Conversational Strategies:

Impact on Search Performance in a Goal-Oriented Task

Mateusz Dubiel
University of Strathclyde
Glasgow, Scotland, UK
mateusz.dubiel@strath.ac.uk

Martin Halvey
University of Strathclyde
Glasgow, Scotland, UK
martin.halvey@strath.ac.uk

Leif Azzopardi
University of Strathclyde
Glasgow, Scotland, UK
leif.azzopardi@strath.ac.uk

Damien Anderson
University of Strathclyde
Glasgow, Scotland, UK
damien.anderson@strath.ac.uk

Sylvain Daronnat
University of Strathclyde
Glasgow, Scotland, UK
sylvain.daronnat@strath.ac.uk

Figure 1: Conversational strategies employed by four conversational agents(CAs): PS (Passive Summary); PL (Passive Listing); AS (Active Summary) and AL (Active Listing). Followup questions employed by Active CAs are highlighted in bold.

ABSTRACT

Conversational search relies on an interactive, natural language exchange between a user, who has an information need, and a search system, which elicits and reveals information. Prior research posits that due to the non-persistent nature of speech, conversational agents (CAs) should support users in their search task by: (1) actively suggesting query reformulations, and (2) providing summaries of the available options. Currently, however, the majority of CAs are passive (i.e. lack interaction initiative) and respond by providing lists of results – consequently putting more cognitive strain on users.

To investigate the potential benefit of active search support and summarising search results, we performed a lab-based user study, where twenty-four participants undertook four goal-oriented search tasks (booking a flight). A 2x2 within subjects design was used where the CAs strategies varied with respect to elicitation (Passive vs. Active) and revealment (Listing vs. Summarising). Results show that when the CA's elicitation was Active, participant's task performance improved significantly. Confirming speculations that Active elicitation can lead to improved outcomes for end-users. A similar trend, though to the lesser extent, was observed for revealment – where Summarising results led to better performance than Listing them. These findings are the beginning of, but also highlight the need for, research into design and evaluation of conversational strategies that active or pro-active CAs should employ to support better search performance.

CAIR 2020, March 18, 2020, Vancouver, Canada

CCS CONCEPTS

• **Information systems** → **Search interfaces**; **Search interfaces**; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Conversational Search, Voice Interfaces, Interactive Study

ACM Reference Format:
Mateusz Dubiel, Martin Halvey, Leif Azzopardi, Damien Anderson, and Sylvain Daronnat. 2020. Conversational Strategies:: Impact on Search Performance in a Goal-Oriented Task. In *ACM CHIIR 3rd Conversational Approaches to Information Retrieval Workshop (CAIR), March 18, 2020, Vancouver, Canada.* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Conversational Agents (CAs) are systems that enable natural language interaction which is not constrained by menus, command prompts and key words [23]. In theory, CAs should give a user a feeling of collaborating with a virtual companion rather than using a system [21]. In practice, however, due to the ambiguous nature of human language and necessity for incremental interpretation of utterances in context, interaction with CAs can be a rather cumbersome and frustrating experience [12, 13, 19].

Currently, most commercially available CAs (e.g. Google Home, Amazon Echo etc.) lack the capacity for asking meaningful follow-up questions to refine the user's intent and, consequently, struggle with tasks that involve interpretation, judgment or opinion. Due to these limitations, CAs have been mostly used for simple goal-oriented tasks that require structured conversation characterised by predictable user input and small number of dialogue turns (cf. [2, 23, 27]). As such, most CAs tend to be based on an inflexible, passive interaction strategy which may leave users uncertain about possible options and functionalities [22]. However, for a CA to be more natural and usable, a CA needs to be able to correctly interpret a user's query, as well as present the information in ways that help the user achieve their goals [11]. With this premise in mind, rather than passively eliciting user information needs, and simply listing search results, this paper focuses on evaluating alternative interaction strategies. Specifically, we aim to explore the influence of: (1) two elicitation strategies (Passive vs. Active), (2) two revealment strategies (Listing vs. Summarising) and (3) the combination of elicitation and revealment strategies on the user's search performance and their search experience.

We hypothesise that:

- **H1:** An Active elicitation strategy will improve search performance compared to a Passive elicitation strategy.
- **H2:** A Summarising revealment strategy will improve search performance compared to a Listing revealment strategy.

