Managing Consistency in Wizard of Oz Studies: A Challenge of Prototyping Natural Language Interactions

Stephan Schlögl
Institut Mines-Télécom
Télécom ParisTech, LTCI
Paris, France
schlogl@enst.fr

Gavin Doherty
Trinity College
University of Dublin
Dublin, Ireland
Gavin.Doherty@scss.tcd.ie

Saturnino Luz
Trinity College
University of Dublin
Dublin, Ireland
Saturnino.Luz@scss.tcd.ie

ABSTRACT

Wizard of OZ (WOZ) is a prototyping method that uses a human 'wizard' to mimic the functions of a prospective system. Although the WOZ technique is widely used in HCI design, surprisingly little research has been done on how wizard performance and consistency may effect the findings of a WOZ trial. In this paper we present a meta-analysis of three WOZ studies that shows how wizard behavior can vary significantly, potentially influencing the user experience of test-participants. In text-based interactions, for example, wizards may be required to send several consecutive utterances before receiving any feedback. In order to deliver a consistent experience they then need to give test participants enough time for reading and processing an utterance before sending the next utterance. In this study we examined this aspect in some more detail and found that these time-windows can vary significantly within as well as between studies. By contrast, a similar experimental setup simulating speech-based system output did not suffer from this type of inconsistency. Our analysis highlights one of the challenges involved in using the WOZ method for designing novel interactive systems (both stationary and mobile), and suggests that additional support for wizards is needed in order to improve consistency.

Author Keywords

Wizard of Oz; Prototyping; Language Technology

ACM Classification Keywords

H.5.2 User Interfaces: Prototyping

General Terms

Human Factors, Experimentation, Measurement

INTRODUCTION

Wizard of OZ (WOZ) is a well-established method for prototyping future products. Using a human 'wizard' to mimic the functionality of a system, either completely or in part, is particularly useful in the development of speech and language enabled applications, where the performance of components may not be sufficient to enable a reasonable user experience without extensive engineering effort. It helps designers to produce appropriate dialog models as well as to improve their understanding of a domain by allowing them to evaluate potential user experiences and interaction strategies without the need for building a fully functional product first. In general the technical requirements for running a WOZ experiment are relatively low, however, Salber & Coutaz [8] point out that the task of the human wizard is highly demanding. Aspects that seem particularly difficult to accomplish include fast response times (i.e. a wizard needs to find an appropriate response among a set of pre-defined choices, generate a suitable response on the fly, or initiate an error routine), a realistic simulation of system behavior (i.e. both too good as well as deficient simulation may lead to bias), and consistency (i.e. inconsistent wizard behavior might influence user responses). In order to build tools that optimally support the WOZ experimentation, it is therefore important to understand the different problems and constraints faced by those who perform the task of the wizard. The research presented in this paper aims to contribute to this understanding by examining a particular aspect of the wizard's task, namely his/her ability to provide consistent timing when sending utterances to a test-participant. This aspect of the wizard task was observed in three WOZ studies that simulated text- as well as speech-based interaction with an intelligent system. Conducting a meta-analysis of those three studies we found that consistency in text-based settings is difficult to achieve. Speech output, however, leads to significantly better wizard performance.

BACKGROUND

WOZ has its roots in the area of Natural Language Processing (NLP) where it was first used by Erdman & Neal [3] to test their concept of a self-service airline ticket kiosk. Later Gould et al. [4] employed WOZ to explore the possibilities of the 'Listening Typewriter' and De Marconnay et al. [7] extended its application area from testing pure speech-based interaction to evaluating gestures and face recognition. This expansion in scope continued with Salber & Coutaz [8] who looked at multi-modal interaction, leading to the introduction of multiple wizards. In more recent years WOZ experiments have been used for a variety of purposes, including the prototyping of a speech-based city guide [5], the testing of a pedestrian navigation system [6] and the simulation of a location-based mobile game [1]. While recent employments increasingly explore novel interaction paradigms (e.g. location-based services) most of them also integrate some sort

PID-MAD 2013, Aug 27 2013, Munich, Germany.
(In Conjunction with MobileHCI '13, Aug 27-30 2013, Munich, Germany.)

of natural language, and this is usually the one aspect which needs the biggest amount of simulation, as existing technology is simply not mature enough to be used without significant upfront investment. Hence, we argue that by improving the support for language-based interaction we can foster the WOZ method as a whole and significantly increase its robustness as an evaluation instrument.

From a methodological point of view WOZ tries to mimic the functionality and performance of a computer system. This should happen in a way such that potential test-participants believe they are interacting with a piece of technology rather than a human. The goal of the wizard must therefore be to imitate the functions of the system as convincing as possible. While it seems obvious that human control might not be able to completely reproduce all the aspects of technology, especially in terms of speed and accuracy, maintaining a level of consistency is important for the effectiveness of the method, in particular when it is used in a controlled experimental setting. Inconsistency may be seen as a confounding variable and therefore lead to a bias in the quantitative (e.g. completion time) as well as the qualitative (e.g. user satisfaction) data that is collected.

