WIZARD-OF-OZ TESTING AS AN INSTRUMENT FOR CHATBOT DEVELOPMENT

An experimental Pre-study for Setting up a Recruiting Chatbot Prototype.

Prof. Dr. Stephan Böhm | Judith Eißer | Sebastian Meurer
RheinMain University of Applied Sciences, Wiesbaden
Faculty Design – Computer Sciences – Media
Degree Program Media Management
AGENDA

1. Introduction
2. Theoretical/Research Background
3. Methodical Approach
4. Preliminary Findings & Implications
5. Conclusion, Limitation & Outlook
01
INTRODUCTION
Introduction: Chatbot Development
Wizard-of-Oz within Chatbot Development

- Chatbots automate repetitive stakeholder inquiries within internal and external business communication processes (Schildknecht et al., 2018; Genwuch et al., 2020)

- A suitable conversational design concept is imperative considering the envisioned users’ requirements and expectations (Jurafsky & Martin, 2018; McTear, 2017)

- Integration of early user feedback is crucial within development (Schlögl et al., 2014; Böhm et al., 2020; Dahlbäck et al., 1993)

- Wizard-of-Oz (WOz) experiment as way to yield early stakeholder feedback and thus necessary input for chatbot creation (Schlögl et al., 2014; Jurafsky & Martin, 2018)
  - Executors lead test subjects to believe that they interact with a fully developed technological system (Dahlbäck et al., 1993)
  - In reality, a human operator in disguise serves as chatbot (Dahlbäck et al., 1993)
Introduction: Study approach
Goals and Design of the Study

* Practical example of an FAQ chatbot for recruiting
* Embedded in broader chatbot user testing scenario

 Goals:

- **evaluate the intent database** of the developed recruiting FAQ chatbot prototype in terms of **relevancy** and answer **suitability**,
- collect **feedback on the conversational design** and specifically the (1) preliminary content, (2) the perceived user satisfaction, (3) the user’s level of acceptance, and (4) utilization limitations, and
- **yield not yet considered but relevant content** in the form of novel chatbot intents, as well as potential training data for the chatbot.
02
THEORETICAL/RESEARCH BACKGROUND
(CATS – CHATBOTS IN APPLICANT TRACKING SYSTEMS)
Chatbot Prototyping and Development

General Overview

Chatbots belong to the field of Human-Computer Interaction (HCI) (Folstad & Brandtzaeg, 2017) and are conversational interfaces (McTear, 2017).

User testings are commonly integrated into the system design process (Nielsen, 2014; Landauer, 1996).

Multiple requirements such as adequate and useful reaction to input, behavioral appropriateness, friendliness.

Frontend user interface is not well influenceable; the content itself and the way of communication are in focus of chatbot designing.
Wizard-Of-Oz Experiments (1)

WOz for Technological Innovations

Term stems from children’s book: character hides behind a curtain to control a scene from remote pretending to be a powerful wizard (Baum, 1900)

Simulation where the researchers interact with the users themselves in a concealed way posing as a fully functioning technology (Eynon & Davies, 2012; Eißer & Böhm, 2017)

Human mediates the conversation to circumvent the contraints of current technology by pretending to showcase an operating technology (Dahlbäck et al., 1993)

Long established method (Schlögl et al., 2014) representing a practical, resource-saving way of early user testing

No need for a full-fledged prototype to yield first feedback

However, WOz is no holistic testing approach but rather gives first ideas within a realistic scenario (Jurafsky & Martin, 2018)
Wizard-Of-Oz Experiments (2)
WOz for Chatbot Development in Specific

🌟 WOz advantages can be well exploited within chatbot development

- Especially AI components can be mimicked without the necessity of sophisticated AI framework implementation

- Chatbots are bound to predefined databases and thus input; early WOz-based prototype tests reveal unexpected and thus otherwise non-considered content (Guerin, 2011)

- WOz approaches have commonly been applied to chatbot research (e.g., Eißer & Böhm, 2017; El Asri et al., 2017; Guerin, 2011; Kearns et al., 2020; Riek, 2012; Quarteroni & Manandhar, 2007)

🌟 In this study, the experiment yields (1) relevant intents and (2) accompanying training as well as test data for a recruiting FAQ chatbot prototype providing detailed insights into the framework and its implementation
03

METHODICAL APPROACH
Wizard-of-Oz Study
Goals & Study Design

(1) Intent matching & answer suitability assessment
Wizard reflects not only the functions but also the limitations of the intended chatbot. We used an initial intent set and predefined answer phrases.

(2) Conversational design evaluation
The FAQ chatbot’s (1) content, and (2) the experience with the chatbot in the specific application area of recruiting FAQ were assessed.

(3) Intent generation
Besides testing of already implemented topics in our recruiting FAQ chatbot concept, further information needs (user intents) need to be identified. Apart from intent generation, potential training data can be derived.

