Comparing User Experience in Search Interaction for Conversational and Conventional Search Systems using Implicit Evaluation Methods [Prototype]

Abhishek Kaushik[*,†], Gareth J.F. Jones

ADAPT Centre, School of computing, Dublin City University, Dublin, Ireland

Abstract

Conversational search offers the prospect of improved user experience in information seeking via agent support. However, it is not clear how searchers will respond to this mode of engagement in comparison to a conventional user-driven search interface, such as those found in a standard web search engine. We describe a laboratory-based study directly comparing user behaviour for a prototype agent-mediated multiview conversational search interface (MCSI) which extends the functionality of a conventional search interface (CSI) with that of an equivalent CSI. User reaction and search outcomes of the two interfaces are compared for a set of scenario-based search tasks using implicit evaluation with two analysis methods: workload-related factors (NASA Load Task) and psychometric evaluation. Our investigation shows the MCSI to be more interactive and engaging, with users claiming to have a better search experience in contrast to the corresponding CSI.

Keywords

conversational search interface, conventional search interface, user satisfaction

1. Introduction

Bringing together the needs of users of search technologies for unstructured information and advances in artificial intelligence, recent years have seen rapid growth in research interest in the topic of *conversational search (CS)* systems. These systems assume the presence of an agent of some form which enables a dialogue interaction between the searcher and the search engine to support them in satisfying their information needs [1]. While there has been much discussion of the potential of CS methods, there is little work reporting on the investigation of operational CS prototypes, and in particular how these compare with conventional search systems used to perform the same search task. Those studies of CS which have appeared generally adopted a human "wizard" in the role of the search agent [2, 3]. Studies with these systems have been conducted with the implicit assumption that an agent can interpret the searcher's actions with human like intelligence. In this study, we take a alternative position using an automatic

DESIRES 2022 – 3rd International Conference on Design of Experimental Search & Information REtrieval Systems, 30-31 August 2022, San Jose, CA, USA

[*]Corresponding author.

[†]Now at Dundalk Institute of Technology, Dundalk, Ireland.

abhishek.kaushik2@mail.dcu.ie (A. Kaushik); Gareth.Jones@dcu.ie (G. J.F. Jones)

0000-0002-3329-1807 (A. Kaushik); 0000-0003-2923-8365 (G. J.F. Jones)

rule-based agent to support the searcher when using a prototype CS interface, and compare this with the effectiveness of a similar conventional search interface (CSI) to perform the same search tasks. In this study, we introduce a prototype agent-mediated multiview conversational search interface (MCSI) which uses a search engine API, shown in operational example videos at link[1]. Our interface combines a CS assistant with an extended standard graphical search interface. The interface agent takes the form of a personal assistant which works beside the user, rather than sitting between the user and the search engine [4].

Previous studies of CS interfaces have focused on have chatbot type interfaces which limit the information space of the search [3, 5], and are very different from conventional graphical search interfaces. Search via engagement with a chat type agent can result in the development of quite different information-seeking mental models to those developed in the use of standard search systems, meaning that it is not possible to directly consider the potential of CS in more conventional search settings. We are interested to consider how user experience differs between use of the MCSI and an equivalent CSI. For our study, we adopt a range of implicit evaluation methods [6]. Specifically we use cognitive workload-related factors (NASA Load Task) [7] and psychometric evaluation for software [8]. Our findings show that users exhibit significant differences for these evaluation dimensions when using our MCSI and a corresponding CSI.

This paper is structured as follows: Section 2 overviews the features of our MSCI, Sectio 3 describes the methodology for our investigation, Section 4 provides details of our results, and includes analysis, findings and hypothesis testing and Section 5 concludes.

2. Prototype Conversational Search System

We developed a fully functioning prototype MCSI, the user interface of which is shown in Figure 1. The interface is divided into two distinct sections. The righthand side which corresponds to a CSI with which the user can interact, while the lefthand side is a text-based chat agent which interacts with both the search engine and the user. Essentially the agent works alongside the user as an assistant, rather than being positioned between the user and the search engine [4]. The Web interface components are implemented using the web python framework flask and with HTML, CSS, and JS toolkits. The agent is controlled by a logical system and is implemented using Artificial Intelligence Markup Language (AIML) scripts. The MSCI interacts with the standard Wikipedia API. The interface includes use of an algorithm which highlights important segments within long retrieved documents to enable the searcher to skim through them. A more detailed description of the components and dialogue scheme is discussed in our previous publication [9].

The prototype user interface provides the user with the flexibility to interact with both the search assistant and directly with the search engine. The system enables the user to explore a chosen document by presenting them with multiple subtopics from the document.

The system also provides support to a user to support them in reading full documents. As described above, important sections in long documents are highlighted to ease reading and reduce the cognitive effort.

