Exploring Conversational Search With Humans, Assistants, and Wizards

Alexandra Vtyurina
University of Waterloo
Ontario, Canada
avtyurin@uwaterloo.ca

Denis Savenkov
Emory University
Atlanta, GA, USA
denis.savenkov@emory.edu

Eugene Agichtein
Emory University
Atlanta, GA, USA
eugene.agichtein@emory.edu

Charles L. A. Clarke
University of Waterloo
Ontario, Canada
claclark@gmail.com

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3053175

Abstract
Chatbots and conversational assistants are becoming increasingly popular. However, for information seeking scenarios, these systems still have very limited conversational abilities, and primarily serve as proxies to existing web search engines. In this work, we ask: what would conversational search look like with a truly intelligent assistant? To begin answering this question empirically, we conduct a user study, in which 21 participants are each given 3 information seeking tasks to solve using a text-based chat interface. To complete each task, participants conversed with three conversational agents: an existing commercial system, a human expert, and a *perceived* experimental automatic system, backed by a human "wizard" behind the curtain. The observations and insights of our study help us understand the aspirations of users and the limitations of the current conversational agents – and to sharpen a frontier of work required to improve conversational assistants for search scenarios.

Author Keywords
Conversational search, intelligent assistants.

ACM Classification Keywords
I.2.1 [Applications and Expert Systems]: Natural language interfaces

Introduction.

Rapid progress in technology is changing the way we interact with the information [8]. Improvements in speech recognition and natural language processing have allowed people to build voice-controlled personal assistants, such as Apple's Siri, Amazon's Alexa and Google Home. These technologies are increasingly popular, and people are integrating them in everyday life, e.g., for simple tasks like setting up a timer, checking the calendar, requesting the latest news, a song, etc.[1]. The popularity of text-based chatbots is also on the rise in many areas of the web [7]. Most of them are template-based and are designed to fulfill a single, often monotonous, job [5, 6].

At the same time, a growing proportion of web search queries is formulated as natural language questions [12, 9, 2], which is partially explained by the increasing usage of voice interfaces [15].

Alas, for information seeking scenarios, existing chatbots and intelligent assistants are usually implemented as simply a "proxy" to existing web search engines, even though question-answering technology has made dramatic progress handling such question-like queries [14]. Furthermore, conversation provides additional opportunities to improve search quality. For example, a conversational system should be able to ask clarification questions [3] to better identify searcher's intent, and incorporate explicit user feedback [13] – something that is not normally available in a traditional web search scenario.

However, before jumping into implementing additional features for conversational search systems, it is important to gain a better understanding what the users' expectations are when interacting with a truly intelligent conversational search agent. It is equally important to anticipate how users might behave when faced with a conversational search system since behavioural feedback is critical for system evaluation and improvements. To this end, we explore the following research questions:

- **RQ1**: What are the main expectations from a conversational search system?
- **RQ2**: What are the differences between human-to-human and human-to-computer conversations?
- **RQ3**: What characteristics prevent existing conversational agents from becoming effective tools for complex information seeking?

As no truly intelligent conversational search systems exist yet, we explore these research questions with a mixture of survey methods and user studies. In the user study, the participants are faced with 3 complex information search tasks, derived from TREC Session track tasks [4]. To eliminate the voice recognition quality variable, we chose to use text messaging as the interface between a participant and conversational systems. We use three different conversational systems answering user requests: an existing commercial intelligent assistant, a human expert and a human disguised as an automatic system.

The results of our exploration suggest: (1) people do not have biases against automatic conversational systems, as long as their performance is acceptable; (2) existing conversational assistants are not yet up to task, i.e., they cannot be effectively used for complex information search tasks; (3) by addressing a few requests from users that we identified, even current search systems might be able to improve their effectiveness and usability, with feasible modifications.

2. Related work

The topic of chatbots and conversational answer seeking has recently become quite popular. Radlinski and Craswell [13]

[1] https://arc.applause.com/2016/09/26/amazon-echo-alexa-use-cases/

Topic 10:
Suppose you are writing an essay about a tax on "junk food". In your essay, you need to argue whether it's a good idea for a government to tax junk food and high-calorie snacks.

Topic 20:
You have decided that you want to reduce the use of air conditioning in your house. You've thought that if you could protect the roof being overly hot due to sun exposure, you could keep the house temperature low without the excessive use of air conditioning.

Topic 21:
Hydropower is considered one of the renewable sources of energy that could replace fossil fuels. Find information about the efficiency of hydropower, the technology behind it and any consequences building hydroelectric dams could have on the environment.

