Engagement and Usability of Conversational Search – A Study of a Medical Resource Center Chatbot

Fergencs, Tamás; Meier, Florian Maximilian

Published in:
Proceedings of iConference 2021

DOI (link to publication from Publisher):
10.1007/978-3-030-71292-1_26

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Engagement and Usability of Conversational Search – A Study of a Medical Resource Center Chatbot

Tamás Fergencs and Florian Meier

Science, Policy and Information Studies
Department of Communication and Psychology
Aalborg University, Copenhagen, Denmark
tfergencs@gmail.com,fmeier@hum.aau.dk

Abstract. Due to advances in natural language understanding, chatbots have become popular for assisting users in various tasks, for example, searching. Chatbots allow natural-language queries, which can be useful in case of complex information needs, and they provide a higher level of interactivity by displaying information in a dialog-like format. However, chatbots are often only used as auxiliaries for a graphical search user interface (SUI). Thus, they must be engaging and usable so that users both want to and are able to use them. In this study, we conduct a controlled interactive information retrieval experiment following a within-subject design to compare a chatbot to a graphical SUI in terms of engagement and usability. Our findings point towards the need for flawless usability in order for conversational search interfaces to (1) be able to provide additional value in information retrieval tasks and (2) elicit a higher level of engagement compared to their SUI-based counterparts.

Keywords: Conversational search · search user interface · usability · user engagement · chatbot

1 Introduction

Conversational interfaces are becoming increasingly popular due to the advancement in natural language understanding technology. They enable human-computer interaction via natural speech or text instead of using buttons and menus and allow for a more human-like dialog with a system. Text-based conversational interfaces, so-called chatbots, have been around for quite some time, but they gained commercial interest only recently, due to digital communication becoming a standard [7]. The number of customer service chatbots is increasing as businesses explore the possibilities of conversational commerce to interact with and provide support for customers [6,11,40]. Mobile health solutions are starting to utilize conversational agents to promote health or facilitate recovery [8,13,23]. Chatbots also find their way into the field of education, where they inform university students about school facilities or act as teaching agents to supplement

classroom learning [12,26]. Regardless of the field of application, conversational agents are seen as a useful tool to facilitate user engagement — they can motivate increased usage of an application, enrich business-to-consumer interactions, or simply serve as "wow factor" for marketing purposes. One specific use case of chatbots is assisting with searching and retrieval of web content — a concept denoted as conversational search [25]. Instead of examining a lengthy FAQ page, a user can simply submit their question to a chatbot, which queries the database and returns a relevant answer [18]. Or, a library chatbot can help in promptly retrieving reading material based on the user's preferences [1,37]. If the natural language understanding framework is robust enough, users can submit even complex search queries, which can be useful in cases where the information need is difficult to formulate – e.g. in the case of non-targeted searching, where exploration of the collection is the main activity [35].

A chatbot is often used as an auxiliary to a website search interface, and not as a standalone search system. If the chatbot is not engaging enough, the initial interest can quickly fade, and users will return to using the website search. Chatbots can increase user engagement by enhancing interactivity, that is, by delivering information in a human dialog-like manner [30]. It is, however, uncertain whether a higher level of interactivity is enough for users to prefer using the chatbot if there is an alternative. Besides, implementing search functionalities to a conversational interface is not a straightforward process and, even if it's successful, users may have trouble transitioning from a traditional graphical search user interface to a conversational interface [38]. This is due to the inherently complex nature of search behaviors, which generally do not adhere to a simple query-answer model, but are rather characterized by constantly evolving information needs [2]. A search chatbot, therefore, should satisfy both the need for enhancing user engagement and serve as a user-friendly supplement (or even substitute) to a graphical SUI. If the chatbot has poor usability, people may not be able to use it. If the chatbot does not motivate engagement, people may not want to use it.

This study aims to compare the conversational search user interface (chatbot) of a medical resource center database, with its graphical search user interface in terms of user engagement and usability. The platform represents an ideal object of investigation as both the chatbot and the website-based SUI taps into the same database of psychiatry and neurology-related resources. Currently, the main users of this platform are mostly healthcare experts, but the providers' aim is to make the collection more accessible to the broader public. The chatbot was considered as an experimental tool to draw in more users by enhancing content interactivity and, subsequently, user engagement. We formulated the following research question:

How does a conversational search interface compare to a graphical search user interface in terms of user engagement and usability?

To investigate this question we conducted a controlled interactive information retrieval experiment (IIR). In this experiment it is hypothesized that the chatbot

will achieve its goal, i.e. it will successfully enhance user engagement. Therefore, we formulated the following hypothesis:

H_1: *The usage of the chatbot for searching has a positive effect on user engagement.*

User engagement is measured using the User Engagement Scale (UES), a defacto standard in the field of IIR [21]. While the main focus of the study will be to compare the overall engagement of users across the two interfaces, one aspect of user engagement will be discussed in greater detail: system usability. This is done by (1) collecting quantitative measures on time on task, task success and overall preference and (2) conducting a thematic analysis on the qualitative data shared by the participants during the experiment and in a post-study questionnaire.

To sum up, the contributions of our work are as follows:

- We present one of few studies that compare a conversational chatbot interface to a graphical SUI.
- We present a detailed analysis of how user engagement and usability issues are related.
- We reveal usability issues of the chatbot, link these issues to design patterns and give recommendations for generic chatbot design.

Before we present our experimental design, section 2 reviews relevant related work.