[1] https://drive.google.com/open?id=1AoS5Nrnj7nGrPIsRAiA96ttwvzzwkpCK

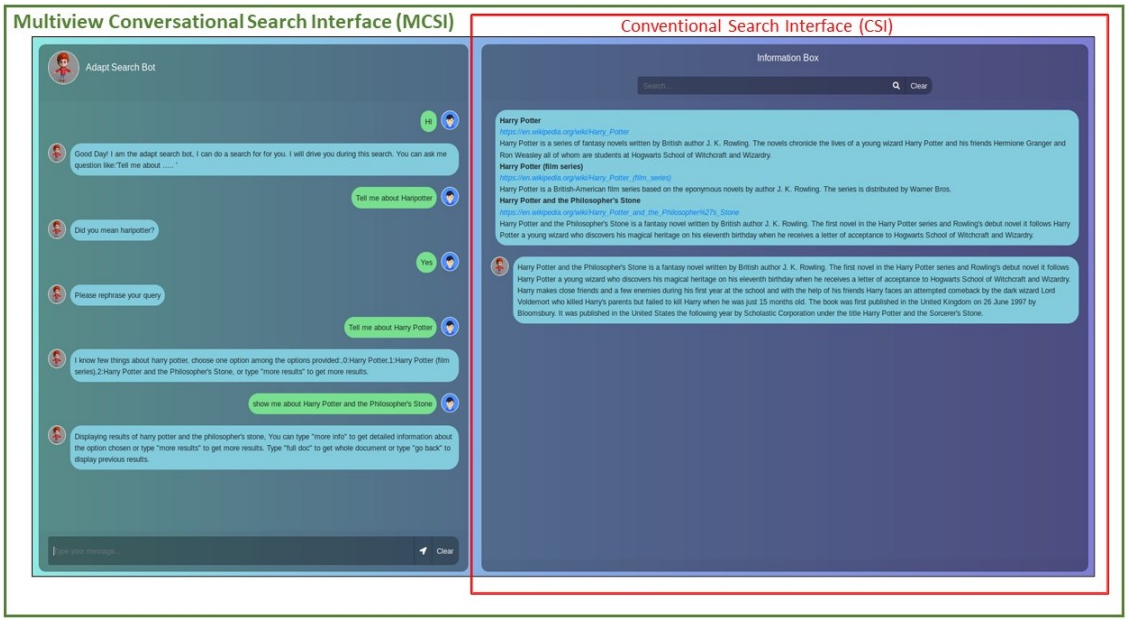

Figure 1: Conversational Search Interface: chat display, chat box, information box, query box with action buttons for Enter and Clear, and retrieved snippets and documents. Green outline indicates the MCSI setting and red block indicates the Conventional Search Interface setting.

`It is late, but you can't get to sleep because a sore throat has taken hold and it is hard to swallow. You have run out of cough drops, and wonder if there are any folk remedies that might help you out until morning.`

Figure 2: Example backstory from UQV100 test collection.

To enable direct comparison with our MCSI, a CSI for our study was formed by using the conversational interface with the agent panel removed and the document highlighting facilities disabled. The searcher enters their query in the query box, and document summaries, are returned by the Wikipedia API, and full documents can be selected for viewing to satisfy the user's information need.

3. Methodology

In this section we describe the details of our user study. The study aimed to enable us to observe and better understand and contrast the behaviour of searchers using our prototype MCSI and the corresponding CSI.

3.1. Information Needs for Study

For our investigation, we wished to give searchers realistic information needs which could be satisfied using a standard web search engine. In order to control the form and detail of these

needs, we decided to use a set of information needs specified within *backstories*, e.g. as shown in Figure 2. The backstories used are taken from the UQV100 test collection [10], whose cognitive complexity is based on the Taxonomy of Learning [11]. 12 of the most cognitively complex backstories were selected for use in our study, using the selection mechanism from our study [12].

3.2. Experimental Procedure

Participants in our study had to complete search tasks based on the backstories using the MCSI and CSI. Each session consisted of multiple backstory tasks, with participants completing pre- and post task search questionnaires.

Participants used a setup of two computers with two monitors side by side on a desk in our laboratory. One monitor was used for the search session, and the other to complete the online questionnaires. The questionnaire was divided into three sections: a) *Basic Information Survey*: assigned user ID, age, occupation and task ID. b) *Pre-Search*: details of the participant's pre-existing knowledge with respect to topic of the search task. c) *Post-Search*: Post-search feedback from the user. The questionnaire used an online Google form. All search activities were recorded using a standard screen recorder tool to enable post-collection review of the user activities. Approval was obtained from the relevant university Research Ethics Committee prior to the data collection.