Figure 1: Description of the tasks used in the study. All the tasks were obtained from TREC Session track 2014 [4].

defined a set of required properties and designed a theoretical model of interactions in conversational search. Our user study complements this work by providing an analysis of real user dialogs. Braslavski et al. [3] studied dialogues on StackExchange community question answering website and analyzed clarification questions. The most closely related work was done by Luger and Sellen [10], where 14 people were interviewed about their experience with an intelligent assistant that they use in their daily life. The authors report on people's experiences, expectations, discuss scenarios of successes and failures of conversational agents. They report that the most frequent types of tasks are relatively simple – weather updates and checking reminders. Our study, on the other hand, focuses on studying similar aspects of user behaviour for a different type of task – for complex search tasks. However, some of their findings overlap with ours.

Much work has been done in the area of comparing user interactions with a human and a computer. There are varying opinions on the subject. Edwards et al. [6] found no significant differences in how Twitter users treated a social bot, whether it was perceived as a human or not. In turn, Clément and Guitton [5] report that the way bots are perceived varies with the role they play. They found that "invasive" Wikipedia bots received more "polarizing" feedback – both positive and negative – compared to the bots that carried out "silent helper" functions. The similar result is reported by Murgia et al. [11] – Stackoverflow bot receives more negative feedback for false answers when its identity as an automatic program is revealed. Another work by Aharoni and Fridlund [1] reports mixed results from participants who underwent a mock interview with a human and an automatic system. The authors report that there were no explicit differences in the interviewer perception described by the participants, although the authors noticed significant differences in people's behaviour – when talking to a human interviewer they made greater effort to speak, smiled more, and were more affected by a rejection.

3. Study design.

We recruited 21 participants (graduate and undergraduate students at a major university), to complete 3 different complex search tasks, taken from the TREC Session track 2014 [4] (shown in Figure 1 in the sidebar). The participants were asked to use an assigned text messenger-based conversational agent. They were not given any instructions on how to use the agent and therefore were free to interact with it in any way they chose. They were allowed to spend up to 10 minutes working on each task, after which they were asked to move on a topical quiz, consisting of 3 questions, designed for the topic. After seeing the topical quiz questions, the participants were not allowed to talk to the agent anymore. By doing so we ensured that the task stayed exploratory in nature, i.e., the participants did not have a set of predefined points to cover. After completing a topical quiz, the participants filled out a preference questionnaire, where they were asked to rate their experience with the agent, provide feedback about advantages and disadvantages of the agent. After completing all tasks they filled out a final questionnaire. The communication was implemented through the Facebook Messenger interface[2]. Participants used a Facebook account created specifically for the purpose of the study. Message history was cleared prior to every experiment.

Wizard agent.

Our first research question explores human behaviour in human-computer communication. There are currently no general purpose intelligent conversational search systems, that we could use for our purposes. Therefore we "faked" one by substituting the backend with a person (two of the

[2] www.messenger.com

authors interchangeably). However, the participants were told that it was an experimental *automatic* system, thus following a general Wizard-of-Oz setup. We will be further referring to this system as the Wizard agent, and the person in the backend as the Wizard. The Wizard had previously done the research about the topics of the 3 tasks prior to the experiment and compiled a broad set of passages covering most of the aspects of each topic. At the time of the experiment, the Wizard tried to find the best passage to reply to the participant's question/comment. However, in cases where such passage could not be found, the Wizard would reply with a passage retrieved from web search, or write a new passage. In case the participant's question or comment was ambiguous, the Wizard was allowed to ask a clarification question to better identify the information need of the participant.
Our Wizard agent was allowed to maintain the context of the conversation, respond to vague questions, understand implied concepts, and provide active feedback in form of clarification questions when needed (all of these capabilities do not yet exist in commercial systems). At the same time, by partially restricting the Wizard to a precompiled set of passages, we could maintain some consistency of answers between participants, i.e., for the same question any participant would receive the same answer. By analyzing the ways the participants communicated with the Wizard agent, we could gain insights about strategies people use in a human-computer dialogue for solving complex tasks and look for design implications for automatic conversational systems.

Human agent.
To answer our second research question, about the differences between human-to-human and human-to-computer communication, we devised our second conversational agent – the Human agent. In this case, the Wizard from the previous setup was still serving as a backend, but the participants were explicitly informed that they were talking to a live person. Another difference was that the Human agent was not restricted to the pre-retrieved set of passages but was free to slightly reformulate or revise the passages to better respond to the question. By including both the Human and Wizard agents in the study, we were able to maintain a constant level of intelligence for both agents, thus comparing not the accuracy of each agent, but rather the participants' attitude and expectations towards a perceived automatic agent compared to a known human.

