Interactive Evaluation of Conversational Agents:

Reflections on the Impact of Search Task Design

Mateusz Dubiel
mateusz.dubiel@strath.ac.uk
University of Strathclyde

Martin Halvey
martin.halvey@strath.ac.uk
University of Strathclyde

Leif Azzopardi
leif.azzopardi@strath.ac.uk
University of Strathclyde

Sylvain Daronnat
sylvain.daronnat@strath.ac.uk
University of Strathclyde

ABSTRACT

Undertaking an interactive evaluation of goal-oriented conversational agents (CAs) is challenging, it requires the search task to be realistic and relatable while accounting for the user‘s cognitive limitations. In the current paper we discuss findings of two Wizard of Oz studies and provide our reflections regarding the impact of different interactive search task designs on participants' performance, satisfaction and cognitive workload. In the first study, we tasked participants with finding a cheapest flight that met a certain departure time. In the second study we added an additional criterion: 'travel time' and asked participants to find a fight option that offered a good trade-off between price and travel time. We found that using search tasks where participants need to decide between several competing search criteria (price vs. time) led to a higher search involvement and lower variance in usability and cognitive workload ratings between different CAs. We hope that our results will provoke discussion on how to make the evaluation of voice-only goal-oriented CAs more reliable and ecologically valid.

KEYWORDS

Conversational Search, Performance Evaluation, User Study

ACM Reference Format:
Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2020. Interactive Evaluation of Conversational Agents:: Reflections on the Impact of Search Task Design. In *The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '20), September 14–17, 2020, Virtual Event, Norway.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3409256.3409814

1 INTRODUCTION

Conversational Agents (CAs) are systems that enable natural language interaction, unconstrained by menus, command prompts and key words. To be effective, a CA needs to present search results in a way that: (1) does not overwhelm users with information and (2) gives them a good overview of the information space cf. [21]. Meeting these two requirements is especially challenging in an audio channel since users need to remember the presented information and reason about it simultaneously [22]. A mixed conversational initiative between the CA and a user was proposed as a solution to tackle this problem - with several conversational theoretical frameworks advocating active involvement of a conversational agent to reduce cognitive workload and enhance performance [1, 17, 19]. However, to date, few interactive evaluations have been carried out to investigate how the active involvement of the CA impacts users' performance, satisfaction and cognitive workload [9].

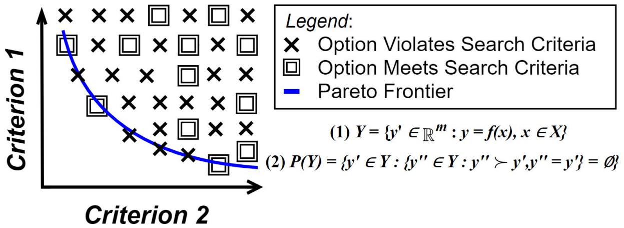
Criterion 1
Criterion 2
Legend:
× Option Violates Search Criteria
▣ Option Meets Search Criteria
Pareto Frontier
(1) $Y = \{y' \in \mathbb{R}^m : y = f(x), x \in X\}$
(2) $P(Y) = \{y' \in Y : \{y'' \in Y : y'' \succ y', y'' = y'\} = \emptyset\}$

Figure 1: A conceptual representation of search space.

The current paper is a reflection on the relationship between the complexity of search task and users' subjective evaluations of voice-only CAs. Specifically, we present the results of two lab-based Wizard of Oz [7] studies that investigated the impact of CAs' search support (active vs. passive) on users' performance (objective measure), satisfaction and cognitive workload (subjective measures). In each study, participants completed four goal-oriented tasks, where they searched for flights. Participant performance was evaluated in terms of meeting search criteria and task completion time. For subjective measures, we looked at participants' self-reported cognitive workload (measured by NASA TLX [11]) and satisfaction with the agent (measured by System Usability Score [SUS] [4]). Both measures allowed us to gauge how effective a CA providing active search support was compared to a passive CA - a system based on the slot-filling architecture (slots specify the information required to process a user's query). In Study 1, we provided participants with two search criteria i.e. 'price' and 'departure time'. While in Study 2, we added another criterion, 'duration of the flight' and changed the 'required departure time' to 'preferred time of arrival.' We also outlined the implications of missing the preferred time of arrival (additional check-in fee, getting stuck in traffic etc.) We noticed that different search criteria led to changes in participants' behaviour and impacted their performance and satisfaction with the CAs. While in Study 1, most of the participants took the fastest path to task completion, in Study 2 participants engaged in a more thorough exploration of the search space. The results of the cross comparison of both studies indicated that the design of the search task has an important impact on behavioural measures and users' search performance. Our paper provides reflections on the role of task design in evaluating of voice-only CAs for goal-oriented tasks.

