Building Real-World Chatbot Interviewers: Lessons from a Wizard-of-Oz Field Study

Michelle X. Zhou[1] Carolyn Wang[2] Gloria Mark[3] Huahai Yang[1] Kevin Xu[4]

[1]Juji, Inc.
San Jose, CA
{mzhou, hyang}@juji-inc.com

[2]Columbia University
New York, NY
carolynw@gmail.com

[3]University of California, Irvine
Irvine, CA
gmark@uci.edu

[4]Univ. of Pennsylvania
Philadelphia, PA
xukevinwork@gmail.com

ABSTRACT

We present a Wizard-of-Oz field study, where a human-assisted chatbot interviewed 53 actual job applicants each in a 30-minute, text-based conversation. A detailed analysis of the chat transcripts and user feedback revealed users' likes and dislikes of the chatbot, as well as the patterns of their interaction with the chatbot. Our findings yield a set of practical design suggestions for building effective, real-world chatbot interviewers that appear intelligent with even limited NLP or conversational capabilities.

CCS CONCEPTS

• Computing Methodologies → Intelligent Agents • Human-centered computing → Interactive systems and tools

KEYWORDS

Chatbot, AI Interviewer, Personality Inference, Wizard-of-OZ study

ACM Reference format:

Michelle X. Zhou, Carolyn Wang, Gloria Mark, Huahai Yang, and Kevin Xu. 2019. Building real-world chatbot interviewers: Lessons from a Wizard-of-Oz field study. In Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 6 pages.

INTRODUCTION

From recruitment to user research, interviewing is a key technique used to collect information from a target audience. Human-driven interviewing however cannot scale to large numbers and introduces potential biases [11]. To address the challenges, researchers have built intelligent agents as interviewers (e.g., [4, 13, 15, 18]). Despite the obvious benefits, building such an agent for real-world use is nontrivial because it must cover varied interview questions and handle diverse user responses [4].

Since natural language processing (NLP) techniques are far from perfect and it is unclear what users' expectations and behavior would be in real-world interview situations, we built a Wizard-of-Oz (WoZ) interviewing system to better understand the associated user and technical requirements. This system lets a human operator (wizard) drive a text-based chat with a user (Figure 1). Inspired by the work in [13], we also wanted to observe how the use of a personality engine might affect user interactions with a human-assisted chatbot interviewer. We thus incorporated a personality engine into the WoZ system to automatically infer a user's Big 5 personality traits based on the user's text given during the interview. We deployed the system to interview 53 real job applicants who applied to an internship program. We analyzed about 1600 minutes of interview transcripts to answer two sets of questions:

- **Q1** What were the users' impressions of a human-assisted chatbot interviewer?
  - What did the users like about the chatbot?
  - What did the users dislike about the chatbot?
- **Q2** How did users interact with a chatbot interviewer?
  - What were the characteristics of user interactions?
  - Which interactions could be supported in practice?

The first set of questions is to understand users' perceptions of a human-assisted chatbot interviewer, especially what they like and dislike. The second set of questions is to uncover a set of practical features that make a chatbot intelligent with even limited NLP or conversation capabilities.

As we discuss shortly, the answers to the above questions reveal new insights into user interactions with a chatbot interviewer in a real-world application. Moreover, the answers help formulate a set of practical design suggestions for building effective, real-world chatbot interviewers.

RELATED WORK

Our work is inspired by recent efforts on using virtual in-

Agatha:	I took the liberty to analyze your personality from the text in your social media account. See it on your right.
User:	Hm, I think there's a bug
Agatha:	What is a bug?
User:	It says data source twitter, and there are no words showing up.

Figure 1. A recorded example interview fragment between the Chatbot Agatha driven by a human (wizard) and a user.

IUI Workshops'19, March 20, 2019, Los Angeles, USA.

Copyright © 2019 for the individual papers by the papers' authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors.

terviewers to aid in information collection. For example, Li et al. show that actual job candidates are willing to confide in and listen to chatbot interviewers [13]. Several studies also indicate how an embodied virtual agent elicits more sensitive information from users than a human interviewer can (e.g., [15, 20]). While prior work demonstrates the benefits of a virtual interviewer, little is known on the requirements for creating a chatbot interviewer for effective real-world use. For example, what NLP skills should a chatbot possess to support practical interview situations? How would a chatbot's ability to understand a user's personality affect their interview experience? The purpose of this study is to find answers to these questions.

