Wizard of Oz: Protocols and Challenges in Studying Searchbots to Support Collaborative Search

Sandeep Avula
University of North Carolina at Chapel Hill
asandeep@unc.edu

ABSTRACT

The Wizard of Oz (WoZ) methodology has gained prominence with researchers using it to understand how users perceive and interact with conversational agents. They favor this methodology to expedite development time and reduce noise due to system performance, such as unreliable automatic speech recognition (ASR). In this paper, we describe the WoZ methodology we used in our work on searchbots (i.e., chatbots that perform specific types of searches) during collaborative information-seeking tasks. We describe our protocols and challenges.

Keywords: Wizard of Oz; Chatbots; Collaborative Search; Intelligent Agents

1 INTRODUCTION

Conversational agents which support users in their everyday tasks are increasing by the day [8]. Their use cases extend from physical spaces, such as kitchens [11], to virtual work spaces [2, 7]. Despite their wide usage, the functionality of these systems is limited and interaction with them is still not *natural* [8]. The goal of seamlessly integrating these technologies into our everyday lives has spurred an active interest in the IR and HCI communities.

A common trend among recent studies in IR and HCI on conversational sytems is the use of the Wizard of Oz (WoZ) methodology [7, 10, 11]. WoZ is a methodology in which a human stands in place of a machine to replace some of its functionalities such as speech recognition, dialog management, etc. This is done to pursue research which does not evaluate how well the technology works but rather its applicability, perception, and interaction by users. This methodology allows researchers to subdue errors from imperfect systems, such as automatic speech recognition (ASR), while also encouraging an iterative design process [3, 10].

In this paper, we describe the WoZ methodology we used in our study on searchbots (i.e., chatbots that perform specific types of searches) [2]. We studied searchbots that intervene dynamically and compared between two intervention types: (1) the searchbot presents questions to users to gather the information it needs to produce results, and (2) the searchbot monitors the conversation among the collaborators, "infers" the necessary information, and then displays search results with no additional input from the users. In this paper, we describe our protocols and challenges in implementing this methodology. We believe that insights from our methodology can support future research on conversational agents.

2 WIZARD OF OZ

The goal of a Wizard of Oz (WoZ) methodology is to mimic a hypothetical system which is not yet in existence or one which is bounded by technical limitations such as imperfect automatic speech recognition (ASR) [6, 10]. The name comes from the popular movie/book of the same, in which a Wizard from the outset behaves as though he is grand and intimidating, but in reality is a normal middle-aged man, who hides behind a curtain. In our study, we use the WoZ methodology to imitate a conversational agent which can dynamically intervene to assist users in an information-seeking task. We adopt this methodology as our objective is to understand how users perceive and utilize the search tools, and not how well the system learns from the conversation.

Recent studies [10, 11] have used this methodology to study how users perceive and interact with conversational agents. Vtyurina and Fourney [11] used this methodology to simulate a conversational agent in guided task completion scenarios, which are scenarios in which tasks have clear and sequential subtasks such as a cooking recipe. In their methodology, the Wizard had access to a list of responses, which depending on the users' request got played back in a computer-synthesized voice. The focus of the study was to understand how and when users provide implicit conversational cues as feedback. Implicit cues are utterances such as *"Okie-doke"* and *"alright"* which convey agreement or interest without explicitly stating the entire intention in an utterance. Current conversational systems are not attuned to integrate such feedback and the WoZ methodology is appropriate to understand this interaction space. Shamekhi et al. [10] used the WoZ methodology to study how embodied conversational agents can support groups. They compared two conversational agent conditions: embodied vs. voice-only. The Wizard played the role of the agent in both conditions. The Wizard in this study was given access to a control panel with specific commands. The Wizard would listen to the conversation between the participants and based on the intent (of the conversation) would select an option on the control panel which would send a response to the group. By making the Wizard play the role of the conversational agent, the researchers were able to avoid noise caused by ASR systems and dialog management.