One major element of consistency, when working with text-based system output, is to give participants time to read output. That is, when sending consecutive utterances, it is up to the wizard to decide when a participant should receive the next one. From a methodological point of view it is not necessarily important to provide an accurate estimation of a participant's reading speed, as this could also be seen as a rather sophisticated and obviously user dependent feature of a simulated system. However, a convincing simulation of a system may very well require a certain level of consistency - one that is associated with the length of a sent text fragment, giving more reading time for longer utterances and less time for shorter ones. Focusing on this aspect of giving test-participants time to read sent text-utterances the following sections describe an analysis of three different WOZ studies in which three different wizards interacted with a number of different test-participants. All three studies pursuit distinct research goals and therefore can be seen as realistic and valid examples for exploring methodological aspects of WOZ.

METHODOLOGY

Focusing on a wizard's consistency in giving test-participants time to read a sent utterance (i.e. a simulated system response), the log-files of three different WOZ studies were analyzed. All three studies employed a similar set-up. The first (herein after referred to as Study A) as well as the second study (herein after referred to as Study B) were looking at the effect of using Machine Translation (MT) in an interactive information retrieval scenario, and the third study (herein after referred to as Study C) was collecting a corpus of dialog utterances for designing an online pronunciation trainer. Even though all three studies followed their own study design and methodology, they were very similar from a wizard's point of view. They simulated a language-based (i.e. speech input and text or synthesized speech output) interaction between a human and a system, where the main task of the wizard was to select a suitable response from a set of pre-defined utterances. All studies were conducted in the same research lab and they all used non-native English speaking test participants. Their only difference was the application domain and the wizard that was used to simulate the system. In terms of tool support all three studies used the same WOZ prototyping platform, albeit different releases. The platform has been developed as part of a research project focusing on generic wizard support. It is based on modern web technologies and offers browser-based interfaces for both the wizard and test participants, allowing for stationary and mobile experimentation.

Study A

The first study simulated a system that understands spoken input in German and produces text-based answers. The scenario was situated in the sales domain, where the system should help potential customers choose a suitable product. In the simulated case the product was an Internet connection bundle. Possible system utterances for this sort of customer-machine interaction were defined and translated into German, using two different MT systems. A WOZ prototype for the scenario was implemented and tested, and a member of our research team was chosen to act as a wizard. The wizard first conducted three trial runs before eight German-speaking test-customers were asked to solve two different tasks with the system (i.e. 11 experiments in total). Further information and results for this study can be found in [9].

Study B

Study B was building upon the results of Study A and aimed at extending them into the spoken language domain. The set-up was similar to the one used in Study A with the difference that for one of the two tasks the system generated spoken output. On the test-participant side the input modality for both tasks remained speech. In order to simulate spoken system output, utterances were pre-recorded using a Text-to-Speech system and linked to the according text utterances already stored in the system. In addition the wizard interface for the study was significantly changed. While Study A was based on an early prototype, the interface for Study B was modified to better support the task of the wizard. For conducting the study a different researcher was chosen to act as a wizard. After one trial run 16 German test-customers were recruited to interact with the system (i.e. 17 experiments in total).

Study C

The final study employed WOZ to collect a corpus of realistic dialog utterances for an online language pronunciation trainer. For this purpose we were working with an external research institution. The team there had built a working prototype of a system that could analyze a test-participant's pronunciation of an English sentence and highlight which words or parts of a sentence were mispronounced. Linking this analysis to actual textual feedback was, however, not supported at the time. The study therefore used a human wizard to produce real-time textual feedback based on the results of the pronunciation analysis. Again a slightly improved version of the WOZ system was used to implement the study. Different text elements were prepared so that they could be assembled

to flexibly form a feedback sentence. The wizard was able to compose the sentence and fill in the specifics or alternatively create a response completely from scratch. Feedback sent from the wizard was displayed in a text box situated in the bottom of the test-participant's screen. A member of the external research team acted as a wizard. One trial run was conducted to test the set-up after which 12 test-participants were recruited to train their pronunciation (i.e. 13 experiments in total). Additional details and results of this study can be found in [2].