Automatic agent.
As a means of comparison to an existing conversational agent, we used the Google Assistant as a backend for our third agent. Every message sent by a participant was forwarded to the Google Assistant app, and the response was forwarded back to the participant. Most of the time, the response consisted of an URL and a text snippet. The participants were told that they were interacting with another experimental conversational system, but were not given any specific information about it. By using a system representative of the state-of-the-art technology, we were able to evaluate its drawbacks, and situations where it failed to respond properly.

5. Results and Discussion

After running the study, we analyzed message logs, answers to topical quizzes, and preference questionnaires and found the most popular trends and answers. This section describes our findings in detail.

Overall satisfaction.
After completing each task participants rated their overall experience of working with each agent on a 1 to 5 Likert scale. Average ratings for each agent are shown in Table 1.

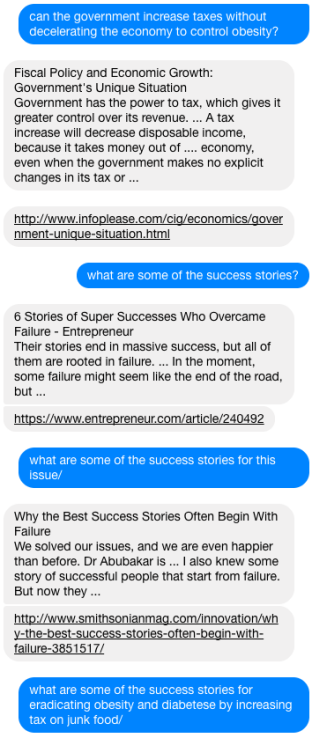

Figure 2: Automatic system (gray background) fails to maintain context, which causes the participant 15 (blue background) to reformulate his question twice.

Agent	Human	Wizard	Automatic
Overall satisfaction	4.1	3.8	2.9
Able to find information	1.5	1.3	1.0
Topical quiz success	1.6	1.6	1.3

Table 1: Row 1: average satisfaction for each agent; row 2: average rate of success for finding desired information; row 3: average rate of success for answering topical quiz questions.

The differences in ratings of Human vs. Automatic systems, and the Wizard vs. Automatic systems, were statistically significant ($p < 0.0001$ and $p < 0.0005$ respectively), while the difference between the Human vs. Wizard systems was not significant. In the final questionnaire, after completing all the tasks, participants were asked which system they liked the most. Out of 21 people, 8 people preferred the Human agent, 6 – the Wizard agent, 4 – the Automatic agent, 2 people said they would use the Wizard or the Human depending on their goals, and 1 person said he would choose between the Human and the Automatic agent depending on her goals.

Able to find information.
After completing each task we also asked participants whether they were able to find all the information they were looking for. We coded each answer on a 0-2 scale (0 - no, I couldn't; 1 - partially; 2 - yes, I found everything I needed). Average results for each agent are shown in Table 1.

Topical quiz success.
After completing each task participants were asked 3 questions about the topic. We evaluated those questions on a scale 0-2, where 0 meant no answer, 1 - poor answer, 2 - good answer. On average, participants showed a similar level of success with each agent. The average user ratings for each agent are shown in Table 1.

These results confirm our initial intuition that human-to-human conversation is more natural for the open-ended problem of the complex search task, compared with automatic conversational agents. This could be because people have experience talking to other people, and the results match their initial expectations. On the other hand, for any system that people have no experience with, they have to learn its functionality and ways to interact with it effectively. We now turn to qualitative results, reporting the comments participants provided in the post-study questionnaire. The participants' comments broke down into the areas of maintaining context, trustworthiness, and social burden.

Maintaining context.
Participant 19 (P19): "It didn't use contextual information so there was no way to expand on the previous answer it gave me." Within a conversation, people expect that the main topic of the discussion is maintained, and they tend to ask short questions, omitting the subject, or referring to the subject using pronouns. Formulating a full question takes effort and is unnatural. For the Automatic system, anaphora resolution did not always work, which annoyed the participants. Similarly, when dealing with the Human and Wizard systems, participants pointed out the ease of use, because their partially stated questions were understood, and relevant answers were returned.

Trustworthiness of the sources is crucial.
P7: "I ... like to be able to verify the credibility of the sources used." Even though the Automatic system did not always respond with a relevant result, it received approval from our participants for providing sources of its answers. Out of 21 participants, 13 people said that being able to access the URL allowed them to assess the trustworthiness of the source and therefore to accept or reject the answer. On the other hand, in spite the Human and Wizard systems

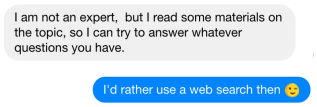

Figure 3: A participant prefers web search to talking to a person. Part of a conversation between participant 7 (blue background) and Human agent (gray background).