3 STUDY DESIGN

In this paper, we describe the Wizard of Oz methodology from our work on searchbots [2]. The goal of the study was to investigate the possibility of integrating collaborative search tools into existing communication channels (e.g., Slack, which we used as our communication platform) rather than building stand-alone collaborative search tools. Specifically, we studied searchbots that intervened dynamically and compared between two intervention types: (1) the searchbot presented questions to users to gather the information it needed to produce results, and (2) the searchbot monitored the conversation among the collaborators, inferred the necessary information, and then displayed search results with no additional

CAIR '18, July 12, 2018, Ann Arbor, MI, USA
2018. https://doi.org/xxxxxx

input from the users. In this section, we describe the Wizard of Oz methodology to implement the two intervention types.

To investigate our research questions we conducted a laboratory study and recruited 27 pairs of participants (34 female and 20 male) who had previously collaborated together. Each pair of participants worked on three different search tasks. During each task, the participant pair had to make three selections (e.g., three restaurants to consider for a night out). Each task was accompanied with a backstory (to contextualize the task) and each participant was given constraints that were unknown to their partner.

3.1 Tasks and Searchbot Conditions

Tasks: Participants completed three search tasks: (1) a restaurant-finding task, (2) a local attractions-finding task, and (3) a book-finding task. Each task had a "background story" and asked the participants to search for and agree on three different items. Participants were given gender-neutral first names (Jamie and Taylor). The personal preferences associated with each task were: restaurants task–location, food preference; local attractions task–location, attraction type; books task–fiction vs. non-fiction, sub-genre.

Searchbot Conditions: We gave participants three tasks and also custom-designed three different searchbots to match the tasks: (1) a searchbot for local restaurants, (2) a searchbot for local attractions, and (3) a searchbot for books. Each searchbot was designed to require *two* key attributes in order to produce search results. The two key attributes were designed to match the two personal preferences given to participants for the corresponding task. In other words, the restaurant searchbot was designed to require location and food preference, the local attractions searchbot was designed to require location and indoor vs. outdoor activity, and the books searchbot was designed to require fiction vs. non-fiction and subgenre.

We experimented with three different searchbot conditions: *no_bot*, *bot_q* and *bot_auto*. In the *no_bot* condition, there was no searchbot and participants had to use out-of-channel search tools (mostly Google) to find information. In the *bot_q* condition, the searchbot intervened and requested the two key attributes needed to produce results. The two key attributes were requested using a scripted dialogue, for example: "It looks like you're trying to find local restaurants?", "What is your location?", "Any food preferences such as Italian, vegetarian, or vegan?" After obtaining the two key attributes, the searchbot produced its search results. The goal of this second condition was to mimic a searchbot that is able to intervene and provide relevant information, but does not learn from the conversation and must explicitly request what it needs in order to produce results. In the third and final condition, *bot_auto*, the searchbot intervened and directly produced search results without asking for any information. The goal of this third condition was to mimic a searchbot that is able to learn from the conversation and provide contextually relevant results without asking for any information.

The search results provided by the searchbot in the *bot_q* and *bot_auto* conditions were exactly the same for each task. In other words, in the *bot_q* condition, the search results provided by the searchbot were the same regardless of participants' responses. All three searchbots returned 15 results that were pre-fetched from Google Maps (restaurants, local attractions) and Goodreads (books). Similarly, all three searchbots embedded the top-3 results directly into the Slack window and provided a "see more search results" link that opened a pop-up window with all 15 results.

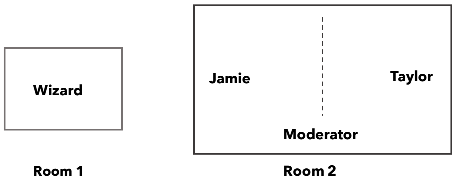


Figure 1: Physical setup of the study.

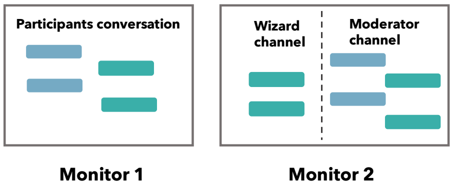


Figure 2: Wizard setup. The Wizard had access to two monitors. In monitor 1, they had access to the participants' communication channel. In monitor 2, they had access to the Wizard channel and the moderator channel.