2 BACKGROUND

The impact of search task complexity on user behaviour and search performance is a long-standing research problem in interactive

ICTIR '20, September 14–17, 2020, Virtual Event, Norway

ACM ISBN 978-1-4503-8067-6/20/09…$15.00
https://doi.org/10.1145/3409256.3409814

information retrieval. Kelly et al. [12] found that while more cognitively demanding tasks required significantly more search activity from participants, they did not impact participants' satisfaction. Choi et al. [5] found that task complexity had an impact on participants' perceptions about temporal demand and satisfaction with the time spent on the task, and showed that participants with lower working memory were more prone to satisfying behaviour (i.e. unwilling to engage in exploration). Moffat et al. [14] argued that the task complexity should be incorporated as a metrics to evaluate the effectiveness of an IR system. They showed that user search strategies vary based on task complexity and therefore require different levels of support from the system in order to access the information required to complete the task. Although insightful, the above studies ([5, 12, 14] focused only on a traditional search where queries are typed in a search box and results examined on the screen.

Another strand of research in the voice-only search domain has explored the role of task complexity on user search behaviour in an information seeking context ([18, 20] among others). Trippas et al. [20] analysed interaction patterns between pairs of seekers and intermediaries and observed that interaction time increased with task complexity. Thomas et al. [18] found that as participants had to exercised more effort, their engagement decreased. It should be noted that while studies of Trippas et al. and Thomas et al. focus on information seeking scenarios, designing goal-oriented scenarios where user decisions have impact on meeting requirements remains a challenge. This challenge arises from a limited understanding of how users would behave in a voice-only search scenario where their decisions can have direct implications on search outcomes and available resources.

A goal-oriented dialogue can also be considered in terms of a cost-benefit trade-off, where the usefulness of a CA is determined by its ability to resolve a user's information need quickly and comprehensibly [1]. A problem with the current generation of CAs is that they provide information in a verbose way which puts strain on the user who needs to retain alternative options in memory. Consequently, due to cognitive overload, users tend to accept the first minimally acceptable option (satisficing) rather than continuing to absorb the cost of interaction in order to find a better option (maximising) [13]. On the other hand, users are unlikely to accept the CA's best suggested option without exploring alternatives [13]. It is thus important to focus on the task design for voice-only search scenarios that would allow for higher participant involvement and lead to more natural interactions. This can consequently aid the development of a more user-friendly CA that promote exploratory behaviour and minimises satisficing (i.e. reduces the likelihood of a user settling for the first minimally acceptable option).

3 METHOD

We conducted two lab-based search studies using a Wizard of Oz (WoZ) methodology [7], where participants interacted with a wizard (a person simulating a conversational agent) in order to find and book flights. Booking a flight is a goal-oriented task that allowed us to measure the performance of participants and evaluate the impact of different CAs. We focused on the most salient factors that impact participants' choice: flight price, flight duration and arrival time (as identified by the IATA Global Passenger Survey [16]). In each study, participants were asked to imagine that they are a traveller looking for a one-way flight. Both studies are discussed below.

Study 1: In total, 22 people participated in the study (13M, 9F; M_{age} = 29, SD = 10). One participant did not complete all the tasks and their data was withdrawn. Two CAs were used: the state-of-the-art voice search agent (Passive CA) based on a 'slot-filling' architecture which represented the current state of the art; a and conversational human-like agent (Active CA). The Passive CA was designed based on the design recommendations outlined in [15]. In total, participants completed four search tasks. After each task participants filled in NASA TLX [11] and SUS [4] questionnaires. An example search scenario is presented below.

> You are planning to visit your friend who lives in **Bristol**. You will be flying from Glasgow Airport. Your total budget is **100 pounds**. You can travel either on the 11th or 12th of November.
> **Indicative Request:** You want to find the cheapest possible deal but your flight needs to leave on, or before 11am.
> **Note:** Please wait for the agent to finish before you start to speak.