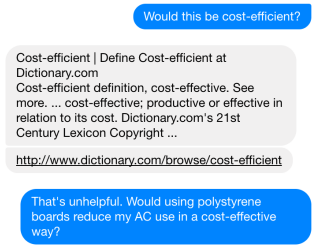

Figure 4: Explicit user feedback could be used to recover from failure. Part of a conversation between participant 12 (blue background) and Automatic system (gray background).

returning more relevant results, they were both criticized for not providing the sources.

Social burden.
P15: "you have to think about social norms, asking too much, being too stupid, not giving them enough time to respond, troubling them." When dealing with the Human system, 4/22 participants reported that they felt uncomfortable talking to a person, thought more about the social norms, were afraid to ask too many questions, were not sure how to start and end a conversation. This additional burden of interacting with humans further motivates research in the area of automated conversational agents as the medium of choice for a notable fraction of use cases.

Discussion and design implications.

Based on our findings we devised a list of recommendations for a conversational agent design, that according to our empirical study will improve user experience significantly.

Context.
Maintaining a context of the conversation to enable short questions and comments is crucial to user experience since formulating long sentences each time feels unnatural and takes longer.

Provide sources of answers.
Finding relevant and precise answers is important. But trustworthy sources are equally important, and their absence may diminish the credibility of the system. While the Automatic agent supported each answer with an URL, Human and Wizard did not, unless specifically asked.

Use feedback.
One crucial difference of conversational setup from web search is the ability of a user to provide the system with explicit feedback. It is likely to contain essential information that may help the system to get back up from failure and improve upon the previous result.

Opinion aggregation.
According to the participants, sometimes what is needed is the *experience* of other people in similar situations. A good conversational system should be able to aggregate opinions and present them to the user in a short summary, perhaps explaining each one. Participant 17 said: *"It would be nice if I could see a summarization of different opinions that there exist – from different sources."*

Direct answers vs. expanded information.
For this aspect, our participants split into 2 camps: those who prefer getting direct answers to the question provided, and those who prefer also getting a broader context. People from Camp 1 complained that the answers returned by the systems were too long (even for the Wizard and Human), and preferred to have their questions answered directly with minimum extra information. Camp 2, on the other hand, said that they prefer talking to a person, who would recognize their true information need (beyond the immediate question) and provide the relevant information.

Conclusions and future work.

In this paper, we investigated human behaviour when using conversational systems for complex information seeking tasks. We also compared participant behaviour when talking to a human expert, vs. a perceived automatic system. We observed that people do not have biases against automatic systems, and are glad to use them as long as their expectations about accuracy were met. Future research directions include further investigating the possibilities for improving existing conversational agents and studying the effect of these changes on user experience.

References

[1] Eyal Aharoni and Alan J Fridlund. 2007. Social reactions toward people vs. computers: How mere lables shape interactions. *Computers in human behavior* 23, 5 (2007), 2175–2189.

[2] Anne Aula, Rehan M Khan, and Zhiwei Guan. 2010. How does search behavior change as search becomes more difficult?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 35–44.

[3] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. What do you mean exactly? Analyzing clarification questions in CQA. In *CHIIR'2017*.

[4] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2014. *Overview of the TREC 2014 session track*. Technical Report. DTIC Document.

[5] Maxime Clément and Matthieu J Guitton. 2015. Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior* 50 (2015), 66–75.

[6] Chad Edwards, Autumn Edwards, Patric R Spence, and Ashleigh K Shelton. 2014. Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior* 33 (2014), 372–376.

[7] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[8] Marti A Hearst. 2011. "Natural" search user interfaces. *Commun. ACM* 54, 11 (2011), 60–67.

[9] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. 2012. When web search fails, searchers become askers: understanding the transition. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 801–810.

[10] Ewa Luger and Abigail Sellen. 2016. Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.

[11] Alessandro Murgia, Daan Janssens, Serge Demeyer, and Bogdan Vasilescu. 2016. Among the Machines: Human-Bot Interaction on Social Q&A Websites. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1272–1279.

[12] Bo Pang and Ravi Kumar. 2011. Search in the lost sense of query: Question formulation in web search queries and its temporal changes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 135–140.

[13] Filip Radlinski and Nick Craswell. A Theoretical Framework for Conversational Search. In *CHIIR'2017*.

[14] C Tsai, Wen-tau Yih, and C Burges. 2015. *Web-based question answering: Revisiting AskMSR*. Technical Report. Technical Report MSR-TR-2015-20, Microsoft Research.

[15] Ryen W White, Matthew Richardson, and Wen-tau Yih. 2015. Questions vs. queries in informational search tasks. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 135–136.