2 Related Work

Conversational search is still a novel branch of IR and HCI, but it is becoming more popular due to the increased acceptance of voice-based intelligent personal assistants (IPA) by the general public [24]. However, as mentioned before, users may have difficulty adapting to conversational search [27], since the majority of search interfaces are based on a graphical user interface. Graphical SUIs set the standards for digital information search, and the majority of IR system design principles are based on graphical representation – e.g. faceted search [34] or SERP control features like sorting, filtering or grouping [39].

While interaction with voice-based interfaces has received much attention [27,24,33], chatbot interaction has been less well studied. Most importantly, there is a lack of studies that compares classic SUI-based information access to dialog-based chatbot interaction. Literature about this field is scarce, and most comparative studies do not focus specifically on search systems. For example, Ischen et al. compared a website, a human-like chatbot, and a machine-like chatbot and studied the effects of the interface on anthropomorphism and privacy concerns via questionnaires [15]. One of their findings was that the website evoked more privacy concerns in users than the machine-like chatbot, which lead to less information disclosure (interestingly, no such difference was found between the human-like chatbot and the website). Celino and Re Calegari [5] investigated whether administering surveys via a conversational interface is a reliable and user-friendly method for data gathering. They tested a website-based survey, a

chatbot with informally formulated questions, and one with formally formulated questions via A/B testing and collected preference data via questionnaires. They found out that users have a preference towards the chatbot-administered survey, and that a chatbot-based method is at least as reliable in terms of inter-rater reliability as the website-based one. The work by Sundar et al. [30] is the only study of this type that focused on an interface that is used for searching. They compared several types of interfaces for a movie search website with varying levels of message interactivity, which they manipulated by adding/removing search history functionalities and a chatbot for assisting users in their browsing. They found that providing interaction history and the possibility of chatting with a live agent significantly increased perceived contingency, and subsequently, interactivity, which affected user engagement positively. Apart from the latter, no literature was found that compares the performance of conversational and graphical search user interfaces – therefore, the focus will be on conversational search interfaces in general.

Vtyurina et al. explored users' preferences towards conversational search interfaces of various sentience [36]. Participants completed exploratory search tasks with three types of chatbots: a commercial chatbot, a human expert (where participants knew they interacted with a human), and a "wizard" where the chatbot was covertly operated by a human but participants thought they interact with a machine. They found that most users preferred the human or "wizard" chatbots as both were able to interpret half-sentences, whereas the machine struggled with reference resolution, which also negatively affected participants' search task performance. Dubiel et al. found similar differences in task performance and user satisfaction in a Wizard-Of-Oz-style study. They explore two hypothetical spoken dialog systems: a standard voice bot using a slot-filling algorithm and an intelligent "conversational search agent" with a memory component for handling contextuality [9]. Participants were significantly more successful with their tasks when they used the agent with a memory component, and they found it less taxing and displayed a more positive sentiment towards it compared to the slot-filling agent. This points towards the users' need for more human-like conversations where chatbots have contextual awareness – preferably without asking too many questions for confirmation [9]. However, user expectations about the capabilities of conversational interfaces are usually met with disappointment. Luger and Sellen conducted a qualitative study using interviews and thematic network analysis to explore the mental models that users have about their voice assistants [19]. They found a "deep gulf of evaluation": users reported their confusion about the capabilities of the voice systems, as their expectations were not met. The in-built playful responses (e.g. the capability of telling jokes) also set unrealistic expectations about the sophistication of the system, and after continued disappointment, users became reluctant to use their voice assistants for complex tasks.

Seeing that the discipline of conversational search still lacks profound research, Thomas et al. collected a rich dataset of search-oriented conversations called MISC (Microsoft Information-Seeking Conversation data) [32]. The par-

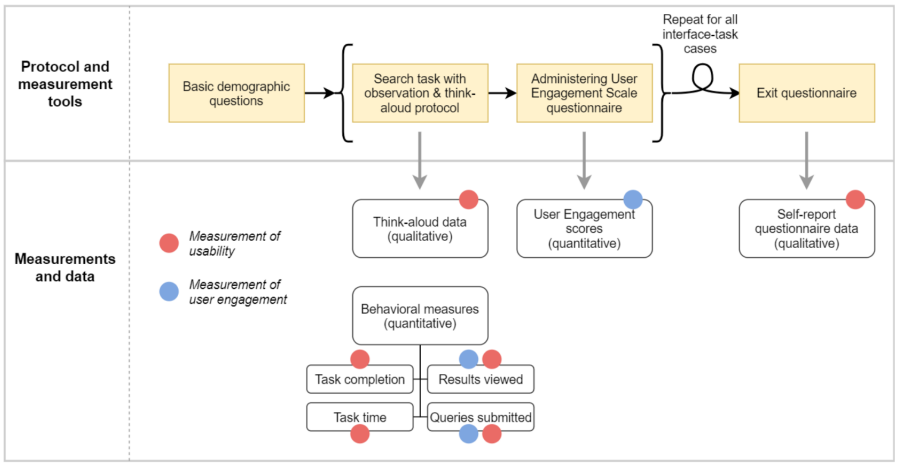

Fig. 1. Experimental Setup: the measurement tools used and the type of data collected

ticipants of the conversations consisted of a searcher, who was given a search task, and an intermediary who had access to the internet and was tasked to follow the searcher's directions and provide feedback only via voice. These conversation recordings are created to help to establish requirements for an optimal conversational search system and demonstrate users' desires for an aligned discourse with conversational interfaces. Alignment means that the user and the system can match each other's style of communication in terms of involvement (chit-chattiness, verbosity, enthusiasm) or considerateness (more listening, hesitance, independence). If alignment succeeds, then task execution becomes more efficient [31].