Meta-Analysis

Looking at the above described studies from a meta-level we were interested in comparing wizard response times. More precisely we were looking at the consistency with which a wizard gives a test-participant time to read a sent utterance before sending the next one. This is an issue that is very specific to the text-based output of those three studies but related to the rather stressful task of the wizard. Generally, a wizard first listens and interprets a test-participant's spoken input, and then searches for an appropriate response utterance to reply (in multi-modal settings several input and output channels might be simulated). In a sense this task can be seen as an information retrieval problem under time pressure in which a majority of unsuitable responses act as distractors for finding the one suitable response. To tackle this problem utterances are often limited to one or two sentences. Short utterances are easier to distinguish from each other and at the same time allow for a higher degree of flexibility. Thus, a wizard can 'feed' them to a test-participant bit by bit and they can be reused for different responses. The consequence is, however, that a wizard might need to send several utterances in a row without receiving any feedback in between. Here it is up to the wizard to estimate the time a test-participant needs to read one utterance so as to decide when to send the next one.

Analyzing the log files of Study A (11 experiments) we were able to identify 156 instances in which the wizard had to send several utterances without waiting for participant input in between them. In Study B (17 experiments) there were 117 instances (i.e. looking solely at text-based interaction) and Study C (13 experiments) contained 218. Next we took the time a wizard waited before sending a follow-up utterance and divided it by the number of words that were used in the preceding utterance (Note: We use the number of words in a sentence as an approximation for the time that is required to read and process it. We are, however, aware that the number of syllables as well as the overall complexity of used words may be a better scale. Nevertheless, it can be argued that for the setting of a WOZ experiment a consistent time/word ratio is sufficient to simulate realistic system behavior). In an optimal setting this time/word ratio would be consistent during the course of one study trial and furthermore stable throughout the whole study. Taking the Interquartile Range (IQR) of these ratios for each trial provides a measurement for in-trial consistency. If the IQR is zero it can be assumed that the wizard acted consistently throughout the trial, any number above zero constitutes variability in the time that was give to a test-participant to read utterances. Figure 1 shows the IQR values for all three WOZ studies and their changes over time.

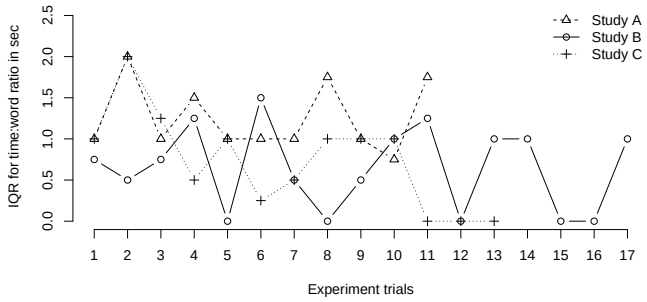

Figure 1. Interquartile Range (IQR) values for text-based interaction.

At this point it needs to be highlighted that the wizards in those three studies were not actively asked to estimate a participant's reading speed. Nor were these studies specifically designed for exploring this aspect of wizard behavior. We simply report on a meta-analysis conducted in order to better understand the challenges that were involved. The results of this analysis show that a wizard's natural actions (i.e. their decisions on when to send an utterance) vary significantly between experiments, and also within an experiment, wizards were inconsistent in their estimation of when a test-participant would have finished reading one utterance and therefore could be confronted with the next one. In Study A the time/word ratio varied in each experiment between 0.5 and 2 seconds. In Study B the wizard was consistent in 5 experiments out of 17 (i.e. IQR = 0.00) and in Study C only the last 3 experiments showed no variation. One could infer that, since consistent behavior on this matter was not explicitly required from the wizards, those results simply reflect a faulty experimental design. Yet, one of the wizards categorically highlighted that she was reading a sent utterance slowly in her mind before sending the next one, which shows a certain awareness of the problem. Furthermore, it could be argued that consistent behavior generally depends on experience, for which a wizard needs to receive sufficient training before being able to run a study with real participants. However, an interview study with 20 researchers from academia and industry, all of whom had experience with WOZ, highlighted that the time spent on wizard training is often less than 30 minutes. Furthermore, our data shows that even with appropriate wizard experience, the consistent timing of utterances remains a challenge. The wizard in Study C seemed to benefit from the experience gained over the course of 13 trials, which led to consistent behavior at the end. The wizards in Studies A and B, however, did not show significant improvements over time. Hence it can be argued that additional wizard support on this matter might be needed, even in cases where the wizard has received an appropriate level of training.

To further reflect on this aspect we looked at speech-based interaction. As already pointed out earlier, in Study B one of the two tasks a test-participant had to complete was using synthesized output. The wizard could hear the output and was therefore able to send the follow-up utterance as soon as the

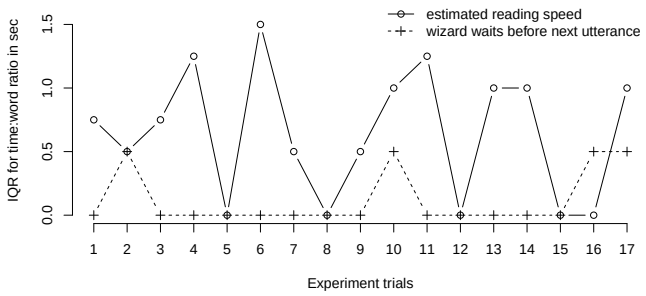

Figure 2. Results of Study B that shows a wizard's differences in consistency (IQR value) when dealing with speech-based utterances (dotted line) compared to dealing with text-based ones (continuous line).