At the beginning of each interaction, the Passive CA provided participants with a welcome message and presented them with its functionalities. Participants were asked to provide their search criteria, namely 'destination airport', 'date of travel', and 'available budget'. The Passive CA also provided an example to help them formulate their query: 'For example, you can say I'm travelling to London on the 2nd of December and my budget is 100 pounds'. The Active CA started with a greeting: 'Hello, how can I help you?' to prompt the participant to provide their search query. Interaction with the Active CA was not constrained in any way. Participants were free to provide information in any order and the CA would ask them follow up questions to clarify their intent. In contrast, the Passive CA system could only process a query once all of the requested information was provided. Participants' performance was evaluated based on two criteria: (1) Does the flight meet the price requirement? (2) Does the flight meet the departure time requirement?

The prompts used in both of the systems were prepared in accordance with the guidelines outlined in [6]. The prompts were made to resemble a natural spoken discourse by use of appropriate cohesive devices (pronouns and discourse markers), adhering to the principles of information structure (providing new information at the end of the utterance, and applying Grice's 'Cooperative Principle' [10] (making assumptions about inferences that users will draw from the prompts). During the interaction, the Wizard (the lead researcher) played pre-recorded prompts using a GUI. For any unexpected participant responses a live-speech synthesis tool was used. Participants were addressing their requests to a stand alone device representing the CA.

Study 2: In total, 24 participants completed the Study 2 (12M, 12F; M_{age} = 26, SD = 6). Study 2 expanded on Study 1 by focusing on different methods of results presentation. During interactive search scenarios, the wizard followed different conversational strategies for providing information: i.e. Listing (a detailed list with *2* elements) and Summarising (aggregating different sets of options and then presenting them to users in ranges). Both strategies were implemented in a Passive and Active mode. In Study 2, we assigned a name to each of the strategies to make them easily identifiable to the participants. The names of the agents were: **Angus:** Passive Summarising, **Blair:** Passive Listing, **Calum:** Active Summarising and **David:** Active Listing. During search sessions the Passive CAs (Angus & Blair) did not make any suggestions with regards to results filtering while the Active CAs (Calum & David) asked participants questions to progressively narrow down the list of flight options. To control for the impact of the agent on search performance, a 4x4 Latin Square [3] was used to rotate the CAs.

In Study 2, participants were instructed to explore the available options to find the shortest and cheapest flight *that meets a certain arrival preference* (e.g. 'You want to reach your destination around 1pm to avoid traffic.'). Contrary to Study 1, we did not provide a strict budget but encouraged participants to explore the trade-off between flight cost and travel time, we also motivated the exploration with a background story (e.g. 'You are a student who is attending the conference in Stockholm, try to save money from your travel fund while making sure that you reach your accommodation on time'). This highlighted the implications of not meeting the provided search criteria, i.e. having to pay for late check-in. A full task description is presented below.

> You will be attending a student conference in **Stockholm**. You will be travelling there on either **Monday the 5th, Tuesday the 6th, or Wednesday the 7th of November**. Your university advised you that you will be allocated money from your conference fund that you will use to fund other events till the end of your academic course. To be able to attend more events in the future, you want to save money while not spending too long getting there. The student dorms where you will be staying charge extra for late check-in, so you will be aiming to arrive at around 7pm to be able to check in to your accommodation on time.
>
> **Indicative Request:** Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
>
> **Note:** Please wait for the agent to finish before you start to speak.

We measured task performance in terms of the distance of the selected flight from the Pareto Frontier (illustrated in Figure 1) and in terms of preference (desired travel time specified in search scenario). For instance, if a participant booked a flight for £100 with a duration of 2 hours, and there was another flight available at the same price which took 1 hour - the distance from the Pareto frontier was 1 hour. In terms of time preference, if the scenario specified the 'preferred arrival time' as noon and the participant arrived at 4pm - the absolute difference was 4 hours. In terms of task outcome, we also looked at how much money each participant spent on flights when using different agents, and how much travel time they potentially wasted by selecting a longer flight. The premise of introducing the Pareto measure was to encourage exploration and minimise satisficing behaviour, since, like in real life, users want to get the best value for money (cf. [2, p. 345]).

Pareto frontier: Figure 1 is a conceptual representation of the search space used in Study 2. The Pareto Frontier is formally represented by equations (1) and (2). Where R^{m} is considered to be a space of flights, X represents feasible flight options in this space, and Y is a set of decision vectors such that given our preferred decisions criteria (short travel time and low price), all points on the Pareto frontier dominate over other points in the search space.