4 METHODOLOGY

4.1 Physical Setup

This study had a moderator and a Wizard. As shown in Figure 1, the moderator was sitting in the same room as the participants, and the Wizard was sitting in a different room. The Wizard never interacted with the participants, and they too were not informed about the Wizard. Similar to the protocol used by Morris and Horvitz [9], participants were seated in the same room, but did not face each other and were asked not to communicate outside of Slack. Participants used the Slack messaging system to communicate and were also provided with a Google Chrome browser in order to perform any desired searches. As shown in Figure 2, the Wizard had access to two monitors and could only perform actions over Slack. On monitor 1, the Wizard could follow the conversation between both participants, but could not participate. On monitor 2, the Wizard had access to two channels: the Wizard channel over which they could trigger the searchbot (Section 4.4), and the moderator channel over which they could communicate with the moderator. The moderator in this study interacted with the participants and was responsible for describing to them the study protocol, Slack, and the basic functionality of the searchbot.

4.2 Guiding Principles

In this section we describe the guiding principles behind our Wizard of Oz methodology.

Minimize human-agent conversation: The focus of our study was not on the conversations between human and agent. Therefore, we decided to keep the dialog between the searchbot and participants to a minimum. In our study, a dialog between the searchbot and participants was possible only in the *bot_q* condition.

In this condition, the searchbot asked participants for information it needed to provide them with relevant results. Soon after, the Wizard would present the users with results. From there on the Wizard could no longer trigger or perform any actions with the searchbot. They could only observe the conversation or communicate with the moderator.

Reduce latency of response: We wanted to create a scenario in which the searchbot would intervene immediately after both the participants exchanged their preferences. To reduce latency in response, we gave the Wizard quick and easy to type trigger commands (Section 4.4).

Maintain consistency: We maintained consistency at three levels: (1) point of intervention, (2) results provided to the user, and (3) the same Wizard throughout the study. Keeping these three factors consistent allowed us to compare participants' perceptions and interaction across the different searchbot conditions.

Manage expectations: Recent work on conversational agents suggests that there exists a gap in expectations between what an agent can do versus what people think they can do [8]. The larger the gap, the higher the likelihood of a negative experience. To prevent a gap in expectations, we explained to participants what the searchbot could and could not do. We also explicitly informed them that they could not explicitly involve the searchbot and that they could only interact with it in one condition, *bot_q*.

4.3 Protocol of Intervention

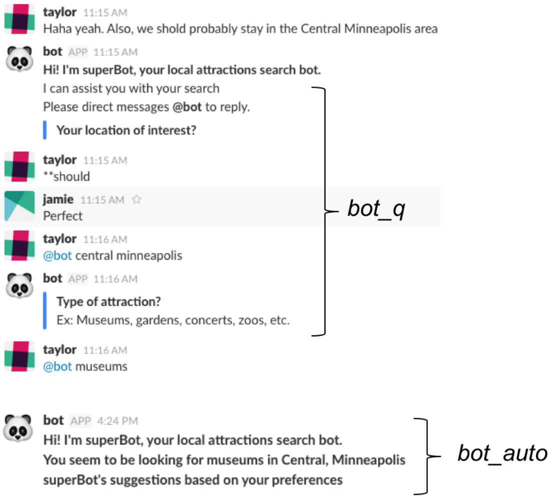
In the *bot_q* and *bot_auto* conditions the searchbot was operated by the Wizard who had access to the participants' Slack channel and was sitting in a different room (Section 4.1). The Wizard monitored the conversation and always intervened immediately after both participants mentioned their personal preferences in the conversation (Figure 3). We did this in order to keep the point of intervention consistent between *bot_q* and *bot_auto*.

By using this point of intervention, we achieved three goals: (1) we maintained a *consistent* point of intervention between the *bot_q* and *bot_auto* conditions, (2) we used a realistic point of intervention for the *bot_auto* condition (a point in which a searchbot would be able to infer the necessary information to produce search results), and (3) we created a situation in the *bot_q* condition in which participants might perceive the searchbot as having "missed" information that would have enabled it to directly produce relevant results. Additionally, we believe that this point in the conversation might often mark a sub-task-transition point (i.e., a point of low cognitive load) in which participants would be less disrupted by the intervention [1, 4, 5]. It should be noted that it was possible for participants to not "trigger" the searchbot in the *bot_q* and *bot_auto* conditions if they did not mention their personal preferences in the conversation.