The WOz approach is utilized to test and validate the recruiting FAQ chatbot prototype from the corresponding perspectives. Based on the findings, the chatbot will be iteratively adapted and enhanced.
Wizard-of-Oz Study
Setup of the experiment / Conceptualization

Roles during the experiment

<table>
<thead>
<tr>
<th>Participant (P)</th>
<th>Wizard (W)</th>
<th>Moderator (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Chatbot users</em> belonging to the target group of potential <em>candidates</em>, who converse with the chatbot during their application process.</td>
<td><em>A researcher</em> operates the WOz framework by sending preformulated messages or creating ad-hoc responses as seemingly AI-based automated answers on remote based on the experimental study framework.</td>
<td><em>Another researcher</em> accompanies the participant through the experiment giving an introduction, instructions, and guidance through the process.</td>
</tr>
</tbody>
</table>

Participant side experiment flow

1. Experiment/Use Case introduction (M)
2. Provision of exemplary application documents for the participants (M)
3. Quantitative survey 1 (M/P)
4. WOz-Experiment – Chatbot utilization (P)
   - Selection of a predefined job ad in portal
   - Start of *application process* – information and application via a specially configured testing application platform
   - During application process, upcoming questions should be directed to the “chatbot prototype”
   - Reply to participants’ inquiries by Wizard (W)
   - Document upload/application submission
5. Qualitative user feedback via a thinking aloud during chatbot usage in application process (P)
6. Quantitative survey 2 (M/P)
Wizard-of-Oz Study
Setup of the experiment / Technical infrastructure: Participant frontend perspective

Participants (potential candidates) operate in exemplary career portal
Chatbot prototype is accessible during the whole application process
All upcoming problems (questions, irritations etc.) can be directed to the chatbot prototype
Messages sent by the respondents via chat window are forwarded to our chat solution (Rocket.Chat)
The researcher acting as wizard can utilize an administration interface to receive and process incoming inquiries while posing as a chatbot
Wizard-of-Oz Study
Setup of the experiment / Technical infrastructure: Wizard frontend perspective

Chatbot Cockpit
- Send predefined answers by pressing buttons sorted by topics on cockpit webpage
- Enter answers in real-time if no predefined answer available

FAQ Chatbot

Wizard (Backend)

... operates WOz cockpit to simulate chatbot

In a specially designed cockpit (administration interface) within a web application the wizard can either

- (1) choose from predefined answers related to a specific intent considered in the predefined intent set
- (2) take a predefined answer and modify it to meet unexpected input or
- (3) enter answers in real-time to create individual content for distribution to the participant
Wizard-of-Oz Study

Setup of the experiment / Technical infrastructure: WOz-Framework

- Overview on WOz-Framework embedded into the overall study design, including different roles, channels, and processes.
- Servers hosting the chat environment (Rocket.Chat) and the career portal as central parts.
- Participants access the framework from front-end perspective (lefthand side) while the wizard operates in secret from the backend perspective imitating the expected FAQ recruiting chatbot (right-hand side).
04
PRELIMINARY FINDINGS & IMPLICATIONS
Findings: Wizard-of-Oz Study

Overview

8 users in total, actively took part in the WOz experiment. One participant did not use the chatbot to get support and thus was excluded from the analysis. Another user had to be excluded as changes in the setup of the WOz environment were required (due to a shift on remote execution). 6 users were part of the analysis.

In Accordance with the study goals (focus on intent matching, answer suitability assessment and conversational design evaluation), the findings of this pre-study can be split up in the following three parts:

(1) Metrics on Chatbot Interaction
Chatbot sessions; interactions; wizard answers (1) via button, (2) edited, and (3) freely created; human handover requests

(2) Quantitative User Experience Survey(s)
(1) Overall and (2) process phase specific satisfaction rating in the field of user experience

(3) Qualitative User Feedback
Latency times, complexity capacity (in terms of multiple intents for one input etc.), degree of answer detail/superficiality, usability, content, scope, authenticity
Findings: Wizard-of-Oz Study
Metrics on Chatbot Interaction

- Participants interacted with wizard in **79 chatbot sessions** (= a coherent sequence of interactions associated with a single user intent)
- The ratio of chatbot interactions per session (c) varied between 1.00 and 1.43 (mean 1.21): Activation for/ intensity of use varied greatly; some respondents expected a prompt answer, where others got more involved in an interactive dialog.
- Wizard’s response behavior (three different response options): **63 percent** (55; d) of the wizard’s responses were given by predefined answers (d), for about **one-third** of the user requests (28; f), there was no matching intent.
- Average response times:
  - 20s for predefined “button answer”
  - 34s for edited, predefined answers
  - 32s for free answers by wizard
Findings: Wizard-of-Oz Study
Quantitative User Experience Survey

Experiment evaluation criteria

<table>
<thead>
<tr>
<th>(Absolute Values; N = 7)</th>
<th>completely satisfied</th>
<th>rather satisfied</th>
<th>moderately satisfied</th>
<th>rather not satisfied</th>
<th>not at all satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer Completeness</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Competence</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Speed and Performance</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>General Added Value</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- User satisfaction with regard to answer completeness is rather indifferent
- Answer quality and thus perceived competence of the chatbot is also considered moderately
- Satisfaction rating with the chatbot performance (speed of the chatbot answers) recognizably poor due to the character of the WOz project (human simulation of chatbot)
- Findings do not seem to have influenced the perception of general added value of chatbots – six of the seven participants consider the tested use case as relevant in applicant support.
Findings: Wizard-of-Oz Study
Quantitative User Experience Survey