3 Experimental Design & Experiments

We conducted a controlled IIR experiment to investigate whether the type of interface used for searching, the independent variable, influences user engagement, the dependent variable. Figure 1 visualizes the experimental setup and shows how and what kind of data got collected during the study. For participants to be able to compare the two systems, we followed a within-subjects design. The two interfaces were compared through a series of tasks that the participant had to complete with the interfaces.

Object of Investigation and Stimuli Our object of investigation is the medical resource center website *Progress in Mind*. *Progress in Mind* is an online, open-access database, where articles and videos about current scientific trends, international news, and congress highlights are hosted. The publications on the website are written and curated by medical writers in a generally informal style,

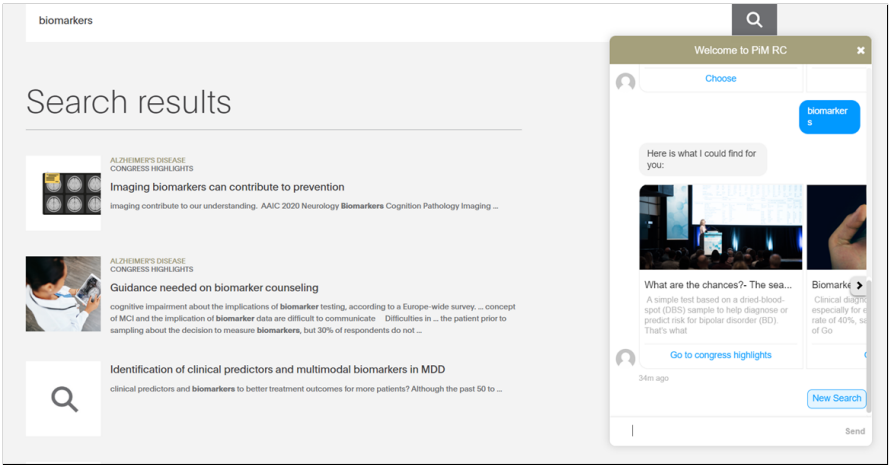

Fig. 2. Screenshot of the Progress in Mind platform's graphic SUI on the left side and its chatbot SUI on the right side

and the content is aimed at healthcare practitioners and academics in the field. Users can filter content by diseases or types of publications or use free-text queries to search across the database.

The company's goal is to transform the platform into a go-to resource center for healthcare professionals of psychology and neurology. Therefore, they are experimenting with new ways to make the content more accessible and interactive – which led to the development of a chatbot. This chatbot interface is an auxiliary tool for the website search and uses a conversational modality to help users search the database, presenting search results in a chat window (Figure 2). This conversational style is aimed to improve interactivity, which, as Lundbeck anticipates, will lead to greater user engagement and promote the usage of the platform. As Sundar et al. [30] have shown, delivering online content in a dialog-like manner can lead to improved interactivity and, in turn, a greater level of engagement.

To gather an adequate amount of information from users, each user interacted with an interface twice, completing two tasks with each interface – therefore, a total of four different tasks were defined which got randomized across the two interfaces for every participant. This is to account for learning effects, as users might initially focus on getting to know the system and concentrate less on the search task.

Task design During each task, the basic goal of the user was to list three diseases that have a connection to the topic of the given task. Tasks were of low-intermediate cognitive complexity, corresponding to the "Understand" category following Kelly et al.'s task classification [17]. These tasks "require the searcher to provide an exhaustive list of items" by identifying "a list or factors in an

information source and possibly compile the list from multiple sources if a single list cannot be found" [17].

The topics of the four tasks were: sleep disturbance, cognitive impairment, biomarkers and mobile health. Tasks have been formulated as "simulated work tasks" according to Borlund [3] the following way: "You have a friend who needs help with a school project where he needs to explore [topic]. He asks you to send him some easy-to-understand material about the topic, so you decide to use the Progress in Mind platform to search for resources. Use the [search interface] to search for publications and find at least 3 diseases that may be linked to [topic]/where [topic] can be applied. When you read a publication, please also decide whether or not you would send it to your friend to help him in his project.".

Measurements and data collection Engagement was measured using the User Engagement Scale - Short Form (UES-SF), developed by O'Brien et. al [21]. It contains 12 items, grouped into four categories: Focused Attention (FA), Perceived Usability (PU), Aesthetic Appeal (AE), and Reward (RW). The participants had to fill out the UES-SF form each time after a task was completed.

Usability was broken up into three constituents according to the ISO 9241-11:2018 standard [14], and measured using various behavioral measures:

- Task time (how much time the participant spent on a task) that measures efficiency;
- Task success (whether the participant successfully completed the task or not) that measures effectiveness; and
- Preference (which interface was preferred by the participant for the given task) that measures satisfaction.

Moreover, we collected search behavior related measures: the number of viewed results and the number of submitted queries. Differences in search behavior across the two interfaces were assessed, as search behavior can have an effect on user engagement [22]. Finally, data about the general user experience was gathered via qualitative think-aloud comments recorded during the experiment and an exit questionnaire, where participants were asked to list their most positive and negative experiences during their interaction.