4.4 Trigger commands

The Wizard could trigger the searchbot in the *bot_q* and the *bot_auto* conditions based on the intervention protocols in Section 4.3. The Wizard used the following commands to trigger the searchbot in both conditions:

***bot_q*:** In this condition, when the Wizard had to intervene, they did so by typing *"@bot h1"* in the Wizard channel. This command triggered the searchbot to first reveal to both participants that it understood the intent of their conversation but lacked enough information to search for them (See *bot_q* in Figure 3). The searchbot also requested the first piece (out of 2) of information it needed from participants. When participants responded to the question, irrespective of the response, the Wizard would trigger the searchbot once again to ask the next question. To do this, the Wizard would once again type *"@bot h1"* in the Wizard channel. If the participants responded to this question, the Wizard would then type *"@bot h7"* in the Wizard channel and the searchbot would present participants with a set of results.



taylor 11:15 AM
Haha yeah. Also, we shold probably stay in the Central Minneapolis area

bot APP 11:15 AM
Hi! I'm superBot, your local attractions search bot.
I can assist you with your search
Please direct messages **@bot** to reply.
Your location of interest?

taylor 11:15 AM
**should

jamie 11:15 AM
Perfect

taylor 11:16 AM
@bot central minneapolis

bot APP 11:16 AM
Type of attraction?
Ex: Museums, gardens, concerts, zoos, etc.

taylor 11:16 AM
@bot museums

bot APP 4:24 PM
Hi! I'm superBot, your local attractions search bot.
You seem to be looking for museums in Central, Minneapolis
superBot's suggestions based on your preferences

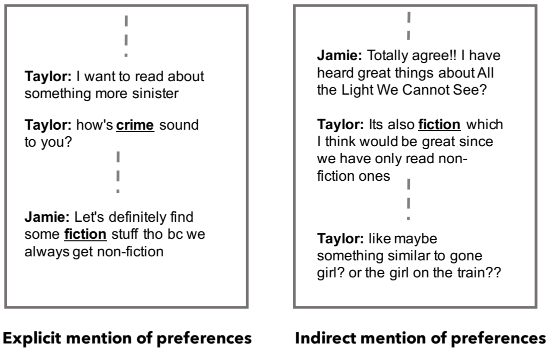
Figure 3: Searchbot intervention in the *bot_q* and *bot_auto* condition.

***bot_auto*:** In this condition, the Wizard would intervene as per the protocols in Section 4.3. To trigger the searchbot, the Wizard had to type *"@bot h1"*. The searchbot would then intervene to inform participants that it understood the intent of the conversation as well as their personal preferences. This message would be accompanied by a list of results.

After presenting results in the *bot_q* and *bot_auto* conditions, anything the Wizard typed inside the Wizard channel was irrelevant and not actionable. In the *no_bot* condition, irrespective of what the Wizard typed in the Wizard channel the searchbot would not be activated.

4.5 Challenges

Indirect reference of preferences: As per the intervention protocol (Section 4.3), the Wizard intervened after participants mentioned their personal preferences. Though participants mentioned their preferences explicitly in most cases, there were instances where this did not happen. For example in Figure 4, we show an exchange between two participants for the book task where the preferences are fiction and crime genre. In the exchange shown on the left (Figure 4), we see that participants exchange their preferences in a clear manner. In the exchange shown on the right (Figure 4), the preference "crime" is indirectly mentioned. Participants exchanged this preference by giving examples of books which are crime related. During such exchanges, where the personal preference is



Taylor: I want to read about something more sinister

Taylor: how's **crime** sound to you?

Jamie: Let's definitely find some **fiction** stuff tho bc we always get non-fiction

Jamie: Totally agree!! I have heard great things about All the Light We Cannot See?

Taylor: Its also **fiction** which I think would be great since we have only read non-fiction ones

Taylor: like maybe something similar to gone girl? or the girl on the train??

Explicit mention of preferences **Indirect mention of preferences**

Figure 4: Conversation between two participants in the book task. Personal preferences are explicitly mentioned in the left conversation and indirectly in the right conversation.

not explicitly mentioned the Wizard did not consider that as an exchange of personal preference.