WIZARD-OF-OZ (WOZ) INTERVIEWING SYSTEM

Our WOZ system is a web-based system that offers a text-based chat interface with which a user can interact. A human operator (wizard) uses the interface to drive an interview. The wizard can ask questions and compose responses, including calling system functions (e.g., calling a function to analyze a user's personality and display the results).

FIELD DEPLOYMENT

We deployed the WOZ system to aid a software startup in hiring three summer interns from over 650 applicants. First, 60 candidates were selected based on their stated technical interests and experience relevant to the positions. Each candidate was invited to participate in a 60-minute interview: a 30-minute chatbot (wizard-of-oz) interview followed by a 30-minute phone interview. The same person served as the wizard and the phone interviewer.

The wizard used an interview agenda with a set of questions to guide each interview. At the beginning of each interview, the candidates were informed that: (1) they would be interviewed by a chatbot, (2) they would be asked about their interview experience, and (3) the chatbot would analyze their conversation and infer their personality.

To start an interview, a candidate logged into the WOZ system with his/her Facebook or Twitter account. The interview included four parts. First, the chatbot named Agatha asked the candidate to make a self-introduction. Agatha then displayed the system-inferred Big 5 personality traits based on the user opt-in Facebook posts (up to 200) or tweets (up to 3200) during the login. The personality inference engine used in the study is similar to the one described in [13]. The candidate was asked to evaluate and comment on his/her inferred traits. The second part of the interview included a set of circumstantial questions where a candidate was asked to provide their assessment of a situation and propose solutions. For example, one question was about handling software defects before a deadline. The third part of the interview included a set of casual inquiries about the candidate (e.g., "*If you had a super power, what would it be*?"). The last part solicited the candidate's impression of the chatbot ("*What's your impression of me*") and input for future improvements ("*What should I improve on*").

During each interview, the wizard intentionally did not interpret long or complex user input and kept her response simple without getting into a deep conversation on any topic. She did so for two reasons. One was to test users' impressions of a capable but realistic chatbot, since even the most advanced chatbot is unlikely to understand or respond to every user input. The other was that the wizard could not afford detailed responses due to time constraints, as it takes time for a person to digest a complex input, compose a thoughtful response, and type it into the system.

The whole process lasted about two and half weeks, during which 53 candidates completed their interviews. All were university students with 20 (37.7%) females and 33 (62.3%) males. On average, each user answered about 18 interview questions and input about 500 words.

RESULTS ANALYSIS AND FINDINGS

To answer the two sets of questions (Q1 and Q2), three coders read and analyzed each interview transcript using an open-coding approach. Below we report the results.

Q1: User Impressions of the Chatbot

At the end of each interview, a candidate was asked of his/her impression of the chatbot interviewer. The top-10 keywords used to describe the chatbot were: *understand*, *responsive*, *natural*, *interesting*, *human-like*, *friendly*, *fluency*, *believable*, *fun*, and *cool*. While unaware of the wizard's presence, most users (92%) described the chatbot as interesting and intelligent, almost like a real person. To extract a list of user likes and dislikes, each coder went through all user comments independently and then worked together to merge their lists. Three categories emerged from the coding, shown in Table 1.

User Likes

In general, the users liked the chatbot's language capabilities and thought it asked questions naturally and responded to them well. Here is what a user said to the chatbot:

> "*Your responses sounded very natural and real, it could have almost been a live human.*"

Regarding the chatbot's conversation capability, the users felt that the chatbot was attentive and engaging. They especially appreciated that the chatbot made an effort to learn during the conversation. For example, when a user asked the chatbot to tell a joke, the chatbot responded "*what is joke*". Upon receiving the user's answer, the chatbot thanked the user. These simple exchanges made the users perceive the chatbot as honest, engaging, and willing to learn. For example, one user told the chatbot:

> "*I think you're doing a pretty good job so far! Sometimes you don't understand my questions, but you are still learning and I can tell you're making an effort to learn more by asking me questions*".

Likewise, another user expressed:

"*You ask more questions than you give answers, indicating that you are focused on me and wish to maintain the conversational flow*".

The interviewees were also impressed by the chatbot's ability to analyze their personality during the interview. For example, one user told the chatbot:

"*It's the first time I've ever done something like this… I'd say your analyses were generally accurate*".