previous recording was finished (as opposed to estimating the time a participant needed to read it). We again calculated the IQR in order to contrast it with the text-based values. Figure 2 shows the IQR values for Study B's speech-based output compared to its text-based output. As can be seen from the plot, the time/word consistency for the speech-based interaction was significantly better than it was in its text-based counterpart (one-tailed paired Student's t-test: $t(16)= 3.9105$, $p=0.0006$). In fact, in 13 out of 17 trials a wizard's response time in relation to the preceding utterance can be seen as entirely consistent. The remaining 4 trials show a variation of only 0.5 seconds, which might be explained by additional time needed to search for a follow-up utterance.

DISCUSSION AND CONCLUSION

WOZ is a valuable prototyping method for designing interactive technology. Especially for systems that involve some sort of NLP the kind of feedback that can be gained helps to shape design. Yet, the dependency on a human wizard makes the method susceptible for errors. Inconsistent wizard performance, as described above, can influence participant behavior and consequently bias results gained from a study. In the here presented analysis it was shown that additional acoustical information provided by an audio (or possibly video) channel clearly improves a wizard's performance, and we would therefore recommend using it by default. In the case of an experiment that aims at simulating 'intelligent' text-based system output (e.g. one could think about prototyping location-based information system) such acoustic feedback may also be provided only to the wizard. Using Text-to-Speech technology to read out sent utterances, for example, could help achieve better consistency. However, one also needs to be aware that such might lead to overly 'intelligent' system behavior in which case alternative, more specific solutions may sometimes be better (e.g. a timer function). On the other hand for the presented studies one could also argue that letting a wizard first collect all the relevant utterances before sending them on would eliminate the problem of estimating timing altogether. Yet, this would likely result in an increased overall response time, potentially influencing experiment results. In addition, a lot of experimental settings, particularly in the mobile domain, need to deal with limited screen space where sending a collection of several utterances is often not suitable. Finally, in case a user is not responding after a certain amount of time, a system (i.e. the wizard) might need to send a check-up utterance, for which again a realistic time-window needs to be estimated. In summary one can therefore say that while wizard support on the one hand needs to be flexible, in order to allow for a variety of different settings, it also needs to help control for as many confounding variables as possible; consistent response timing being one of them. A useful combination of flexibility and control can thus be seen as the key challenges for designing better WOZ prototyping tools whereupon more studies of wizards at work, such as the one presented here, may help in achieving the right balance.

ACKNOWLEDGMENT

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin, and by ARHOME, a French national research project.

REFERENCES

1. Bernhaupt, R., Jenisch, S., Keyser, Y., and Will, M. Capture The Flag: Simulating a Location-Based Mobile Game Using the Wizard-Of-Oz Method. In *Proceedings of ACE* (2007), 228–229.
2. Cabral, J. P., Kane, M., Ahmed, Z., Abou-Zleikha, M., Szekely, E., Zahra, A., Ogbureke, K., Cahill, P., Carson-Berndsen, J., and Schlögl, S. Rapidly testing the interaction model of a pronunciation training system via wizard-of-oz. In *Proceedings of LREC* (2012).
3. Erdmann, R. L., and Neal, A. S. Laboratory vs. field experimentation in human factors: An evaluation of an experimental self-service airline ticket vendor. *Human Factors 13* (1971), 521–531.
4. Gould, J. D., Conti, J., and Hovanyecz, T. Composing letters with a simulated listening typewriter. *Communications of the ACM 26* (1983), 295–308.
5. Howell, M., Love, S., and Turner, M. The impact of interface metaphor and context of use on the usability of a speech-based mobile city guide service. *Behaviour & Information Technology 24*, 1 (2005), 67–78.
6. Krüger, A., Aslan, I., and Zimmer, H. The Effects of Mobile Pedestrian Navigation Systems on the Concurrent Acquisition of Route and Survey Knowledge. In *Proceedings of MobileHCI* (2004), 446–450.
7. Patrice, D. M., James, L. C., and Daniel, S. Visual interpretation of faces in the NEIMO multi-modal test-bed. In *Proceedings of IJCAI. Workshop Looking at People: Recognition and Interpretation of Human Action* (1993).
8. Salber, D., and Coutaz, J. A wizard of Oz platform for the study of multimodal systems. In *Proceedings of INTERACT and CHI* (1993), 95–96.
9. Schneider, A., Van der Sluis, I., and Luz, S. Comparing intrinsic and extrinsic evaluation of mt output in a dialogue system. In *Proceedings of IWSLT* (2010), 329–336.