Neurocomputing 163 (2015) 56–68

Contents lists available at ScienceDirect

Neurocomputing

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

# A proposal for the development of adaptive spoken interfaces to access the Web

David Griol [a,*], José Manuel Molina [a], Zoraida Callejas [b]

[a] Applied Artificial Intelligence Group, Department of Computer Science, Carlos III University of Madrid, Spain
[b] Spoken and Multimodal Dialogue Systems Group, Department of Languages and Computer Systems, University of Granada, Spain

ARTICLE INFO

*Article history:*
Received 29 October 2013
Received in revised form
16 July 2014
Accepted 23 September 2014
Available online 17 April 2015

*Keywords:*
Dialog systems
Spoken interaction
User modeling
Adaptation
Neural networks
Statistical methodologies

ABSTRACT

Spoken dialog systems have been proposed as a solution to facilitate a more natural human–machine interaction. In this paper, we propose a framework to model the user's intention during the dialog and adapt the dialog model dynamically to the user needs and preferences, thus developing more efficient, adapted, and usable spoken dialog systems. Our framework employs statistical models based on neural networks that take into account the history of the dialog up to the current dialog state in order to predict the user's intention and the next system response. We describe our proposal and detail its application in the Let's Go spoken dialog system.



## 1. Introduction

Continuous advances in the development of information technologies and the miniaturization of devices have made it possible to access information, web services, and artificial intelligence systems from anywhere, at anytime and almost instantaneously through wireless connections [1]. Although devices such as smartphones and tablets are widely used today to access the web, the reduced size of the screen and keyboards makes the use of traditional graphical user interfaces (GUIs) difficult, especially for motor-handicapped and visually impaired users. This way, although mobile phones are designed to provide ubiquitous access to Internet, the present challenge is to make the enormous web content accessible to all mobile phone users by means of more natural communication metaphors.

Dialog systems go a step beyond GUIs by adding the possibility to communicate with these devices through other interaction modes such as speech [2–5]. These systems can be defined as computer programs designed to emulate human communication capabilities including several communication modalities. To successfully manage the interaction with the users, spoken dialog systems usually carry out five main tasks: automatic speech recognition (ASR), spoken language understanding (SLU), dialog management (DM), natural language generation (NLG) and text-to-speech synthesis (TTS).

The goal of speech recognition is to obtain the sequence of words uttered by a speaker [6]. Once the speech recognizer has provided an output, the system must understand what the user said. The goal of spoken language understanding is to obtain the semantics from the recognized sentence [7]. This process generally requires morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge.

The dialog manager decides the next action of the system, interpreting the incoming semantic representation of the user input in the context of the dialog [8]. In addition, it resolves ellipsis and anaphora, evaluates the relevance and completeness of user requests, identifies and recovers from recognition and understanding errors, retrieves information from data repositories, and decides about the next system's response. In order to complete these tasks and decide "what to say", the dialog manager needs to track the dialog history and update some representation of the current state of the dialog. In addition, the DM needs a dialog model that defines the conversational behavior of the system, for example when to take the initiative in a dialog, when to confirm a piece of information, how to identify and recover from recognition and understanding errors, and so forth.

Natural language generation is the process of obtaining sentences in natural language from the non-linguistic, internal representation of information handled by the dialog system [9]. Finally, the TTS module transforms the generated sentences into synthesized speech [10].

During the last years, Internet is playing an increasingly important role for making speech technology available anywhere.

* Corresponding author.
*E-mail addresses:* david.griol@uc3m.es (D. Griol), josemanuel.molina@uc3m.es (J.M. Molina), zoraida@ugr.es (Z. Callejas).

http://dx.doi.org/10.1016/j.neucom.2014.09.087
0925-2312/

By giving the user the chance to interact with the web via natural language, users are provided with the possibility to come up with less restricted input. Initial application domains included simple solutions to provide a vocal interface to an existing web browser [11] or to access information in limited on-line domains [12].

The performance of spoken language dialog systems to access web contents has improved over time, extending these initial application domains to more complex information retrieval and question answering applications [13,14], e-commerce systems [15], surveys applications [16], recommendations systems [17], e-learning and tutoring systems [18,19], in-car systems [20,21], remote control of devices and robots in smart environments [22,23], healthcare and Ambient Assisted Living systems [24,25], or embodied dialog systems and companions [26,27].

Many companies and public institutions have also taken advantage of using Voice Portals as a cheap and effective way of supporting customers [28]. Besides spoken and written language have become popular with the incorporation of chatbots for web-based customer support [29]. The spread of information through social web media has also made possible to generate models for conversation that take profit of the data's vast size and conversational nature of web applications like Twitter or Wikipedia [30,31], and also allow spoken interaction in 3-D immersive virtual environments like Second Life or Open Simulator [32,33]. Mobile devices have also extended the possibilities of integrating speech interaction to develop advanced apps that access the web [2].