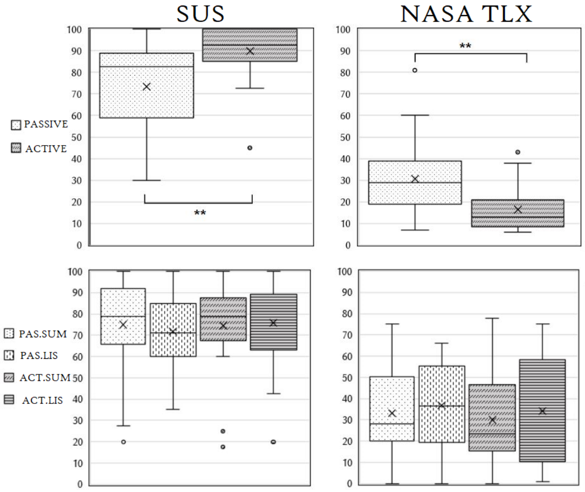

Figure 2: Comparison of Objective Metrics for Study 1 (top) and Study 2 (bottom). Active CA(s) are highlighted in grey. Note: for SUS the higher score indicates better performance, for NASA TLX, the lower score indicates that the system is less cognitively taxing.

Table 1: Subjective Measures, ** indicates p < 0.01

Study	CA Type	SUS Med (IQR)	NASA TLX Med (IQR)
1	*Passive*	81(22)	30(18)
	Active	**93(17)****	**14(13)****
2	Passive Sum.	79(26)	28(30)
	Passive List.	71(25)	37(36)
	Passive Overall	76(29)	33(31)
	Active Sum.	79(20)	26(30)
	Active List.	81(26)	27(44)
	Active Overall	80(22)	26(33)

4 RESULTS

Comparing the results of Studies 1 and 2 indicated that differences in task design regarding the number and type of search criteria lead to substantial differences in performance and satisfaction.

Subjective Measures: Subjective measures are shown in Figure 2 and Table 1. Since most of the data was not normally distributed in either study, for pair-wise comparisons, unless stated otherwise, we used the Kruskal Wallis H Test. We observed notable differences in participants' perceptions of the CAs. In Study 1, there was a statistically significant difference for both overall perception of cognitive workload between Passive and Active CAs (p < .001) and their usability (p = .003). However, In Study 2, although Active CAs were also perceived as less cognitively taxing and more usable, the differences between the agents were not significant. This result indicates that more nuanced tasks with competing search criteria required more focus and participant involvement which was manifested in higher TLX scores (as participants had to exercise more effort) and, consequently, lower satisfaction with CAs.

Objective Measures: Table 2 shows a comparison of objective measures for both studies. Since task success was considered in binary categories, i.e. a selected flight is either optimal or not, we used Cochran's Q Test to compare our CAs. The performance of

Table 2: Objective Measures, *** indicates p < 0.001

Study	CA	Found	Task Time
	Type	Optimal Flight	Med(IQR)
1	*Passive*	22/42(52%)	127(42)
	Active	**38/42(90%)*****	**58(28)*****
2	Passive Sum.	3/24(13%)	243(113)
	Passive List.	2/24(8%)	210(180)
	Passive Overall	5/48(10%)	232(122)
	Active Sum.	8/24(33%)	271(161)
	Active List.	5/24(21%)	229(88)
	Active Overall	13/48(27%)	252(132)

participants varied radically between the two studies. In Study 1, where participants had to satisfy two search criteria, The Active CA led to significantly better performance (p < 0.001). Yet, this was not the case in Study 2 (p = 0.07) where participants had to evaluate the trade-off between flight cost and travel time while trying to meet the required arrival time. A similar pattern was observed for task completion time. In Study 1 participants completed their search tasks significantly faster when using an Active CA (p < 0.01), while in Study 2 participants invested more time trying to find the optimal fight. This comparison indicates that more nuanced scenarios, in Study 2, led to a more in-depth exploration. In particular, when faced with additional constraints regarding the desired time of arrival, participants seemed to put in more effort to mitigate negative implications (e.g. additional check-in fee, etc.)