Overall, users' positive perceptions were encouraging especially as the wizard intentionally did not interpret many user inputs and kept the chat simple and shallow due to time constraints as well as to set realistic expectations. This suggests that practical solutions could satisfy users with even limited NLP.

User Dislikes

The users pointed out several aspects they did not like about the chatbot and were in need of improvements.

Concerning language, 66% of users mentioned that the chatbot could be improved to carry out a deeper, more interesting conversation. For example, per one user:

"*You need to have more knowledge, … your responses will become interesting, not just some simple answers*".

For example, the wizard simply chose not to answer a user question like "*What do I need to know about myself*?"

On the conversation capabilities, the main complaint was about the conversation timing—untimely interruptions during an interview. After a user texted the first response to a question, the wizard often continued with the next question without waiting for more user input, mainly due to the time constraint. In some cases, the users might still be typing or wanting to give more input, but felt that their thoughts or responses were interrupted untimely. In reality, the human wizard found it difficult to determine the response timing especially since she had little knowledge of a user' habit (e.g., fast or slow in response).

Several complained that the chatbot was unable to "remember" and learn from their exchanges ("*I already told you that I like basketball…*")

Concerning the chatbot's personality, some users felt that the chatbot was unlike a real person because it was too emotional:

"*You are a bit too emotional when you respond. People don't really use the punctuations that you use*."

However, others felt that the chatbot lacked personality or strong opinions:

"I see the lack of personality within your sentence structure or word choices."

"*you need to have more strong, personal opinions… and the ability to keep the conversation going… maybe gives me more opinionated feedback*"

Since all the users interacted with the same wizard, we suspected that such impressions might be affected by the users' own personality. We however did not have sufficient data to validate this hypothesis.

	Likes	Dislikes
Language Capability	Asks and responds to questions well (92%)	Unable to ask specific follow-up questions and interpret certain input (66%);
Conversation Capability	Natural flow, attentive and engaging, listens well (45%)	poor timing (15%), a lack of slang use (10%), didn't explain itself (10%), no memory (8%)
Personality	Friendly, polite, cute, receptive (15%)	A lack of personality or strong opinions (8%)

Table 1. User likes and dislikes of the chatbot interaction.

Q2: User Interactions with the Chatbot

Extracting user likes and dislikes helped answer the first set of questions on users' impressions of a human-assisted chatbot. To support the user likes and avoid the dislikes, we must answer the second set of questions to discover what takes to build an effective chatbot interviewer.

As shown in Table 1, 92% of users thought the chatbot was capable of understanding them, but yet 66% of them hoped such a capability to be further improved. Existing work shows that how an interviewer responds to interviewees during an interview largely affects one's interviewing experience [5, 14]. We thus analyzed each WoZ interview transcript to identify how the wizard responded to users' questions/requests during an interview, which would help explain the users' impressions and expectations.

Each coder first extracted all user questions/requests from the transcripts independently and then merged their lists. They identified a total of 328 user requests and classified them into six categories (Table 2). The wizard responded to 200 such requests (response rate 61%) during the interviews. The top three user questions/requests asked about the chatbot's personality analysis (32.6%) and the chatbot itself (27.7%), and requested conversation continuation (20%). The chatbot responded to these three types of questions 66.4%, 75.8%, and 56.7% of the time, respectively. In most of these cases, the wizard used canned responses (e.g., answering about the chatbot's personality analysis). To avoid an unbounded conversation, the wizard answered few general user requests ("*Tell me a joke*").

As indicated by their questions, the users showed great interest in their personality analysis results regardless the accuracy of the results. In fact, the system did not perform analysis for nearly half of the users (26) because of a glitch in getting their social media data. Although the system displayed random results and indicated zero words analyzed, most users except one did not notice the glitch, and inquired and argued about the results. For those who obtained an actual result, 67% thought their result was accurate.

Category	Example	Total	Answered
Personality Analysis	*"How do you discover the traits of a person?" "How did you know I am open?"*	107	66.4%
About Chatbot	*"Do you have a Facebook profile?" "How old are you?"*	91	75.8%
Conversation Continuation	*"What's next?" "Do you have any more questions?"*	67	56.7%
General	*"Do you know what an idiom is?" "Tell me a joke"*	43	39.5%
Rhetorical	*"Well, bugs aren't going to fix themselves, right?*	9	22.2%
User	*"What do I need to know about myself?"*	11	27.3%

Table 2. Types of user question or request extracted from the interview transcripts.