A pilot study was conducted with two undergraduate students in Computer Science using two additional backstory search tasks. This enabled us to see how long it took them to complete the sections of the study using the MCSI and CSI, to gain insights and debug the experimental setup. Results from the pilot study are not included in the analysis.

Based on the result of the pilot study, each participant in the main study was assigned two of the 12 selected task backstories with the expectation that their overall session would last around one hour. Pairs of backstories for each session were selected using a Latin square procedure. After every six tasks the sequence of allocation of the interface types was rotated to avoid sequence effects [13].

In total, 27 subjects (18 Males and 9 Females) participated in our study (excluding the pilot study), we examined the data of 25 subjects, since 2 subjects were found not to have followed the instructions correctly. The study was conducted in two phases. Each user had to perform search tasks using the CSI and MCSI with the sequencing of their use of the interfaces varied to avoid learning or biasing effects.

As well as completing the questionnaires, the subjects also attended a semi-structured interview after completion of their session of two tasks using both interface conditions. The videos and interviews were thematically labelled by two independent analysts and Kappa coefficients were calculated (approx mean .85) [14]. Disparities in labels were resolved by mutual agreement by analysts. The questionnaire in the interview dealt with user search experience, software usability and cognitive dimensions and was quantitatively analyzed. Based on the interview analysis, out of 25 participants 92% of the total subjects were happy and satisfied with the MCSI. This shows that there was no sequence effect arising from the order of the interfaces in the search sessions. In all conditions, subjects preferred the MCSI.

Each hypothesis of the study was tested using a T-Test (since the number of samples was less

Table 1
Task load index to compare the load on user while using MCSI and CSI systems with independent T two tailed test .

Task Load Index	CSI Mean	MCSI Mean	Percentage Change	P-Value
Mentally Demanding	4.16	3.68	11.54	.273795
Physically Demanding	3.12	2.76	11.54	.441676
Hurried or Rushed	3.34	2.76	14.81	.213878
Successful Accomplishing	4.28	5.32	-24.3	.016199
How hard did you have to work to accomplish?	4.44	3.96	10.81	.270243
How insecure, discouraged, irritated, stressed, and annoyed were you?	3.32	2.40	27.71	.071443

than 31). Since the subject sample was small and the power to detect an influence was low, a p 0.10 level was considered significant [15, 16]. Each hypothesis was evaluated on a number of factors which contribute to the examination in each dimension as discussed below.

4. Study Results

The MCSI was compared with CSI using an implicit evaluation method examining: cognitive dimensions and usability.

4.1. Cognitive dimensions

CSIs impose a significant cognitive load on the searcher [17]. An important factor in the evaluation of interactive systems is measurement of the cognitive load experienced by users while using the system. To measure the user's workload we adopted the NASA Ames Research Centre proposed the NASA Task Load Index [7, 6]. In terms of cognitive load, the user was asked to evaluate the CSI and MCSI in 6 dimensions [7] as shown in Table 1.

HO: Users experience a similar task load during the search with multiple interfaces: The user evaluated the system based on six parameters (Table 1). The grading scale lies between 0 (Low) - 7 (High). We compared the mean difference of both systems on all six parameters. In all aspects, subjects experienced lower task load using the MCSI. Subjects claimed more success in accomplishing the task using the MCSI. Results for accomplishing the task with the MCSI were found to be statistically significantly different. Subjects felt less insecure, discouraged, irritated, stressed, and annoyed, while using the MCSI with a significant difference ($P<0.10$). This implies the null hypothesis was rejected on the basis of the Task Load index. Although the four-factors were not significantly different, the mean difference between both the systems on these factors was more than 10%. We conclude that the user experienced less subjective mental workload while using the MCSI.

Table 2
Post Study System Usability Questionnaire (PSSUQ).

Topic	CSI Mean	MCSI Mean	Percentage Change	P value
Easy to use*	4.04	5.96	47.52	.000059
Simple to use	4.48	5.92	32.14	.003526
Effectively complete my work*	3.92	5.64	43.88	.000226
Quickly complete my work*	3.72	5.76	54.84	.00003
Efficiently complete my work*	3.88	5.76	48.45	.000045
Comfortable using this system*	4.16	5.88	41.35	.000471
Whenever I make a mistake using the system, I recover easily and quickly*	4.04	5.44	34.65	.006827
The information is clear*	4.16	5.92	42.31	.000072
It is easy to find the information I needed*	4.00	5.48	37	.000706
The information is effective in helping me complete the tasks and scenarios*	4.20	5.68	35.24	.000675.
The organization of information on the system screens is clear*	4.44	5.92	33.33	.000184
The interface of this system is pleasant*	4.28	6.08	42.06	.00002
Like using the interface*	4.20	6.12	45.71	.000014
This system has all the functions and capabilities I expect it to have*	4.08	5.72	40.2	.000168
Overall, I am satisfied with this system*	4.16	5.92	42.31	.000029

4.2. Usability

Usability is an important evaluation metric of interactive software. The IBM Computer Usability Satisfaction Questionnaires are a Psychometric Evaluation for software from the perspective of the user [8], and were used in this study. The grading scale lies between 0 (Low) - 7 (High). We compared the mean difference of both systems on all parameters. In all aspects, subjects experienced less task load when using the MCSI, as shown in Table 2.