4 Results

A total of 10 participants were recruited through snowball sampling, 8 female and 2 male, all had Hungarian as their native language, but were fluent in English. Their age ranged from 22 to 32 with a mean age of 24,5 years. The participants were highly educated as each participant had at least an undergraduate degree. Almost all participants reported that they never used chatbots or only once or twice in their life; one participant used chatbots more than once a month for booking flights and as online shopping assistance. During the exit questionnaire,

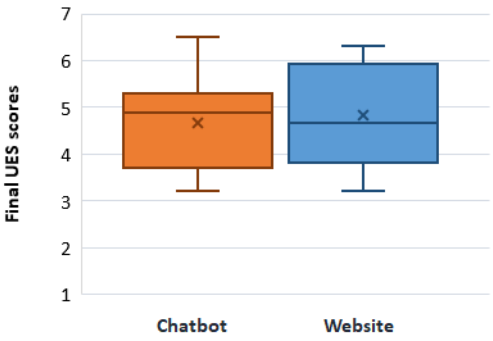

Fig. 3. Distribution of the average UES scores for the chatbot and the website SUI. The 'x' in the boxplot denotes the mean.

9 participants reported that they would use the website for searching across the collection, and only one participant said that she/he would prefer to use the chatbot. On average, an entire experiment was 60 minutes long with a minimum length of 39 minutes and a maximum of 73 minutes.

4.1 User engagement

The first part of the analysis focused on understanding how the independent variable, the search interface, influenced user engagement, the dependent variable. According to Figure 3, most UES scores lie in the upper half of the scale, which suggests that the majority of participants were engaged throughout the study. Comparing the average UES scores of the chatbot ($M = 4.65$, $SD = 1.05$) and the website's ($M = 4.83$, $SD = 1.12$) via a Student's t-test, we did not find a statistically significant difference between the two interfaces ($t(9) = -0.4$,

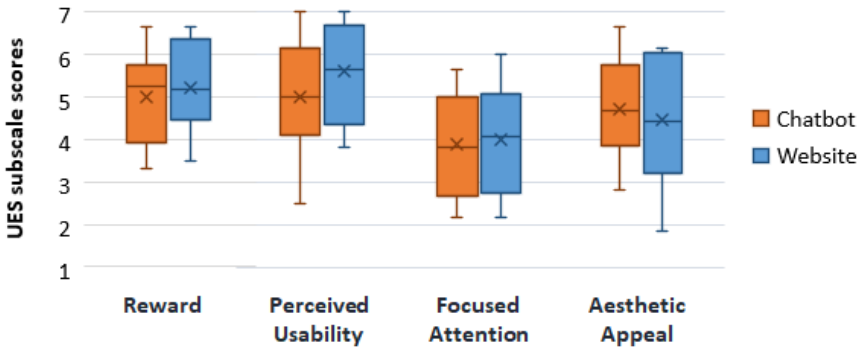

Fig. 4. Distribution of the average UES scores for the chatbot and the website SUI.

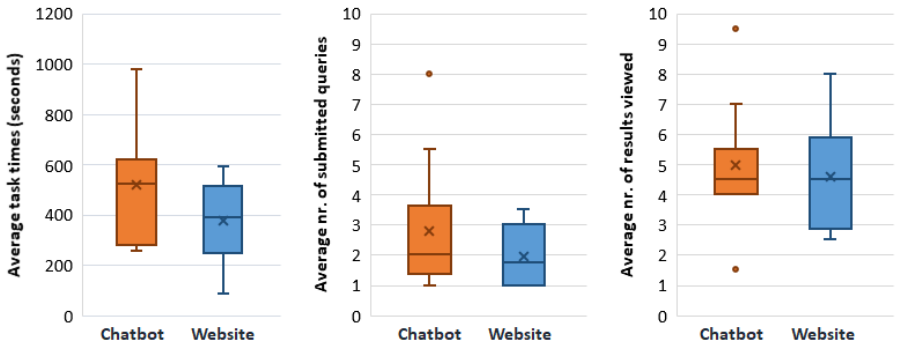

Fig. 5. Summary of behavioral measures per interface: task time (left), number of queries submitted (middle), and number of results viewed (right).

$p = 0.69$). Moreover, the mean UES score of the chatbot was slightly lower than the website's. As we did not find substantial evidence to say that the usage of the chatbot results in higher user engagement, we can not reject the null hypothesis (H_0).

In order to investigate whether the type of interface influenced any specific aspect of engagement, the UES scores have been broken down to subscales. Figure 4 shows the mean subscale scores per interface. No significant difference was found between the two interfaces in terms of subscale scores. The website SUI outperformed the chatbot interface in all but one aspect: Aesthetic Appeal (AE), where the chatbot got a 0.2 points higher score. The largest difference between the two interfaces can be observed in Perceived Usability (PU), where the website ($M = 5.6$, $SD = 1.19$) outperformed the chatbot ($M = 5$, $SD = 1.34$) by more than half a point. Both interfaces received a relatively low score for focused attention (FA). Reward (RW) received the second-highest scores after PU, with only a slight difference between the two interfaces.

4.2 Behavioral measures

Behavioral measures were analyzed (Figure 5) to see how the two interfaces compare in terms of task performance. On average, participants took more time (approximately 2 more minutes) to complete the tasks with the chatbot compared to the website, which indicates a lower efficiency within the chatbot. Regarding task success, there were three instances where the user was not able to successfully complete the task, each time while using the chatbot. In these cases, time until task abandonment was measured instead of time to completion. Users submitted on average almost 1.5 times more queries when using the chatbot (see Figure 5 left). The number of viewed results, shown in Figure 5 right, was approximately equal across the two interfaces, with chatbot users viewing slightly more results than website users. No statistically significant difference was found between the two interfaces in terms of any behavioral measures.