Moreover, we observed *interaction reciprocity*—a user asked a chatbot the same questions that s/he was asked during an interview. For example, the users were asked "*what's your super power*?" Quite a few users asked the chatbot the same question when they were invited to ask a question. Users asked a small number of random general questions (43 out of 328).

DESIGN SUGGESTIONS AND DISCUSSION

Although our WoZ study has its limitations (e.g. only 53 users in one specific use), our findings described above often two valuable insights. First, our analysis shows that the wizard who used limited NLP and did not respond to every user input still impressed the users being very human-like. This suggests that a chatbot interviewer could be built with limited NLP. However, it should focus on responding to frequently asked user questions/requests related to an interview. Such responses even if they are canned or simple would still make users feel the chatbot understand and respond to them well. Second, the types of user questions/requests in an interview can be *anticipated* (see below), which in turn helps prepare a chatbot interviewer to handle user input robustly with even limited NLP.

Below we outline a concrete set of design suggestions for building effective chatbot interviewers. These suggestions aim at enabling a set of practical features that make a chatbot interviewer appear intelligent with limited NLP or conversation capabilities.

Active Listening Skills

Effective human interviewers actively listen to their interviewees to better engage with them [22, 14]. One way to build an effective chatbot interviewer is to empower it with active listening skills, such as repeating a user's input to make the user feel s/he is heard [3, 14, 22]. As one user suggested, the chatbot should just "*repeat the last three words they say*". To do so, a chatbot incorporates a user's expressions in its response. For example, if a user mentioned "*I love to cook*", the chatbot could ask: "*I know you like to cook. Why do you enjoy it?*" Although this feature will require a chatbot to parse a user's input, it does not require perfect NLP and a partial understanding of a user input (e.g., parts of speech) will go a long way.

Being Honest and Humble by Asking Questions

Users liked the attentive and honest behavior of the human-assisted chatbot (Table 1). The wizard posed questions to avoid getting into a deep conversation. Such behavior can be robustly supported when a chatbot encounters unknown words or expressions in user input. For example, in the WOZ study, a user asked the chatbot "*Do you know idioms*?" The chatbot asked "*What is it idiom*?" Not only will such a question make a user feel engaged, but it will also help the chatbot "learn" new concepts. In the above example, the unknown words ("*idiom*") and the associated user explanation can be recorded and later used by the chatbot to answer similar user questions in the future.

Anticipating User Questions/Requests

Instead of providing general NLP capabilities, we can build targeted NLP capabilities by anticipating user interactions during an interview. Table 2 shows that 81% of user questions/requests fell into three categories, of which corresponding answers could be prepared in advance. For example, we can anticipate users' asking about the interview context, such as the chatbot's origin and capability. We can also anticipate user questions based on interaction reciprocity and prepare a chatbot with answers to all its interview questions (e.g., "*what is your super power*").

While it might not be feasible to anticipate all user behavior or to make a chatbot interviewer understand and respond to every user input, our findings suggest that user interactions are not random during an interview and many of them can be anticipated and handled effectively.

Pacing a Conversation Intelligently

Learning from our WOZ study, we suggest that a chatbot use three sources of information to pace a conversation. One source is to detect a user's keystroke activities. If a user is still typing, the chatbot could then wait until the typing is done. Another source is to model a user's pace, such as tracking her average response time, and then use the information to pace a conversation with this user. The third source is to detect the completeness of the content in the current response. If the current response is fairly complete, the chatbot can then move on without waiting for additional input from a user. However, judging response completeness may not be easy as it may be question-dependent. Alternatively, the system could assess the *informativeness* of a response based on information entropy [7].

Personalizing an Interview

Effective human interviewers personalize a conversation to better engage with their interviewees [5, 19]. Our study also showed that users were interested in the system analysis of their personality regardless its accuracy. One way to build an effective chatbot interviewer is to power it with a personality inference engine like the one used in our WOZ study. The analysis result could help a chatbot personalize a conversation and encourage a user to open up. For example,

a chatbot can analyze a user opt-in social media content at the start of an interview and use the inferred personality to pose tailored questions. Assuming that the chatbot infers a user high on creativeness, it could ask "*It seems you are very creative, what's the most creative thing you have done*?" Such a personalized conversation makes users stay engaged and motivates them to cooperate [12, 17]. Moreover, such information could be used to adapt the chatbot's personality to a user's [17] or fit for an interview task [13].