We consider our hypotheses in the context of using a CA to search for flight options. In a series of interactive search tasks, we asked our participants to find flights that meet several competing criteria that require exploration of the search space (i.e. space of possible flights). We evaluate the search experience in terms of performance (i.e. whether the selected flight meets the provided criteria and satisfies requirement with regards to price and travel time). The above metric provides us with empirical insights into suitability and impact of different CAs for our goal-oriented tasks.

2 RELEVANT WORK

Over the past few years there has been a growing interest and resurgence in the design and development of conversational agents due to the maturation of speech and natural language processing technologies that facilitated their development. For example, with the recent advent of deep learning, we have seen a number of studies where neural ranking models were used to support interactive, multi-turn conversational search [1, 15, 16, 25, 38]. In this paper, however, rather than exploring the underlying mechanisms and automatic evaluation of conversational systems, we focused on the design of CAs and how the strategies that CAs employ impact and influence how people interact with CAs and how well people perform using them. With current CAs being rather passive in nature, emulating the query-response search paradigm of search engines, various information retrieval researchers [6, 26, 32] have been calling for a shift in paradigm to transform search engines from "passive query matchers" into "active search partners". Mixed-Conversational initiative (switching of initiative between user and the system) has been identified as a crucial prerequisite for making such a transition from a passive into an active agent. However, the transition poses several challenges relating to the design of conversational interfaces.

Studies on design of CAs and their reception by users have recently been attracting increasing attention (cf. [7, 8, 22, 31]]) and have led to development of guidelines and recommendations regarding the level of conversational initiative required by the agent. The guidelines provide generic suggestions on the design of CAs, e.g. "agents should inform user about their capacities" [22], "agents should use command and control for simple functional interactions" [20] etc. However, the suggestions fall short of providing details on how the conversation with the agent should be conducted (i.e. "what to say?", "when to say it?" and "how to say it?") or exploring its impact on user's search performance.

In terms of interaction with the user, two of the major challenges for developing conversational agents are: (1) Choosing the correct sequence of actions so as to help to resolve a user's information need [3, 35]. (2) Presenting search results effectively: i.e. "not overwhelming the users with information nor leaving them uncertain whether what they heard covered the information space" [33].

The above points are especially challenging in an audio channel (since users need to remember presented information and reason about it simultaneously cf. [36]). When information is presented verbally, lack of persistence makes speech easy to miss and forget since - "Almost everyone is quicker to absorb written text than speech" [37], and "Despite being easy to produce, speech is much more difficult to analyse" [28]. Performing a goal-oriented dialogue can also be considered in terms of a cost-benefit trade-off, where usefulness of a CA is determined by its ability to resolve a user's information need in a quick and comprehensive manner [3].

A problem with the current generation of CAs is that they provide information in a verbose way which puts strain on the user as they need to retain alternative options in their memory. For example, if a user asks about restaurants in the area, the CA will list a sequence of possible options, which the user will need to commit to memory (in order to compare or consider them later on). Consequently, due to cognitive overload, users tend to accept the first minimally acceptable option (satisficing behaviour) rather than continuing to absorb the cost of interaction in order to find a better option (maximising behaviour) [18]. On the other hand, users are unlikely to accept the CA's best suggested option without exploring alternatives [18]. Thus, a right balance needs to be struck between presenting enough options so that the user is satisfied and confident with their selection, and communicating the results in a manner which minimises the user's cognitive load.

Recently, several theoretical frameworks [26, 32, 34] have emerged to address the challenges of CA design. Trippas et al. [32] suggested that in Spoken Conversational Search (SCS) interaction, responsibilities should be shared between the user (who submits their information need) and the system (that actively decides which results to

present back to the user via audio channel). As postulated by Trippas et al. [ibid], system initiative is crucial in the auditory setting, where the effective transfer of information depends on an active exchange between the user and the system. Otherwise, with a passive system, there is a risk of overloading users with information. The role of the active system is thus to provide incremental responses and create a common ground for collaboration via interaction with the user. Similarly, Feldman [14], highlighted that a conversational system needs to actively support a user by: (1) efficiently navigating through search space, (2) offering hypothesis based on knowledge bases and ontologies and (3) tackling the complexities and ambiguities of human language. More recently, Vakulenko et al. [34] proposed QRFA (Query, Request, Feedback Answer) model. The model divides the conversational process into actions taken by User (who submits queries and provides feedback) and Agent (answers queries or requests additional information form user.) Although the above-mentioned theoretical frameworks [14, 26, 32, 34] acknowledge the importance of interactivity between the users and the system in the search task, they do not provide detailed information on how system should elicit and provide information during search task; which is the aim of this work.