Table 1. Pearson-product moment correlations between behavioral metrics and the mean UES subscale scores and the mean UES score per interface.

	Chatbot			Website		
	Average task time	Avg. number of queries submitted	Avg. number of results viewed	Average task time	Avg. number of queries submitted	Avg. number of results viewed
RW	-.047	.002	-.529	-.603	-.333	-.717*
PU	-.411	-.323	-.571	-.426	-.248	-.346
FA	-.252	-.156	-.658*	-.438	-.035	-.546
AE	.158	.292	-.218	-.536	-.277	-.384
Final UES	-.178	-.061	-.571	-.555	-.231	-.551

Symbol * denotes significant r values ($p < .05$). Coloring represent the strength of the correlation, where white color represents very weak correlation ($|r| < .19$), and the darkening colors represents weak ($.2 < |r| < .39$), moderate ($.4 < |r| < .59$), and strong ($|r| > .6$) correlation.

In order to see whether there is any correlation between behavioral measures and user engagement, the Pearson-product moment correlations have been calculated between the behavioral measures and the UES scores (Table 1). Correlations are almost exclusively negative, apart from three cases, which means that task time, number of submitted queries and number of results viewed elicit lower UES scores. Significant correlations have been found between website RW score and number of results viewed, and chatbot FA score and the number of results viewed. Behavioral metrics have overall stronger correlations with UES scores in the case of the website, and weaker correlations in the case of the chatbot.

4.3 Thematic analysis

This section details the themes that emerged from the thematic analysis, a qualitative data analysis method. Our approach for conducting the thematic analysis followed the recommendations by Braun and Clarke [4]. The analysis was conducted in a deductive manner, meaning that it was built around two encompassing themes that dictated which participant remarks can be considered relevant: usability and search behavior. Usability strongly ties to the research question of the study, while remarks about search behavior can both highlight differences in engagement [22] or explain usability-related issues. The think-aloud protocol and the post-study questionnaire were selectively transcribed – only quotes pertaining to the preliminary themes were written up, not the entirety of the interview. A researcher scrutinized the transcripts to find participant quotes pertaining to usability issues (or good usability practices) and quotes that describe or explain search behaviors. Quotes were codified as short sentences, and similar codes

were collated into sub-themes for easier overviewability. Table 2 summarizes the themes, sub-themes and codes that emerged from the qualitative data.

Table 2. Thematic analysis table

Theme	Sub-theme	Code
Usability	User interface	Overlapping UI elements
		Interface aesthetics (visuals and sound)
		Ambiguity of UI elements
	Inconsistencies within the system	Newsletter among the results
		Website SUI's inconsistency with thumbnail images
	Chatbot utility	Uncertainty about when to use the chatbot
		Chatbot fails to handle complex input
		Doubtful disposition towards chatbots
	Importance of overviewable results	Viewing more results at once is important
		Chatbot window is too small
	-	Importance of response speed
Search behavior	Assessing relevance of documents	Relying on field knowledge
		Assessing results and content using metadata and keywords (within SERP and within content)
	Partitioning and query tactics	Filtering and faceting features are missing (in the SUIs and in the homepage)
		Phrase search is useful for experienced searchers
		Preferring keyword-search in chatbot
	-	Finding related items is difficult
	-	Lack of search transparency within the system
	-	Searching for explanations in case of unfamiliar topics

Usability The majority of remarks about usability concerned the visual structure of the user interface, highlighting the differences between the SUI of the website and the chatbot. The main issue seemed to be the relatively small size of the chatbot window, which caused all navigational elements to be placed closely together. One user remarked how "the [chat] window was too small, and

you had to click this small right-arrow which was annoying", reflecting on the difficulty to navigate between results.

Apart from navigational problems, the small chatbot window also hindered the relevance assessment of results. Almost all users noted the inadequacy of the chatbot to display several results at once, which, as one participant noted, "was weird because I couldn't have as much of an overview". In contrast, "the website was better because it showed the results below one another and I could overview them more easily".

Another negative aspect of the chatbot was the slow response time, as "that 3 seconds waiting was strange. It took some time to react". In contrast, the website was "faster [compared to the chatbot] and I didn't have to wait for an answer", as one participant remarked.

Users had ambiguous feelings towards the visual style of the platform, as some pointed out that it "was quite dull and colorless". Interestingly, one participant remarked positively about the poor visual appeal of the website, as "for some reason, it's important to me that if something [deals with] scientific [material], it should look a bit lame. This website looked trustworthy for me [...] because of this". Apart from visual aesthetics, the sound the chatbot made every time it sent a message was found to be "weird and annoying" by two participants, whereas one participant expressed their fondness of it.

Participants also made comments about the chatbot's utility. Users who tried to communicate with the chatbot with complete sentences realized that the chatbot is not capable of handling complex inputs, therefore all resorted to keyword inputs (discussed in the next theme).

Some participants were skeptical about the chatbot's overall usefulness and articulated that they "would not even think about using the chatbot for searching". One reason behind the doubtful disposition was the lack of faith in conversational technology: one participant shared his/her negative experiences with voice assistants "which made it clear that I don't want to use them again". Another participant reflected that "using [chatbots] only makes sense if you're talking to a real human".

Search behavior Throughout their search, users demonstrated various tactics to choose which results to click and to assess whether the content they are reading is relevant for them or not. Most of the participants could be observed scanning the metadata in the search snippets for relevant keywords using either the title, topical tags, or query highlights – which was a fairly limited tactic in the chatbot. Users were missing the rich metadata from the chatbot snippets, because in the website SUI "you could see above the titles the diseases the article is about [...] whereas the chatbot does not display these keywords".