In our study such exchanges are problematic. For example, in the exchange shown on the right in Figure 4 there is no explicit mention of the word "crime" or a variant of it. Suppose later in the task, one of the participants mentioned "crime". The Wizard would have to intervene due to the intervention protocol, at which point participants may perceive the intervention to be late.

Intervention–>Interruption: The rationale for our intervention protocol is based on prior work on interruptions, which suggests that sub-task transition points are well suited for interruptions [1, 4, 5]. But our protocol overlooks scenarios in which users may be engaged in active exchange (quick to and fro between participants), typing, or another activity such as browsing. Intervening in such situations may make users perceive the searchbot's intervention as an interruption. Another scenario in which an intervention may become an interruption is if participants describe their preferences much later in the task. By this point, participants may have already engaged in sufficient search activity by themselves and may perceive the intervention to be badly timed.

Ambiguity on who should respond: At no stage were participants informed over who should interact with the searchbot. At the beginning of the study participants were informed that either of them could interact, but only when the searchbot requests for information. In most cases, participants self-selected themselves to interact with the searchbot, meaning, they interacted with the searchbot without discussing with their partner. In a few cases, they discussed who should interact and in some other cases, there was ambiguity about who should interact. This ambiguity sometimes resulted in both of them responding to the searchbot. This was not a problem in scenarios in which both responded with the same information. However, there was a problem when they responded differently. As the searchbot always responded with the same results irrespective of the user's responses, it was not possible for the searchbot to consider these different responses. This made some of the participants feel ignored and frustrated with their interaction with the searchbot.

Limited Interactions: Though the WoZ methodology is meant to overcome limitations in existing technologies, such as ASR and dialog management, the study setup does introduce limitations which may not reflect a *natural* setting. For example in our system, the Wizard has a fixed set of questions which, though appropriate for our research question, may not reflect a *natural* interaction space. Such a limitation also impacts users' expectations of the system, which we discuss below.

Unrealistic expectations: At the beginning of the study, we informed users that they could not invoke the searchbot and that they could only interact with it by responding to its questions. Despite this, during the task, a few users tried invoking the searchbot or querying it for information. Some participants quickly remembered our instructions and realized the limitations, whereas others did not. This led some to report that their interactions with the searchbot were not satisfactory.

5 SUMMARY

In summary, we report on a WoZ methodology we used in our study on searchbots. In this paper, we first highlighted the importance of this methodology and how it has been employed in recent work on conversational agents. Next, we described our study and methodology. Finally, we reflected upon the challenges we faced with this methodology. We believe that the protocols we mentioned in the study, along with the challenges, will inform future lab studies on conversational systems.

Acknowledgements: This work was supported in part by NSF grant IIS-1451668. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Piotr D. Adamczyk and Brian P. Bailey. 2004. If Not Now, when?: The Effects of Interruption at Different Moments Within Task Execution. In *CHI.* ACM, 271–278.

[2] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. SearchBots: User Engagement with ChatBots during Collaborative Search. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval.* ACM, 52–61.

[3] Jay Bradley, David Benyon, Oli Mival, and Nick Webb. 2010. Wizard of Oz experiments and companion dialogues. In *Proceedings of the 24th BCS Interaction Specialist Group Conference.* British Computer Society, 117–123.

[4] Shamsi T. Iqbal and Brian P. Bailey. 2006. Leveraging Characteristics of Task Structure to Predict the Cost of Interruption. In *CHI.* ACM, 741–750.

[5] Shamsi T. Iqbal and Brian P. Bailey. 2008. Effects of Intelligent Notification Management on Users and Their Tasks. In *CHI.* ACM, 93–102.

[6] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3 (2009), 1–224.

[7] Vera Q Liao, Muhammed Masud Hussain, Praveen Chandar, Matthew Davis, Marco Crasso, Dakuo Wang, Michael Muller, Sadat N Shami, and Werner Geyer. 2018. All Work and no Play? Conversations with a Question-and-Answer Chatbot in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18). ACM, New York, NY, USA*, Vol. 13.

[8] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, 5286–5297.

[9] Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: an interface for collaborative web search. In *Proceedings of the 20th annual ACM symposium on User interface software and technology.* ACM, 3–12.

[10] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the Effects of Embodiment for a Group Facilitation Agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 391.

[11] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 208.