3 CONVERSATIONAL STRATEGIES

Expanding on the theoretical framework of Radlinski and Craswell [26], Azzopardi et al. [3] proposed a more detailed conceptual framework, where the actions and responses of the user and CA are enumerated for each of the different properties. For example, agent revealment actions are to: list the results, summarize the results, and compare the results, while agent elicitation actions are to; extract specific details (i.e. slot-fill), clarify details, or not ask for further details and passively await requests from the user, The different actions (and their combinations) that the CAs can take, and how they implement these actions form the CAs strategy. Taken together, a CA can be described with respect to the approach that it takes when revealing and eliciting information. In terms of elicitation, they can be broadly categorised into three types:

- **Passive:** CA does not ask any questions - it leaves query refinement to the user.
- **Active:** CA asks questions to help a user make their query more specific.
- **Pro-Active:** CA suggests query reformulations that go beyond the scope of the original query and can thus change and/or expand the current information need.

While in terms of revealment, three different approaches can also be taken:

- **No Revealment:** CA does not provide any results (either because no results are available, or the agents may decide not to disclose any results, in lieu of another action, i.e. to elicit more details about the information need, instead).
- **List:** CA provides a detailed list with k elements: where k is the information retention threshold that varies between the users.
- **Summary:** CA aggregates different sets of options and then presents them to user in ranges.

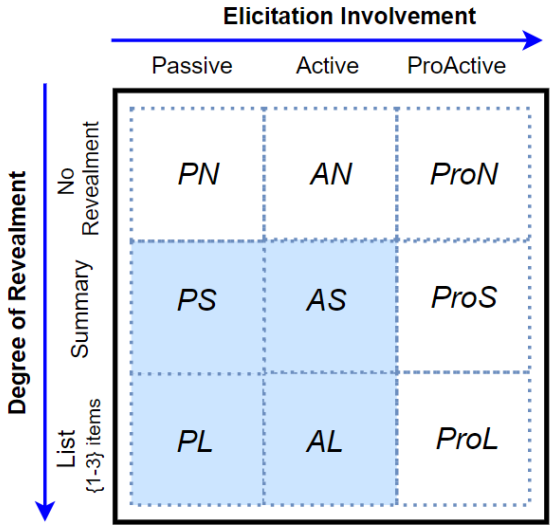

Figure 2: Spectrum of Conversational Agent Strategies. The agents' conversational involvement increases from left to right and top to bottom. In the current study we consider two agents that passively elicit information needs (PS and PL), and two agents that actively elicit information needs (AS and AL).

The illustration of conversational strategies is provided in Figure 1, while the spectrum of conversational strategies is presented in Figure 2 for reference.

In this work, we consider a CA's strategy as the combination of the type of elicitation approach and the type of revealment approach taken i.e. Passive/Active and List/Summary combinations. We will leave Pro-Active, No Revealment and Mixed approaches where CAs adopt a mixture of approaches for future work, and thus study the influence of taking a purely Passive vs. Active approach or Listing vs. Summary.

Present-day CAs tend to focus on well-defined, goal-oriented tasks for recurrent information needs (e.g. someone who already knows which flight/hotel they would like to book and are not concerned about exploring any alternatives etc.). Given the above classification, current CAs tend to be Passive (only asking a pre-defined set of questions to fill in slots) before Listing the results available. For example, on the Amazon Echo, KAYAK [17] and SkyScanner Fly Search [30] provide simple CAs. The KAYAK flight finder asks users a series of questions (slot filling), before providing a list of the cheapest available flights meeting those criteria, while the SkyScanner agents acts in a similar manner but reveals only the cheapest flight. Neither provide detailed information such as departure times, etc. nor provide any summary information such as ranges of prices, etc. Clearly, such agents lack support for exploratory search tasks, where users would like to explore and compare options (a phenomenon known as Comparative Shopping Notion [4]). Thus, in the current work, we will consider the context of flight booking where the user can explore a number of different options in order to select the best flight possible (given the search criteria).