Apart from relying on keywords, another tactic of assessing relevance was relying on their own knowledge. Users would collate the information they read with their own knowledge to assess the relevance of the result, e.g. "this one shows major depressive disorder, which makes sense as that [and cognitive impairment] go hand in hand". Some participants expressed a certain level of confusion when

they met with information which seemingly contradicted their former knowledge, with comments like "based on what the article states, I would not connect it [to the disease] but I know [by myself] that they are related".

More than half of the subjects expressed their frustration that the system does not provide adequate search functionality. The biggest issue was the lack of filtering and faceting options as there are "no options in the search bar to filter, like which source is it from, [or] when was it published". One participant remarked positively that the chatbot provides at least some level of categorical browsing, saying that "I really liked at the beginning that it asked whether I want to read articles or listen to podcast".

In terms of query formulations, users almost exclusively resorted to keyword-based search, usually using the task topic as the query (e.g. "cognitive impairment") – even within the chatbot, despite its conversational interaction. One reason behind this could be that, as one participant stated, "the chatbot phrased the question in a way that it didn't even occur to me to reply in full sentences". The chatbot phrased its welcome message as "type what you are looking for", which the users might have interpreted as a prompt for a keyword or search phrase. Nevertheless, users liked that "it was enough to write keywords and you got all types of publications".

Three participants raised the issue of search transparency – saying that "it wasn't really clear to me how [the chatbot] selects those articles". One participant even remarked about this distrust, saying that "I had a bit of distrust in me about whether [the platform] actually shows me the relevant results". This issue was even more relevant in the chatbot, where only a limited number of results were displayed. Users "didn't really know how to expand the number of results", and one participant mentioned that they were curious how those articles were selected, as they "couldn't really see any pattern in it".

5 Discussion and Conclusion

Based on our research question, we discuss our findings in two section: (1) How does the type of interface influence user engagement and (2) What role did usability play during the experiment? The identified usability problems can also be interpreted as design recommendations as they represent issues in human-chatbot interaction that should be avoided, especially in the context of IR tasks.

5.1 How does the type of interface influence engagement?

The analysis revealed that using the chatbot for searching does not lead to greater engagement – the null hypothesis could not be rejected. In fact, the chatbot underperformed in all but one aspect of engagement, aesthetic appeal (AE). According to the thematic analysis, the aesthetics of the interface seems to be of a subjective matter, as a stylistic choice can elicit both negative and positive reactions from participants. Therefore, the reason behind the higher AE score may be attributed simply to the novelty of the interface – the chatbot might

have grabbed the users' visual attention because of its unique way of searching, which might have resulted in an initial interest and a more favorable AE score. Still, the attractiveness of the interface was not enough to counterbalance the other aspects the interface was lacking – especially perceived usability (PU), which is going to be discussed separately.

Interestingly, both interfaces received a relatively low score for focused attention (FA), which suggests that neither interface managed to hold the attention of the participants to such an extent which could have led to deep involvement. The reason behind the low scores could be that the protocol of the experiment gave little room for substantial immersion: the Understand type tasks we used did not require high-level cognitive processing, only identifying and compiling information [17]. The online format of the experiment might have also played a role, as participants' focus could be easily disrupted by their external environment – which prevented them from being absorbed in the experience. The chatbot's slightly lower FA score could be due to its slow response time, which participants occasionally commented about. Participants might find it self-evident that search systems are generally quick to respond (like Google), thus the chatbot with such a response delay (approximately 2 seconds) may seem sluggish and it can interrupt the user's flow of thoughts [20].

Reward (RW) received the second-highest average score among the subscales with only a small score difference between the two interfaces, which indicates that participants usually found their search experience interesting, worthwhile, and rewarding – regardless of the platform. This highlights the importance of the content, which – interface-independently – enhanced the reward factor of using the platform. Observations also reinforce this assumption, as many participants made sporadic comments about the platform's interesting content (e.g. "the content is [extremely] good... the articles were great and contained relevant information").

Regarding the behavioral measures, participants were less efficient in their tasks when using the chatbot, with higher task times, more queries sent, and more results clicked. The almost exclusively negative correlations between the UES scores and the behavioral measures also show that a higher "interaction cost" leads to lower engagement (see Table 1). This is in accordance with O'Brien, Arguello and Capras' [22] results, who found that a higher task effort correlates negatively with engagement. Edwards and Kelly [10] also found that increased search behaviors signify frustration, rather than engagement. The number of viewed results seems to be a good indicator of low engagement. Participants who clicked on a large number of results might have experienced impatience and frustration, which led to lower engagement. In the case of the website, the time that users spend on a task seems to be a good indicator of low engagement, with correlations ranging from moderate to high. However, in the case of the chatbot, only PU had a moderate negative correlation with task time. Sauro and Lewis [28] draw similar conclusions. They interpret higher task times as an indicator of poor usability. The number of query submissions was not shown to be a good indicator of engagement as correlations were either weak or very weak.