The described systems are usually designed ad hoc for their specific domain using rule-based models and standards in which developers must specify the steps to be followed by the system. This way, the adaptation of the hand-crafted designed systems to consider specific user requirements or deal with new tasks is a time-consuming process that implies a considerable effort, with the ever-increasing problem of dialog complexity [34,35].

In addition, although much work emphasizes the importance of taking into account user's models not only to solve the tasks presented to the dialog system by the user, but also to enhance the system performance in the communication task, this information is not usually considered when designing the dialog model for the system [36,37]. For this reason, in most dialog applications, the dialog specification is the same for all cases: users typically have no control over the content or presentation of the service provided.

Incorporating intelligence into a spoken language based communication system requires, among other things, careful user modeling in conjunction with an effective dialog management. With the aim of creating dynamic and adapted dialogs, the application of statistical approaches to user modeling and dialog management makes it possible to consider a wider space of dialog strategies in comparison to engineered rules [38,8].

The main reason is that statistical approaches for dialog management can be trained from real dialogs, modeling the variability in user behaviors. Although the parameterization of the model depends on expert knowledge of the task, the final objective is to develop dialog systems that have a more robust behavior, better portability, and are easier to adapt to different user profiles or tasks [39]. This would help to create user adapted speech-enabled interfaces for the wide-range of web-based applications previously described, reducing the effort and time required by hand-crafted designed systems to consider specific users requirements or deal with new tasks with the ever-increasing problem of dialog complexity [34,35].

In this paper we describe a framework to develop user-adapted spoken dialog systems. Our proposal is based on the definition of a statistical methodology for user modeling that estimates the user intention during the dialog. The term user intention expresses the information that the user has to convey to the system to achieve their goals, such as extracting some particular information from the system. It is a very useful and compact representation of human–computer interaction that specifies the next steps to be carried out by the user as a counterpart in the human–machine conversation.

This prediction, carried out for each user turn in the dialog, makes it possible to adapt the system dynamically to the user's needs. To do this, a statistical dialog model based on neural networks is generated taking into account the predicted user's intention and the history of the dialog up to the current moment. The next system response is selected by means of this model. The codification of the information and the definition of a data structure which takes into account the data supplied by the user throughout the dialog make the estimation of the dialog model from the training data and practical domains manageable.

The remainder of the paper is organized as follows. In Section 2 we describe the motivation of our proposal and review main approaches focused on key aspects related to it, such as user modeling techniques when interacting with dialog systems and the application of statistical methodologies for dialog management. Section 3 presents in detail our proposal to develop adaptive dialog systems. Section 4 describes the application of our proposal in the CMU Let's Go spoken dialog system, a system that has been used during the last years by the dialog systems community as a common ground for comparison and verifiable assessment of the improvements achieved. In this section we also discuss the evaluation results obtained in this application. Finally, in Section 5 we present the conclusions and outline guidelines for future work.

## 2. Related work

The design and development of a comprehensive adaptive spoken dialog system can be conceptually composed of two interconnected components; the user modeling, and the corresponding adaptation that in our proposal is implemented on the dialog manager.

Research in techniques for user modeling has a long history within the fields of language processing and dialog systems. A thorough literature review on the application of how data mining techniques to user modeling for system personalization can be found in [39–41]. It is possible to classify the different approaches with regard to the level of abstraction at which they model dialog. This can be at either the acoustic level, the word level or the intention-level. The latter is a particularly useful representation of human–computer interaction [39].

Intentions cannot be observed, but they can be described using the speech-act and dialog-act theories [42,43]. The notion of a dialog act plays a key role in studies of dialog, in particular in the interpretation of the communicative behavior of the participants; in building annotated dialog corpora; and in the design of dialog management systems for spoken human–computer dialog. A dialog act has two main components: a communicative function and a semantic content. A standard representation for dialog act annotation is proposed in [44], which uniformizes the semantic annotation of dialog corpora. Thus, it provides a standard representation for the output provided by the SLU module in dialog systems and its communication with the dialog manager (e.g., *Yes-No-Question*, *Reject*, *Conventional-Closing*, or *Thanks*).

In recent years, simulation on the intention-level has been most popular [39]. This approach was first used by [45] and has been adopted for user simulation by most research groups [46–49]. Modeling interaction on the intention-level avoids the need to reproduce the enormous variety of human languages on the level of speech signals [50,51] or word sequences [52,53].