ALGORITHM 1: Summarising Strategy - **Agents Angus & Calum**

```
while Flight has not been selected do
    wait for a query
    for a query do
        if n of flights > 2, then
            provide n of flights
            provide flights range (price)
            provide flights range (travel duration)
            Case 1. say: 'Would you like to filter by price?'
            Case 2. say: 'Would you like to filter by duration?'
            Case 3. say: 'Would you like to filter by departure time?'
        else
            provide detailed flight result(s) (airline, departure and arrival
              times, price and duration)
            say: 'Would you like to select this flight/any of these flights?'
        end
    end
end
```

ACTIVE (Case 1–3 say)

ACTIVE (say: 'Would you like to select this flight/any of these flights?')

ALGORITHM 2: Listing Strategy - **Agents Blair & David**

```
while Flight has not been selected do
    wait for a query
    for a query do
        if n of flights > 2, then
            provide n flights
            provide details of cheapest flight (price,duration)
            provide details of fastest flight (price, duration )
            Case 1. say: 'Would you like to filter by price?'
            Case 2. say: 'Would you like to filter by duration?'
            Case 3. say: 'Would you like to filter by departure time?'
        else
            provide detailed flight result(s) (airline, departure and arrival
              times, price and duration)
            say: 'Would you like to select this flight/any of these flights?'
        end
    end
end
```

ACTIVE (Case 1–3 say)

ACTIVE (say: 'Would you like to select this flight/any of these flights?')

4 METHOD

Our study was conduced using Wizard of Oz (WoZ) methodology [10], where participants interacted with a wizard (human subject) in order to find and book flights. Booking a flight is an exemplar goal-oriented task that allowed us to measure performance of participants and evaluate the adequacy and impact of the different conversational strategies. In the context of flight booking, numerous factors such as the cost of the flight, its duration, departure/arrival time, the airline, fare class etc. impact which flight is selected. For our tasks, we focus on the most salient variables: flight price, flight duration and arrival time (as identified by the IATA Global Passenger Survey [24]). In our experiment, participants were asked to imagine that they are a traveller looking for a one-way flight from Glasgow Airport. There were four search scenarios in total (one per conversational strategy).

We recruited a wizard with an extensive call centre experience who was used to conducting structured goal-oriented conversations. None of the principal investigators acted as the wizard to reduce any potential unconscious experimental bias, i.e. the possibility of carrying out the task in the way that supports research hypotheses. During search tasks, in response to participant's queries the wizard searched a flight database and provided results back to them verbally. The wizard could narrow down the pool of results by applying different filters (e.g. price up to £150) or by ordering them by attributes (e.g. flight duration, arrival time, etc.) For sake of consistency, wizard followed conversational strategies presented in Algorithm 1. (Summarising) and Algorithm 2. (Listing).

Each of the algorithms was implemented in a Passive and Active mode. In the study we assigned a name to each of the conversational agents to make them easily identifiable to the participants, the names of the agents were:

(1) **Agent Angus:** Passive Summarising
(2) **Agent Blair:** Passive Listing
(3) **Agent Calum:** Active Summarising
(4) **Agent David:** Active Listing

When designing the agents, we considered that the suitability of different information presentation methods may vary between different participants based on their ability to retain information. Commarford et al. [9] showed that people with short working memory spans prefer longer lists of options as providing less options within a conversational turn leads to a larger number of conversational turns. When using lists the number of presented elements needs to be capped to prevent overloading a participant's memory. In their study, Demberg et al. [11] limited the number of provided options to 3 at a time. Demberg et al. refrained from presenting complex information in a single conversational turn. Instead, each time that there were more than three flights in the results cluster, only attributes that distinguished the flights were presented to the user (e.g. "The three direct flights are on Continental, Lufthansa, and Delta. They arrive at 9:55 am, 10:15 am, and 11:15 am"). In our study, in order to avoid overloading the working memory of our participants, we have taken a slightly more conservative approach than Demberg et al. [11] and limited the number of presented options to the maximum of two flights at a time. This decision was taken to facilitate retention of information for the broader spectrum of participants In order to make the interaction more realistic, and to exclude the impact of body language on communication, the wizard and participant were separated by a barrier and were not visible to each other.