- Assessment of the added value
- The three areas of greatest attributed added value are
  - Questions about the job advertisement,
  - Questions about the registration process and
  - Questions concerning the application process in general as well as the further procedure

- Interestingly, the three parts
  - Data protection and privacy,
  - Application submission and
  - Feedback/status inquiry after application

have been found to yield least added value
Findings: Wizard-of-Oz Study
Quantitative User Experience Survey

- I can very well imagine using the chatbot regularly when applying.
- I perceive the chatbot as unnecessarily complex.
- I consider the chatbot to be easy to use.
- I think I would need technical support to use the chatbot.
- I think that the different functions of the chatbot are well integrated.
- I think there are too many inconsistencies in the chatbot.
- I can imagine that most applicants quickly learn to master the chatbot.
- I experience the operation of the chatbot as very cumbersome.
- I felt very safe when using the chatbot.
- I had to learn a lot of things before I could work with the chatbot.

- Questions concerning the user experience and usability of the chatbot
- Generally found to be
  - Easy to use and
  - Fast to be learnt
- Hence, chatbots can potentially be quickly adopted and easily mastered
Findings: Wizard-of-Oz Study
Qualitative User Feedback

Qualitative input was given concerning

- **Latency times**: Chatbot answers perceived as delayed due to human operation
  - Participants adapted question behavior (e.g., by reformulating or reducing the number of questions)
  - One participant suspected human behavior to be the reason for delay

- **Question complexity**
  - No identification of multiple intents in a single user prompt considered as intended in real prototype
  - Ignoring question portions led to misunderstandings and confusions

- **Perceived superficiality** of the answers
  - Not satisfactory as to the degree of detail
  - Improvements regarding contexts and specifics needed

- **Usability issues**
  - Typing indicator as requested feature for chatbots
  - Positioning of the interface window (hidden too low on the right for some participants)

- **Missing intents**
  - Several missing intents were uncovered, for example concerning the career portal
05

CONCLUSION, LIMITATION & OUTLOOK
Conclusion and Outlook

Conclusion

❖ This research demonstrates how WOz experiments can be utilized for FAQ recruiting chatbots

❖ In WOz scenarios, participants can be credible convinced to interact with a chatbot

❖ It became apparent that users do not automatically accept support offered by a chatbot and do not necessarily enter into more comprehensive dialogues with such as system

❖ The utilization of an FAQ recruiting chatbot is seen as easy to master and overall valuable

❖ Not only does the chatbot need to provide suitable answers, but it also needs to point out necessary simplifications in case of complex inquiries
Conclusion and Outlook
Limitations and Implications for Further Research

Limitations

- Only 8 (6) participants → larger test or survey necessary to generalize results
- General difficulty to maintain consistent wizard behavior and to mimic errors or suboptimal technical system performance
- Response times must be reduced through further wizard trainings in order to offer a realistic test scenario
- 30 percent of the answers needed to be typed freely/anew without content from the predefined database → database needs to be updated and enlarged

Outlook/Suggestions for future research

- Response times must be reduced through further wizard trainings in order to offer a realistic test scenario
- Further research profits from these findings through chatbot infrastructure optimization
- Other studies might look at other domains or speech-based dialogue systems
DO YOU HAVE ANY QUESTIONS?

THANK YOU!

CENTRIC Paper Presentation 2020

Sebastian Meurer | Judith Drebert | Prof. Dr. Stephan Böhm
judith.drebert@hs-rm.de; stephan.boehm@hs-rm.de

RheinMain University of Applied Sciences, Wiesbaden
Faculty Design – Computer Sciences – Media
Degree Program Media Management
CONTACT

Judith Eißer, M.Sc.
Sebastian Meurer, M.A.
Research/Scientific Assistant
Faculty Design – Computer Sciences – Media
Degree Program Media Management

RheinMain University of Applied Sciences
Postal address:
Postbox 3251 | 65022 Wiesbaden
Visiting adress:
Unter den Eichen 5 | 65195 Wiesbaden | Building F
(Officio II), 1st Floor, Room 110/112

E-Mail: judith.eisser@hs-rm.de
sebastian.meurer@hs-rm.de
Phone: +49 611 9495-2290
+49 611 9495-2306
www.hs-rm.de

Authors
Prof. Dr. Stephan Böhm
Sebastian Meurer
Judith Eißer
Publicity regulation
CATS – Chatbots in Applicant Tracking Systems

This project (HA project no. 642/18-65) is funded in the framework of Hessen ModellProjekte, financed with funds of LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).

For more information: www.innovationsfoerderung-hessen.de
BACKUP & APPENDIX
Literature/Bibliography


Literature/Bibliography