5 REFLECTIONS ON TASK DESIGN

Based on the results of our comparative analysis of two WoZ studies, we make the following observation regarding the impact of task-complexity on the evaluation of voice-only, goal oriented CAs. **(1) Tension between search criteria:** When the wording of the task highlighted the trade-off between the provided search metrics (e.g. flight cost vs. flight duration), participants were more involved in the task and explored the search space more thoroughly. Consequently, participants were less likely to satisfice and spent more time interacting with the CA. The more restrictive set-up that included a larger number of search criteria, provided us with more conversational material for analysis and , in turn, richer insights into participants' behaviour. **(2) Realistic Constraints:** When the search task provided participants with a soft constraint (e.g. preferred time of arrival), motivated by the background story (i.e. reasons for meeting arrival time), we observed a higher involvement in the task - as participants tried to mitigate the negative consequences of missing the recommended arrival time, (e.g. getting stuck in the traffic, paying extra for check-in, etc.) Post-task interviews for Study 2 [8] indicated that participants found the tasks relatable which indicates higher validity of our method.

Our observations of participants' search behaviour provided us with some preliminary insights regarding the design of goal-oriented, voice-only search tasks. However, there are still many open questions that merit further exploration. These include: (1) How to account for personal differences in cognitive abilities and expertise? (2) What other realistic goal-oriented tasks can be suitable for evaluation in the voice-only domain? (3) Can performance-based incentives be used to effectively motivate participants?

6 CONCLUSIONS

The cross-comparison of two voice-only, interactive WoZ studies indicates that conversational search tasks that add tension between different search criteria encourage more thorough exploration of the search space and higher participant involvement. Interestingly, we observed that more constrained tasks led to more exploration. Although less constrained tasks that did not involve a trade-off between search criteria made users' perceptions of the differences between the active and the passive agents more explicit, they offered less motivation and realism. This, in turn, resulted in less realistic conversational data. While the wording and selection of search criteria for voice-search tasks remain an open problem, we hope that our study stimulates debate on how to make an interactive evaluation of CAs effective and insightful.

REFERENCES

[1] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *2nd International Workshop on Conversational Approaches to Information Retrieval.*

[2] Hans H Bauer, Nicola E Sauer, and Christine Becker. 2006. Investigating the relationship between product involvement and consumer decision-making styles. *Journal of Consumer Behaviour* 5, 4 (2006), 342–354.

[3] James V Bradley. 1958. Complete counterbalancing of immediate sequential effects in a Latin square design. *J. Amer. Statist. Assoc.* 53, 282 (1958), 525–528.

[4] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[5] Bogeum Choi, Robert Capra, and Jaime Arguello. 2019. The Effects of Working Memory during Search Tasks of Varying Complexity. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval.* 261–265.

[6] Michael H Cohen, Michael Harris Cohen, James P Giangola, and Jennifer Balogh. 2004. *Voice user interface design.* Addison-Wesley Professional.

[7] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—why and how. *Knowledge-based systems* 6, 4 (1993), 258–266.

[8] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, Damien Anderson, and Daronnat Sylvain. 2020. Conversational Strategies: Impact on Search Performance in a Goal-Oriented Task. In *3rd International Workshop on Conversational Approaches to Information Retrieval (CAIR'20).* ACM, 1–7.

[9] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2018. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The Second International Workshop on Conversational Approaches to Information Retrieval.*

[10] Grice et al. 1975. Logic and conversation. *1975* (1975), 41–58.

[11] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index). *Advances in psychology* 52 (1988), 139–183.

[12] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval.* 101–110.

[13] Page Laubheimer and Raluca Budiu. 2018. Intelligent Assistants: Creepy, Childish, or a Tool? Users' Attitudes Toward Alexa, Google Assistant, and Siri. *Nielsen Norman Group* (2018).

[14] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 1–38.

[15] Ditte Mortensen. 2017. How to Design Voice User Interfaces. *Interaction Design Foundation* (2017).

[16] PWC. 2015. 2015 IATA global passenger survey. (2015).

[17] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *CHIIR 2017.* ACM, 117–126.

[18] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *SIGIR CAIR.*

[19] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: perspective paper. In *Proceedings of the ACM CHIIR 2018.* ACM, 32–41.

[20] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the ACM CHIIR 2017.* 325–328.

[21] Johanne R Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. 2015. Results presentation methods for a spoken conversational search system. In *Proceedings of the First International Workshop on Novel Web Search Interfaces and Systems.* ACM, 13–15.

[22] Maria Wolters, Kallirroi Georgila, Johanna D Moore, Robert H Logie, Sarah E MacPherson, and Matthew Watson. 2009. Reducing working memory load in spoken dialogue systems. *Interacting with Computers* 21, 4 (2009), 276–287.