An example search scenario is presented in Figure 3. Participants were instructed to explore the available flight options over three days and to find the shortest and cheapest flight possible. The rationale was that when searching for flights most people prefer to get to their destination as quickly as possible for the least amount of money. We measured task performance in terms of distance of the selected flight from the Pareto frontier illustrated in Figure 4 - hard constraint, and in terms of preference (desired travel time specified in search scenario) - soft constraint. For instance, if a participant booked a flight for 100 pounds with the 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

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 speaking before you start to speak

Figure 3: Search Scenario. The destination, time of travel of the required flight and participants' instructions are provided in bold for emphasis. During the experiment, participants were provided with a printout of a search scenario so that they do not have to memorise the search instructions.

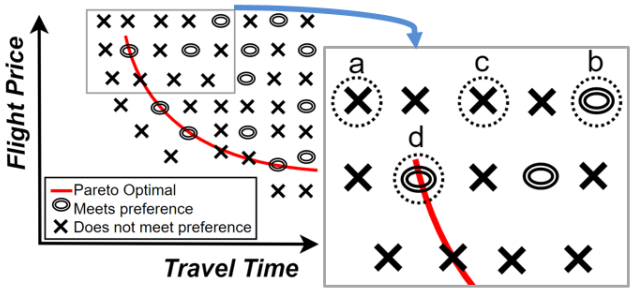

Figure 4: Trade-off between cost and travel time (left) and a close-up with example flight selections (right). All selected flights (a, b, c, d) are considered in reference to the closest flight on Pareto Optimal: a – is an option that wastes time and money, b – is an option that wastes time and money but meets arrival preference, c – is an option that save time but wastes money and d – is an optimal option.

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.

During search sessions Passive agents (Angus and Blair) did not make any suggestions with regards to results filtering while Active agents (Calum and David) asked participant questions to progressively narrow down the list of flight options. Questions 8-10 (See Algorithm 1 and Algorithm 2) were always presented in chronological order, unless a participant made a specific request (e.g. "Tell me about the earliest flight on Monday.") If a query returned no results (∅), the agent would suggest changing filtering criteria. For instance, if there were no flights within the specified price range, the agent would suggest relaxing search criteria (e.g. "There are no flights for less than £200 that leave before 2 pm, try to increase your price, change travel time" etc.). If the query issued by the participant was outwith the scope of the agent, it would inform the user that the given functionality is not supported. The agent could search only one day at the time so that participants had to perform a number of searches to explore the space (and is consistent with flight booking systems.) By using these algorithms to dictate agent's interactions, consistency across participants was ensured.

Task performance measures are presented in Table 1 and Table 2. We consider Task Performance in terms of Primary Indicators ('Meeting Time Preference' and 'Hitting Pareto Optimal), and Secondary Indicators ('Losing Money' and 'Wasting Time'). Primary Indicators concern meeting flight arrival requirements specified in each search scenario and indicate if the selected flight was Pareto Optimal (i.e. it offered the best combination of price and duration while meeting the arrival requirement). Secondary Indicators concern the impact of participant's flight selections on their resources, i.e. time and money (i.e. if participant selected a more expensive flight when a cheaper one of the same duration was available - they lost money; if they selected a flight with longer duration when a shorter alternative was available at same price – they lost time).

5 RESULTS

Since all of our performance indicators are considered in binary categories, i.e. a selected flight is either on the Pareto frontier or not, the selected flight either meets the time preference or does not, Cochran's Q Test [29] was used to compare different conversational strategies. Bonferroni adjusted α-level (.008) was used for all post-hoc analyses.

Meeting Time Preference: There is a statistically significant difference between the conversational strategies (Cochran Q =23.368, p <.001). For pair-wise comparisons, McNemar post-hoc test indicated a statistically significant difference between Active and Passive agents (p <.001) but not between Summary and Listing agents (p =.79).

Hitting Pareto Optimal: There is no statistically significant difference between strategies (Cochran Q = 6.667, p =.83). There is, however, a noticeable difference between the Active and Passive agents for booking flights on Pareto optimal, 5 and 13 respectively (p =.077). The difference is less pronounced between Summary and Listing strategies, 11 to 7 respectively (p =.388).