H0: User Psychometric Evaluation for the conversational interface and conventional search has no significant difference: A T Independent test was conducted. It was found that for all the parameters the MCSI outperformed the CSI. The null hypothesis was rejected and the H1 hypothesis was accepted, which is that the MCSI performs better than the CSI.

5. Conclusions and Observations

We described a prototype conversational search system using agent-mediated search support to users, and compared this with an equivalent conventional entirely user-driven search interface. Our study indicates that subjects found our MCSI more helpful than the closely matched CSI. Most previous studies of user behaviour in conversational search have used Wizard-of-Oz type agents [5], in contrast, our study use of an automated search support agent.

We have also validated the current system in knowledge expansion, user interactive experi-

ence and search experience metrics which are not included in this prototype paper for reasons of space.

Clearly our existing rule-based search agent can be extended in terms of functionality, and going forward we aim to examine basing its functionality on machine learning and reinforcement learning based methods, but this will require access to sufficient suitable training data, which is not available at this prototype stage.

Acknowledgement

This work was supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) at Dublin City University.

References

[1] F. Radlinski, N. Craswell, A theoretical framework for conversational search, in: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, ACM, 2017, pp. 117–126.

[2] J. R. Trippas, D. Spina, L. Cavedon, M. Sanderson, How do people interact in conversational speech-only search tasks: A preliminary analysis, in: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR '17, ACM, New York, NY, USA, 2017, pp. 325–328. URL: http://doi.acm.org/10.1145/3020165.3022144. doi:10.1145/3020165.3022144.

[3] S. Avula, G. Chadwick, J. Arguello, R. Capra, Searchbots: User engagement with chatbots during collaborative search, in: Proceedings of the 2018 Conference on Human Information Interaction&Retrieval, ACM, 2018, pp. 52–61.

[4] P. Maes, Agents that reduce work and information overload, Communications of the ACM 37 (1994) 30–40. URL: http://doi.acm.org/10.1145/176789.176792. doi:10.1145/176789.176792.

[5] S. Avula, J. Arguello, Wizard of oz interface to study system initiative for conversational search, in: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, 2020, pp. 447–451.

[6] A. Kaushik, G. J. Jones, A conceptual framework for implicit evaluation of conversational search interfaces, Mixed-Initiative ConveRsatiOnal Systems workshop at ECIR 2021 (2021) 363–374.

[7] S. Hart, L. Staveland, Development of nasa-tlx (task load index): Results and theoretical research, human mental workload, 1988.

[8] J. R. Lewis, Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use, International Journal of Human-Computer Interaction 7 (1995) 57–78.

[9] A. Kaushik, V. Bhat Ramachandra, G. J. F. Jones, An interface for agent supported conversational search, in: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20, Association for Computing Machinery,

New York, NY, USA, 2020, p. 452–456. URL: https://doi.org/10.1145/3343413.3377942. doi:10.1145/3343413.3377942.

[10] P. Bailey, A. Moffat, F. Scholer, P. Thomas, UQV100: A test collection with query variability, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, ACM, New York, NY, USA, 2016, pp. 725–728. URL: http://doi.acm.org/10.1145/2911451.2914671. doi:10.1145/2911451.2914671.

[11] D. R. Krathwohl, A revision of bloom's taxonomy: An overview, Theory into practice 41 (2002) 212–218.

[12] A. Kaushik, G. J. F. Jones, Exploring current user web search behaviours in analysis tasks to be supported in conversational search, in: Second International Workshop on Conversational Approaches to Information Retrieval (CAIR'18), July 12, 2018, Ann Arbor Michigan, USA, 2018.

[13] J. V. Bradley, Complete counterbalancing of immediate sequential effects in a latin square design, Journal of the American Statistical Association 53 (1958) 525–528.

[14] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174.

[15] F. Hair Jr Joseph, C. Black William, J. Babin Barry, E. Anderson Rolph, Multivariate data analysis 7th ed, 2009.

[16] M. A. Hardy, Handbook of data analysis (2004).

[17] A. Kaushik, Dialogue-based information retrieval, in: European Conference on Information Retrieval, Springer, 2019, pp. 364–368.