CONCLUSIONS AND FUTURE WORK

We are building fully automated chatbot interviewers that can support diverse real-world interview situations. To develop an effective chatbot, we conducted a WOZ field study where a human-assisted chatbot interviewed 53 actual job applicants. Our findings revealed what the users liked or disliked about the chatbot along with a set of user interaction patterns coincident with such opinions. Based on the findings, we formulated a set of practical design suggestions for building effective, real-world chatbot interviewers with even limited NLP capabilities. Based on these design suggestions, we are building chatbot interviewers that can function in varied interview contexts, such as job interviews and customer interviews.

REFERENCES

1. Adali, S., & Golbeck, J. Predicting personality with social behavior. *ASONAM'2012*, 302–309.
2. Bickmore, T., Gruber, A., & Picard, R. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, 2005, 59(1): 21-30.
3. Decker, B., 1989. How to communicate effectively, Page, London, UK.
4. DeVault, D., Artstein, R. Benn, G., Dey, R. Fast, E., Gainer, A., Georgila, K. Gratch, D., Hartholt, A., Lhommet, A., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, S., Suri, A., Traum, D., Wood, W., Xu, Y., Rizzo, A., and Morency, LP. SimSensei kiosk: a virtual human interviewer for healthcare decision support, *Proc. AAMAS 2014*.
5. DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical education*, *40*(4), 314-321.
6. Digman, J. M. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology 1990,* 41(1): 417–440.
7. Ebrahimi, N., Maasoumi. E., and Soofi. E. Measuring Informativeness of Data by Entropy and Variance, 1999, *Advances in Economics, Income Distribution and Scientific Methodology* 61-77.
8. Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology, 97*, 866-880. doi: 10.1037/a0026655.
9. Grieve, R. and Watkinson, J. The Psychological Benefits of Being Authentic on Facebook. *Cyber Psychology, Behavior, and Social Networking*, 2016, 19(7): 420-425.
10. Gou, L., Zhou, M.X., & Yang, H. KnowMe and ShareMe: Understanding automatically discovered personality traits from social media and user sharing preferences. *CHI '14*, 955–964.
11. Jackle, A., Lynn, P., Sinibaldi, J., & Tippping, S. The effect of interviewer experience, attitudes, personality, and skills on respondent cooperation with face-to-face surveys. *Survey Research Methods*, 2013, 7(1): 1
12. Lee, M., Forlizzi, J., Kiesler, S., Rybski, P., Antanitis, J., and Savetsila, S. Personalization in HRI: a longitudinal field experiment, Proc. ACM/IEEE international conference on Human-Robot Interaction, 2012.
13. Li, J., Zhou, M.X., Yang, H., and Mark, G. Confiding in and listening to virtual agents: The effect of personality. *Proc. ACM IUI 2017*, 275-286.
14. Louw, S., Todd, R. W., & Jimakorn, P. (2011). Active listening in qualitative research interviews. *Proceedings of the International Conference: Doing Research in Applied Linguistics*, 71-82. Retrieved from http://arts.kmutt.ac.th/dral/.
15. Lucas, G., Gratch, J., King, A., & Morency, L. It's only a computer: virtual humans increase willingness to disclose. *Comp. in Human Behavior*, 2014, vol. 37: 94-100.
16. McCrae, R. and Costa, P. (1999) The five factor theory of personality. in *Handbook of Personality: Theory and Research*, L.A. Pervin, O.P. Johns, NY: Guilford, 139-153.
17. Nass, C., Steuer, J. and Tauber, E. "Computers are social actors." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 72-78. ACM, 1994.
18. Nunamaker, J., et al. "Embodied conversational agent-based kiosk for automated interviewing." Journal of Management Information Systems 28.1 (2011): 17-48.
19. Okun, B. Effective Helping: Interviewing and Counseling Techniques. 7th Edition, Cengage Learning, 2007
20. Pickard, M., Roster, C., and Chen, Y. Revealing sensitive information in personal interviews: Is self-disclosure easier with humans and avatars and under what conditions? Computers in Human Behavior, 2016, vol. 65: 23-30.
21. Turk, M. Multimodal interaction: A review. Pattern Recognition Letters 36: 189-195 (2014).
22. Weger, H., Jr., Castle, G. R., & Emmett, M. C. (2010). Active listening in peer interviews: The influence of message paraphrasing on perceptions of listening skill. *International Journal of Listening*, 24, 34-49.