Time Wasted: We observe a statistically significant difference between the conversational strategies (Cochran Q = 13, p =.005). Post-hoc test indicated a statistically significant difference between Active and Passive agents (p =.004) but not between Summary and Listing agents (p =.219).

Money Lost: No statistically significant differences were observed between the conversational strategies (Cochran Q = 4.286, p =.232). Pair-wise comparisons indicate that there is little difference between Active and Passive agents (p =.625) and between Summary and Listing agents (p =.219).

Overall, in terms of performance, Active conversational strategy consistently outperforms Passive conversational strategy for all performance aspects under consideration. An analogical trend can be observed for Summarising strategy that outperforms Listing strategy for all aspects but 'Meeting Time Preference'. At the level of individual CA, Active Summary yields the best performance for both Primary and Secondary performance indicators. The most

Table 1: Performance Measures (Primary Indicators): ** signifies p < .001.

	Passive	Active	Summarising.	Listing	Passive Summarising	Passive Listing	Active Summarising	Active Listing
Met Time	28/48	**40/48****	33/48	35/48	13/24	15/24	20/24	20/24
Preference	(58%)	(83%)	(69%)	(72%)	(54%)	(63%)	(83%)	(83%)
Pareto	5/48	13/48	11/48	7/48	3/24	2/24	8/24	5/24
Optimal	(10%)	(27%)	(23%)	(15%)	(13%)	(8%)	(33%)	(21%)

Table 2: Performance Measures (Secondary Indicators): ** signifies p < .001. The lower score indicates better performance.

	Passive	Active	Summarising	Listing	Passive Summarising	Passive Listing	Active Summarising	Active Listing
Money	11/48	9/48	8/48	12/48	6/24	5/24	2/24	7/24
Lost	(23%)	(19%)	(17%)	(25%)	(25%)	(21%)	(8%)	(29%)
Time	27/48	**18/48****	21/48	24/48	13/24	14/24	8/24	10/24
Wasted	(56%)	(38%)	(44%)	(50%)	(54%)	(58%)	(33%)	(41%)

notable differences are observed when it comes to 'Hitting Pareto Optimal' and 'Money Lost'.

For elicitation strategies, we observe that while using Active agents, participants were significantly more likely to select a flight that meets the preferred arrival time and wasted significantly less travel time in the process. We also observe that participants who used Active agents selected more Pareto optimal flights and lost less money overall (see Table 1 and Table 2 for full results). This finding supports our H1 with regards to search performance and provides an empirical evidence that validates the assertion (advocated in previous research [6, 14, 26] that active involvement of CA can boost search performance. With regards to revealment strategies, Summarising CAs outperform Listing CAs with all aspects but Meeting Time Preference. While using Summarising CAs, participants selected more Pareto optimal flights and wasted less money. However, the differences did not reach statistical significance and therefore we cannot support our H2.

6 LIMITATIONS AND FUTURE WORK

In the present study, we have only considered one specific context in which a CA may be used i.e. goal directed search where the user needs to select one item among many in order to satisfy several constraints – and through a voice only channel. While limited, this context is more generally applicable to other related search tasks, and provides a controlled domain in which we can evaluate more precisely how the conversational strategies impact on user experience. We also restricted how the CAs behaved by having them adopt one elicitation approach and one revealment approach - and thus adopt pure strategies (i.e. strategies that rely on one form of elicitation and revealment).

We acknowledge that there are other approaches that could have been taken, and that the strategies could to be varied to adapt to the context and user. In the current study, our interest is understanding how these pure strategies impact task performance, and leave mixed strategies for future work.

Another limitation of this work that needs to be acknowledged is learning effect involved in working with CAs. In current work, we tried to control for learning effects by using a Latin Square [5] rotation of tasks and agents. It is worth noting that we did not observe any significant differences in behaviour or performance stemming from the ordering. However, we do acknowledge that further investigation is required to understand how quickly participants can learn to efficiently and effectively use CAs to perform tasks. Thus, we plan to explore different conversational strategies employed by participants to check if their level of expertise and exposure to conversational agents impact their preferred CA strategies.

7 CONCLUSIONS

To the best of our knowledge, our study is the first empirical investigation of different elicitation and revealment strategies for voice based search. We propose a performance-based evaluation metric and provide insights on how users would interact with different audio-only CAs in a multi-turn, goal-oriented task, when trying to fulfil competing search criteria.

