than others widely used such as *naturalness* or *flexibility*.

Several usability guidelines that should be taken into account in the design of dialogue systems and their evaluation, especially for multi-modal systems, have been reviewed in Dybkjaer et al. (2004). Therefore, besides quality and efficiency metrics, automatically logged or computed, subjective tests have also been carried out in order to assess the impact of the capabilities of the system on user satisfaction and to get a valuable insight into the shortcomings and advantages of the proposed solutions.

The paper is organized as follows: first, the home dialogue system used to answer the aforementioned research questions is described. A couple of subsections, 2.6 and 2.7, introduce the two different versions of our developed prototype, HIFI-AV1 and HIFI-AV2, discussing alternative approaches for the management of the actuation order. The following section describes the experimental framework used to evaluate the performance of the proposed solutions. In the sections which follow, we successively present and discuss the results obtained for both versions of our system (i.e. HIFI-AV1 vs. HIFI-AV2). Finally, the paper concludes by highlighting some conclusions specifically addressing the aforementioned research questions. Some possible future lines of research are also proposed.

2. The dialogue management solution

The major advantage of classic knowledge-based dialogue management solutions (Bui, 2006; Lee et al., 2010) like: finite state automata or FSMs, script based systems or dialogue plans, etc., is the simplicity. They are suitable for simple dialogue systems with well-structured task. However, these approaches lack of flexibility, naturalness, and applicability to other domains.

As an alternative to these we are presenting a dialogue solution based on Bayesian Networks (BNs), that allows a greater flexibility and naturalness by appropriately defining dialogue as the interaction with an inference system (Meng et al., 2003).

This solution can be classified as a data-driven dialogue management approach (Lee et al., 2010) that, although requires time consuming data annotation, enables training to be done automatically and requiring little human supervision.

The framework applies statistically data-driven and theoretically principled dialogue modelling to dynamically allow changes to the dialogue strategy. Stochastic dialogue modelling using reinforcement learning (RL) based on Markov decision processes (MDPs) (Levin and Pieraccini, 1997) or partially observable MDPs (POMDPs) (Williams and Young, 2007) are another alternative approaches within this framework.

2.1. A spoken dialogue interface for a Hi-Fi audio system

The conversational interface that we are presenting (Fernández-Martínez et al., 2005) was included as part of the EDECAN project.[1] It allows users to control a Hi-Fi system from natural language sentences, differentially to other typical control systems based on simple commands. Thus, users can feel free to give several complex commands from a single sentence. Moreover, they neither have to memorize any command list nor use specific vocabulary cum syntax in order to control the system successfully.

The Hi-Fi audio system we are controlling is a commercial system made up of a compact disc player (with a charger of three discs), two tapes deck and a radio receiver. This system can be controlled by an infra-red (IR) remote control. Instead, users are going to control the Hi-Fi system using a microphone. Our interface translates the speech into IR commands in order to carry out different operations or actions on the system. This translation is made so that the appropriate IR commands are sent according to the intention of the user.

2.2. The spoken dialogue system

A dialogue can be defined as the verbal interaction that the user has with the system with the purpose of achieving some goals related to the control of the Hifi equipment. This interaction takes place on a turn basis (a dialogue turn can be defined as one user input action and the corresponding system output). Its length, typically measured either in terms of time or simply as the number of turns, basically depends on the situation. Particularly, we assume a new dialogue to begin as soon as the user starts addressing the system with whatever intention. Then we assume that dialogue to be finished as soon as the system manages to satisfy every goal that may have been positively identified (and hopefully requested by the user during the dialogue) or just as soon as the user decides to abandon it (e.g. by using a "cancellation" voice command).

Fig. 1 shows a block diagram of our conversational interface. The system consists of an automatic speech recognition module (ASR), which translates the audio signal into a text hypothesis of what the user has said; a language understanding module (NLU), that extracts the semantics of the user's utterance; the dialogue manager (DM), which makes use of the extracted semantic information, together with the information available at the context manager module, to determine the actions on the system that the user wants to fulfil, and to provide the user with feedback regarding the current dialogue turn; the context manager (CM), which holds the information of not only the ongoing dialogue but also of the past ones between the same user and the system;

[1] EDECAN Project Web page: http://www.edecan.es