5.2 What role does usability play in engagement?

Quantitative data did not show any significant differences between behavioral measures. Nevertheless, (1) the chatbot elicited higher task times (related to efficiency), (2) the preference data shows a higher satisfaction in the case of the website, and (3) task success (related to effectiveness) also shows that users were slightly more successful in completing their tasks with the website. Considering all three aspects of usability, the website SUI performed better compared to the chatbot. The thematic analysis also revealed that participants found the chatbot less user-friendly than the website SUI. The greatest problem of the chatbot is the limited amount of information it displays due to its small size. Though a horizontally scrollable result list needs less effort to navigate through compared to vertical scrolling, in exchange of displaying the results in a compact area the possibility for an overview is greatly impaired – and since this issue was mentioned by almost all users, it might be the greatest contributor to the reduced PU score of the system. The lack of overview ties closely to Shneiderman's Visual Information Seeking Mantra, which stipulates that a system must first provide the user a proper overview of the collection, before zooming in on items of interest and providing details on demand [29]. The chatbot violated this mantra as users could only see one result at a time.

Result presentation in the chatbot also omitted certain metadata, which made assessing their relevance even more difficult. The lack of metadata and sorting functions also made users question how the chatbot chose which results to display. Jackson et al. [16] raised this issue of lack of search transparency, stating that a search system should provide information according to which criteria the results are ranked, otherwise users become "instinctively distrustful of any mechanism they don't understand" [16].

Lack of filtering and faceting also impaired search efficiency for both interfaces. Although the chatbot does provide faceted browsing to some extent, accessing it is not straightforward, and none of the users managed to figure out how to search within facets. Topical tags are also accessible for each article, but they are not integrated enough in the search system and not salient enough so that users could find them easily. The possibility of issuing phrase-search or using search operators was also not communicated effectively. The system seems to provide more utility to experienced searchers who are already familiar with the platform and who can leverage the system's less visible functionalities (e.g. search operators or tags).

Further indicators of the chatbot's poor usability are the higher task times [28], and the lower PU score of the chatbot (see Figure 4). The higher number of submitted queries and viewed results may also indicate lower search efficiency, as participants had a harder time finding relevant results with the chatbot. However, it must be noted that the chatbot displays only a limited number of items on the SERP, thus users had to submit further queries if they wanted to see more results. This could be another reason behind the large difference in the number of query submissions. It must also be noted that, since the chatbot omits certain metadata and thus makes relevance assessment difficult, users might have

been more inclined to open the result and check the content itself to determine its relevance – hence the higher number of viewed results. Nevertheless, the behavioral measures also show that the chatbot required greater effort from the participants, which translates into poor usability.

The almost exclusive preference for the website SUI shows that users can hardly recognize any value the chatbot could add to their search process. For example, a participant commented that "a chatbot can create added value where social interaction with a human needs to be substituted...and searching is not a social interaction". This signifies that the chatbot performs poorly not only in terms of usability but also utility. However, this stands in contrast to the chatbot's fairly high UES score ratings. Possible explanations for this observation could be biases like social desirability bias which lead participants to rate the chatbot higher in the UES questionnaires. Our study is limited in the sense that the chatbot has serious usability issues, which made our main hypothesis — whether a chatbot could create significantly more user engagement than a more traditional SUI — difficult to assess. However, it has become clear that for chatbots to be able to successfully facilitate user engagement and be a real alternative in information retrieval tasks, flawless usability is an essential quality.

References

1. Allison, D.: Chatbots in the library: Is it time? Library Hi Tech **30**(1), 95–107 (2012)
2. Bates, M.J.: The design of browsing and berrypicking techniques for the online search interface. Online Review **13**(5), 407–424 (1989), `http://www.emeraldinsight.com/doi/10.1108/eb024320`
3. Borlund, Pia: The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information Research **8**(3), 1–31 (2003), `http://informationr.net/ir/8-3/paper152.html`
4. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa
5. Celino, I., ReCalegari, G.: Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. International Journal of Human Computer Studies **139**, 1–16 (2020)
6. Chung, M., Ko, E., Joung, H., Kim, S.J.: Chatbot e-service and customer satisfaction regarding luxury brands. Journal of Business Research pp. 1–9 (2018)
7. Dale, R.: The return of the chatbots. Natural Language Engineering **22**(5), 811–817 (2016)
8. Denecke, K., Hochreutener, S.L., Pöpel, A., May, R.: Self-Anamnesis with a Conversational User Interface: Concept and Usability Study. Methods of Information in Medicine **57**(5-6), 243–252 (2018)
9. Dubiel, M.: Towards human-like conversational search systems. In: Proceedings of the 2018 Conference on Human Information Interaction and Retrieval. p. 348–350. CHIIR '18, Association for Computing Machinery, New York, NY, USA (2018)
10. Edwards, A., Kelly, D.: Engaged or frustrated? disambiguating emotional state in search. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 125–134. SIGIR '17, Association for Computing Machinery, New York, NY, USA (2017)