While we have only examined the impact of conversational strategies within a specific context (flight booking), we have shown that the strategy does impact user performance, and we believe that it is likely to be similar in other contexts such as product/service search scenarios. That being said, our findings on how to elicit and present information can help to pave way to active search support and development of more useable goal-oriented CAs in the future.

8 ACKNOWLEDGMENTS

We would like to thank all of the participants who took part in our study. We also thank anonymous reviewers for their helpful comments and suggestions which helped us to improve this article.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 475–484.

[2] Pierre Andrews and Silvia Quarteroni. 2011. Extending Conversational Agents for Task-Oriented Human-Computer Dialogue. In *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. IGI Global, 177–202.

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

[4] Enrique Bigné, Joaquin Aldás, and Antonio Hyder. 2015. Engagement with Travel Web Sites and the Influence of Online Comparative Behaviour. In *Cultural Perspectives in a Global Marketplace*. Springer, 26–33.

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

[6] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 815–824.

[7] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2018. The State of Speech in HCI: Trends, Themes and Challenges. *arXiv preprint arXiv:1810.06828* (2018).

[8] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 475.

[9] Patrick M Commarford, James R Lewis, Janan Al-Awar Smither, and Marc D Gentzler. 2008. A comparison of broad versus deep auditory menu structures. *Human Factors* 50, 1 (2008), 77–89.

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

[11] Vera Demberg, Andi Winterboer, and Johanna D Moore. 2011. A strategy for information presentation in spoken dialog systems. *Computational Linguistics* 37, 3 (2011), 489–539.

[12] Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. 2018. A Survey Investigating Usage of Virtual Personal Assistants. *arXiv preprint arXiv:1807.04606* (2018).

[13] 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*.

[14] Susan E Feldman. 2012. The answer machine. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 4, 3 (2012), 1–137.

[15] Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational AI. *Foundations and Trends® in Information Retrieval* 13, 2-3 (2019), 127–298.

[16] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2019. A deep look into neural ranking models for information retrieval. *arXiv preprint arXiv:1903.06902* (2019).

[17] KAYAK. 2018. Kayak Skill (Amazon Echo application software]). Acessed: 10th September 2019, Retrieved from: https://www.amazon.co.uk/KAYAK/dp/B01EILLOXI.

[18] 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).

[19] 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.

[20] Jesse Mu and Advait Sarkar. 2019. Do We Need Natural Language?: Exploring Restricted Language Interfaces for Complex Domains.. In *CHI Extended Abstracts*.

[21] Nicole Novielli and Carlo Strapparava. 2011. Dialogue act classification exploiting lexical semantics. In *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. IGI Global, 80–106.

[22] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences*. " O'Reilly Media, Inc.".

[23] Diana Perez-Marin. 2011. *Conversational agents and natural language interaction: Techniques and effective practices: Techniques and effective practices*. IGI Global.

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

[25] Chen Qu, Liu Yang, W Bruce Croft, Yongfeng Zhang, Johanne R Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, 25–33.

[26] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. ACM, 117–126.

[27] Gisela Reyes-Cruz, Joel Fischer, and Stuart Reeves. 2019. An ethnographic study of visual impairments for voice user interface design. *arXiv preprint arXiv:1904.06123* (2019).

[28] Christopher Schmandt. 1994. *Voice communication with computers: conversational systems*. Van Nostrand Reinhold Co.

[29] Siegel Sidney. 1957. Nonparametric statistics for the behavioral sciences. *The Journal of Nervous and Mental Disease* 125, 3 (1957), 497.

[30] Skyscanner.net. 2018. Skyscanner Flight Search (Amazon Echo application software]). Accessed on 10th September, Retrieved from: https://tinyurl.com/y8fx2n3o.

[31] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 603.

[32] 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 2018 Conference on Human Information Interaction & Retrieval*. ACM, 32–41.

[33] 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.

[34] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A data-driven model of information-seeking dialogues. In *European Conference on Information Retrieval*. Springer, 541–557.

[35] Nigel G Ward and David DeVault. 2015. Ten challenges in highly-interactive dialog system. In *2015 AAAI Spring Symposium Series*.

[36] 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.

[37] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in speech user interfaces. In *CHI*, Vol. 95. 369–376.

[38] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 177–186.