The Wizard of Oz meets multimodal GUI interfaces: new challenges

LISOWSKA, Agnès, et al.

Abstract

As part of our work in the Interactive Multimodal Information Management (IM2) project [6], we developed a multimodal dialogue-driven interface for browsing and searching recorded and annotated meeting data in a multimedia database. The challenge lay in creating an interface that was easy to use and which smoothly blended direct manipulation and natural language interaction (both voice and keyboard based).

Reference

The Wizard of Oz meets multimodal GUI interfaces: new challenges

Introduction
As part of our work in the Interactive Multimodal Information Management (IM2) project [6], we developed a multimodal dialogue-driven interface for browsing and searching recorded and annotated meeting data in a multimedia database. The challenge lay in creating an interface that was easy to use and which smoothly blended direct manipulation and natural language interaction (both voice and keyboard based).

Investigative Goals
In order to meet this challenge we had to consider three interrelated research goals. The first was to determine which input modalities to choose - among voice, pen, mouse and keyboard - and in which combinations they would be the most useful for an interface in this domain. The other two goals relate directly to the choice of including natural language interaction. One was to determine the linguistic coverage of the natural language processing modules. The other was to determine the optimal dialogue strategies for smooth interaction between the system and the user. While much research has gone into examining these latter two points, most involves the use of natural language and dialogues in voice-only systems, rather than in a multimodal context. Existing
multimodal systems tend to stay in limited domains that deal with spatio-temporal tasks rather than broad domain browsing and retrieval applications.

**Why traditional methods didn’t apply**
Traditionally, data for the development of voice-only natural language dialogue systems is gathered using Wizard of Oz (WOz) studies [3, 5]. The user interacts with a system that they believe to be fully functional, but which in fact is only partially implemented. The parts of the system that have not been implemented - usually the language processing components - are replaced by a 'wizard', a human being who sits in another location and simulates the missing functionalities. We found that to meet our needs, the traditional WOz environment was insufficient.

In order to determine which modalities were most convenient for certain tasks, we had to design a flexible multimodal system. This means that users can access every functionality using any modality. Consequently, the wizard needed to be able to control not only the vocal interaction with the system, but also to ensure that the graphical interface is synchronized with the vocal dialogue context and vice versa. This level of complexity is often absent in traditional WOz experiments.

**Linking graphics and dialogue**
Due to the need for efficient synchronization of the user’s graphical interface with the state of the vocal interaction, the wizard’s interfaces had to be carefully incorporated into the system. This implied that the wizarding environment itself became a set of system components. The components were designed in a highly modular way to ensure that wizard components could be replaced by automatic components at later times.

We based our work on the Rapid Dialogue Prototyping Methodology (RDPM) [11] developed by some of the authors of this paper, which integrates WOz directly into the design of the system. The RDPM was initially developed for voice-only systems and was extended to cover multimodal situations for work with our interface [2, 4]. It decomposes a complex dialogue model into a set of dialogue sub-models which control a set of simpler vocal interactions with the user.

Since the dialogue model and the graphical interface needed to be tightly coupled, every dialogue sub-model was linked to a corresponding graphical component and vice-versa. Proper linking assures that every possible user action on any of the elements of the graphical component (e.g. clicking on a button, selecting a value from a list) corresponds to a user action in the linked vocal dialogue sub-model (e.g. saying "Do <this>/Go <there>" or "Select <value from list>/It is <value from list>"). Every user manipulation of a graphical component thus translates into an action in the linked vocal dialogue sub-model and can therefore be supervised or responded to by the wizard on his interface.

Finally, we had to ensure that the necessary data was logged in the appropriate way in order to gather the types of data that we needed to meet our research goals of studying the use of modalities, natural language coverage and dialogue strategies.
Implications for WOz environment design

Pilot experiments with early versions of the system and WOz environments showed that the cognitive load on the wizard had increased significantly. They not only had to keep track of and process natural language input through both voice and keyboard, but also oversee the input from the graphical user interface (mouse and pen driven selections). This resulted in the need for multiple wizards who had to work in an efficient and coordinated manner to quickly and correctly process user input and provide the appropriate output.

The first wizard, whose role was similar to that of a wizard in traditional WOz interfaces, interpreted the user's multimodal input. The wizard had to monitor input from all modalities, and in cases other than the sole use of direct manipulation, the wizard had to produce the correct set of actions in response to that input in "real" time. Difficulties were encountered when performing tasks such as quickly finding sought-after information in the database, or replicating point-and-click functionalities in the user's interface that were provided orally. After several runs of pilot experiments, these difficulties were overcome by adding shortcut buttons on the wizard's interface to replicate common actions to manipulate the user's interface, as well as a quick-search mechanism for finding items in the database. This last function also helped the wizard disambiguate uncertain situations in the sense that it would only display items that were possible in the given context.

Moreover, the structure of the wizard's interface permitted fewer 'human-like' interpretations of the user's input, since it limited the types of actions that they could perform. This was particularly useful given the amount of information that the wizard had to process in a short amount of time. Part of the overhead of thinking about what the system would or would not be capable of processing was thus eliminated.

The role of the second wizard was to control the vocal output given by the system – i.e. the natural language prompts. This was necessary in cases where the default prompt suggested by the existing dialogue management module was insufficient or inappropriate given the current dialogue situation. The wizard would select a new prompt from a predefined list of prompts, or create a new one on the fly, which was then automatically added to the list. This ensured that the same set of prompts was available during the experiments and thus introduced less unwanted variety in what the system could 'answer'. Since this role was not originally included in the traditional WOz, we had to develop and incorporate it into the methodology ourselves.

These two interfaces were connected to the system in such a way that the interaction appeared smooth and coordinated from the user's perspective. In most cases however, several runs of pilot experiments were required before the optimal layout and functionalities were found for each of the wizard interfaces.

Finally, it is important to note that multiple wizards, combined with a graphical interface for the user also implies a more sophisticated hardware setup. In order to do their job, it is important for the wizards to have a clear view of what is going on in the user's room, how they are reacting to the system and what is happening
on their screen, in both video and audio form, in real time.

**Assessment of the methodology**

In our experience, the new Wizard of Oz methodology worked quite well and allowed us to gather the types of information we needed to meet our research goals. Moreover, it facilitated a smooth and interlinked design and evaluation process since parts of the system (such as the graphic interface) could be evaluated by real users even while the linguistic components were under development.

Finally, there are several important elements that must be taken into account when using this methodology.

- The wizarding components must be linked into the system under development from the design stage; otherwise the wizards will not have a sufficient degree of control over the system and interactions with the user.

- A significant amount of time and effort is required to develop the wizard's interfaces to both assure consistency in user responses and to refine issues of ergonomy, system control and rapid database access.

- Several runs of pilot experiments may be necessary before the wizard’s interfaces are sufficiently refined to allow the wizards to quickly and efficiently play their roles.

**Acknowledgements**

The work presented here is part of the Swiss NCCR on “Interactive Multimodal Information Management” (IM2, http://www.im2.ch), funded by the Swiss National Science Foundation.

**References**