11. Exalto, M., De Jong, M., De Koning, T., Groothuis, A., Ravesteijn, P.: Conversational commerce, the conversation of tomorrow. In: Proceedings of the 14th European Conference on Management, Leadership and Governance, ECMLG 2018. pp. 76–83 (2018)
12. Graesser, A.C., Li, H., Forsyth, C.: Learning by Communicating in Natural Language With Conversational Agents. Current Directions in Psychological Science **23**(5), 374–380 (2014)
13. Gratzer, D., Goldbloom, D.: Open for Business: Chatbots, E-therapies, and the Future of Psychiatry. Canadian Journal of Psychiatry **64**(7), 453–455 (2019)
14. International Organization for Standardization: Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts (2018)
15. Ischen, C., Araujo, T., Voorveld, H., van Noort, G., Smit, E.: Privacy Concerns in Chatbot Interactions. In: Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.C., Granmo, O.C., Luger, E., Brandtzaeg, P.B. (eds.) Chatbot Research and Design. CONVERSATIONS 2019., chap. 7, pp. 34–48. Lecture Notes in Computer Science, Springer International Publishing, Amsterdam, Netherlands (2020)
16. Jackson, A., Lin, J., Milligan, I., Ruest, N.: Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16. pp. 103–106. ACM Press, New York, New York, USA (2016). https://doi.org/10.1145/2910896.2910912, http://dl.acm.org/citation.cfm?doid=2910896.2910912
17. Kelly, D., Arguello, J., Edwards, A., Wu, W.C.: Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval pp. 101–110 (2015)
18. Lee, K., Jo, J., Kim, J., Kang, Y.: Can Chatbots Help Reduce the Workload of Administrative Officers? - Implementing and Deploying FAQ Chatbot Service in a University. In: Stephanidis, C. (ed.) HCII: International Conference on Human-Computer Interaction 2019. pp. 348–354. Springer Nature, Orlando (2019)
19. Luger, E., Sellen, A.: "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. pp. 5286—-5297. CHI '16, Association for Computing Machinery, New York, NY, USA (2016)
20. Nielsen, J.: Usability Testing. In: Usability Engineering, chap. 6, pp. 165–206. Morgan Kaufmann Publishers, Mountain View, California (1993)
21. O'Brien, H., Cairns, P., Hall, M.: A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. International Journal of Human Computer Studies **112**(December 2017), 28–39 (2018)
22. O'Brien, H.L., Arguello, J., Capra, R.: An empirical study of interest, task complexity, and search behaviour on user engagement. Information Processing & Management **57**(3), 102226 (2020)
23. Perski, O., Crane, D., Beard, E., Brown, J.: Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digital Health **5**, 1–13 (2019)
24. Porcheron, M., Fischer, J.E., Reeves, S., Sharples, S.: Voice interfaces in everyday life. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. p. 1–12. CHI '18, Association for Computing Machinery, New York, NY, USA (2018)

25. Radlinski, F., Craswell, N.: A theoretical framework for conversational search. CHIIR 2017 - Proceedings of the 2017 Conference Human Information Interaction and Retrieval pp. 117–126 (2017)
26. Reed, K., Meiselwitz, G.: Teacher agents: The current state, future trends, and many roles of intelligent agents in education. In: Lecture Notes in Computer Science. vol. 6778 LNCS, pp. 69–78. Springer Berlin Heidelberg (2011)
27. Reeves, S., Porcheron, M., Fischer, J.: ŕthis is not what we wanted.ˀ Designing for conversation with voice interfaces. Interactions **26**(1), 46–51 (2018)
28. Sauro, J., Lewis, J.R.: Correlations among prototypical usability metrics: Evidence for the construct of usability. In: Greenberg, S., Hudson, S.E., Hinckley, K., Morris, M.R., Olsen, D.R. (eds.) CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1609–1618. ACM Press, Boston (2009). https://doi.org/10.1145/1518701.1518947
29. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. The Craft of Information Visualization pp. 364–371 (2007)
30. Sundar, S.S., Bellur, S., Oh, J., Jia, H., Kim, H.S.: Theoretical Importance of Contingency in Human-Computer Interaction: Effects of Message Interactivity on User Engagement. Communication Research **43**(5), 595–625 (2016)
31. Thomas, P., Czerwinski, M., McDuf, D., Craswell, N., Mark, G.: Style and alignment in information-seeking conversation. In: CHIIR 2018 - Proceedings of the 2018 Conference on Human Information Interaction and Retrieval. pp. 42–51. ACM Press, New York (2018)
32. Thomas, P., McDuff, D., Czerwinski, M., Craswell, N.: Misc: A data set of information-seeking conversations. In: Proceedings of the 1st International Workshop on Conversational Approaches to Information Retrieval (2017)
33. Trippas, J.R., Spina, D., Cavedon, L., Joho, H., Sanderson, M.: Informing the design of spoken conversational search: Perspective paper. In: Proceedings of the 2018 Conference on Human Information Interaction and Retrieval. p. 32–41. CHIIR '18, Association for Computing Machinery, New York, NY, USA (2018)
34. Tunkelang, D., Marchionini, G.: Front-End Concerns. In: Marchionini, G. (ed.) Faceted Search, chap. 7, pp. 57–68. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool (2009)
35. Vakulenko, S., Markov, I., de Rijke, M.: Conversational exploratory search via interactive storytelling. CoRR **abs/1709.05298** (2017), `http://arxiv.org/abs/1709.05298`
36. Vtyurina, A., Savenkov, D., Agichtein, E., Clarke, C.L.A.: Exploring conversational search with humans, assistants, and wizards. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. p. 2187–2193. CHI EA '17, Association for Computing Machinery, New York, NY, USA (2017)
37. Ward, D.: Why Users Choose Chat. Internet Reference Services Quarterly **10**(1), 29–46 (2005)
38. White, R.W.: Opportunities and challenges in search interaction. Communications of the ACM **61**(12), 36–38 (2018)
39. Wilson, M.L.: Modern Search User Interfaces. In: Marchionini, G. (ed.) Search User Interface Design, chap. 4, pp. 29–79. Morgan & Claypool (2011)
40. Zhu, P., Zhang, Z., Li, J., Huang, Y., Zhao, H.: Lingke: A Fine-grained Multi-turn Chatbot for Customer Service. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 108–112. Association for Computational Linguistics, Santa Fe, New Mexico (2018)