How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis

Johanne R. Trippas
johanne.trippas@rmit.edu.au

Damiano Spina
damiano.spina@rmit.edu.au

Lawrence Cavedon
lawrence.cavedon@rmit.edu.au

Mark Sanderson
mark.sanderson@rmit.edu.au

RMIT University, Melbourne, Australia

ABSTRACT

We present preliminary findings from a study of mixed initiative conversational behaviour for informational search in an acoustic setting. The aim of the observational study is to reveal insights into how users would conduct searches over voice where a screen is absent but where users are able to converse interactively with the search system. We conducted a laboratory-based observational study of 13 pairs of participants each completing three search tasks with different cognitive complexity levels. The communication between the pairs was analyzed for interaction patterns used in the search process. This setup mimics the situation of a user interacting with a search system via a speech-only interface.

Keywords

Spoken Conversational Search; Observational Study; Voice Search

1. INTRODUCTION

Speech-based web search is becoming ubiquitous, particularly through the use of mobile devices. However, presenting lists of search results in a speech-only setting presents a number of challenges and simply speaking the textual component of a standard search results list has been shown to be ineffective [7]. Intelligent Assistants such as Siri, Google Now, or Cortana can reply to factoid queries. However, when non-factoid queries are posed the search engine result page (SERP) is still displayed on the screen.

In this paper, we present preliminary results of an empirical laboratory study designed to understand how users search in a setting where all communication is over speech. This study observed pairs of participants speaking in order to accomplish a search goal, allowing us to understand the users' conversational patterns. We used quantitative and qualitative research designs for our analysis. The purpose of this observational study was to explore techniques used by people to search over a speech-only communication setting. Thus this paper presents initial identified interaction themes which inform a coding scheme for user-system conversational speech-only search. These themes contribute to our broader goal which is to understand effective response generation techniques for search tasks with different degrees of complexity. Overall we are interested in how spoken interactive information retrieval (IIR) conversations develop over multiple turns and whether there are recurring patterns in response generating techniques depending on task difficulty.

CHIIR '17, March 07-11, 2017, Oslo, Norway

DOI: http://dx.doi.org/10.1145/3020165.3022144

2. METHODOLOGY

A controlled laboratory study was conducted in June 2016 with 30 people of whom 70% were university students (recruited via the RMIT University Behavioural Business Lab, https://orsee.bf.rmit.edu.au). The participants were divided into pairs with two pairs used in the pilot studies (not included in this analysis). The setup was reviewed and approved by RMIT University's Ethics Board (ASEHAPP 08-16). Thus 13 pairs conducted a search where one participant acted as the *User* (participant with the search task) and the other participant acted as the *Retriever* (participant with the search engine); i.e. the User acted as the searcher and Retriever simulated the voice interface. Users and Retrievers did not have access to each others' search task or search engine interface, could not see each other, and could communicate only verbally. Each pair completed three search tasks with different dimensions of Anderson and Krathwol's Taxonomy of Learning [1]. There were nine search tasks in total, rotated using a Latin square design. Retrievers were navigated to use Google but were not stopped if they changed search engine. Retrievers were instructed to type in exactly what they heard Users instruct them to search for.

Each search ended when Users believed they had enough information to answer the backstory or when a 10-minute limit was reached. All participants completed three questionnaires (Pre-test, Post-task and Exit), and an exit interview. Users also completed a Pre-task questionnaire. The complete task took no longer than 90 minutes. No sample search task solution was given to avoid biasing the results.

2.1 Tasks

We describe an evaluation of nine search tasks based on the cognitive complexity framework of the Taxonomy of Learning [1]. The search tasks used in this study were based on TREC Q02, R03, and T04 and described in [2]. The following three of the five cognitive dimensions were used: *Remember* (Retrieving, recognizing, and recalling relevant knowledge from long-term memory); *Understand* (Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining); and *Analyze* (Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differenti-

Table 1: Examples of task complexity [1] with corresponding title and sample backstory [2] and example user queries.

Complexity	Title (1) and Sample Backstory (2)	Example Queries Submitted by Users
Remember	(1) Where does cinnamon come from? (2) The other day you were eating some spiced biscuits from Europe, when it occurred to you that cinnamon probably isn't native to that part of the world. You would like to know where it comes from.	-Where does cinnamon come from -Which country does cinnamon which country grows the most cinnamon
Understand	(1) Recycle automobile tires (2) You need to buy new tires for your car, and the local dealer has offered to take the old ones for recycling. You didn't know tires could be recycled and you wonder what new uses they are being put to.	-Can you just type in tire uhm... car tire recycling -Uses for old car then the query or... passenger vehicle tyres... (user spells tyres) or... in caps tires... (user spells tires) and I wanna uhm... do a date-range so the data is from a recent 12 months... so uses for old car caps or passenger vehicle or tyres (user spells tyres) caps or tires (user spells tires) and data in the last 12 months that's the query ▼
Analyze	(1) Per capita alcohol consumption (2) You recently attended a big party and woke up with a hangover, and have decided to learn more about the average consumption of alcohol. You are particularly interested in any information that reports per capita consumption, and want to compare across groups, for example at the country, state, or province level.	-What's the average alcohol consumption of an Australian -OK so... uhm... in general I sort of want to try to find out the average consumption of alcohol uhm by... by the country state... for local level... so maybe just start of with type in uhm alcohol consumption by country... tell me if anything related to statistics uhm... about alcohol consumption uhm... between countries ▲

▼denotes teleporting and ▲denotes query babbling (see Section 3.1.1)

ating, organizing, and attributing). The dimensions are ranked in task complexity from Remember (least complex) to Analyze (most complex). Table 1 presents title and backstory examples. Example queries spoken by Users are shown in the last column.

2.2 Annotation and Analysis

Users, Retrievers, and the Retriever's screen were video recorded. The recordings were synchronized and merged. The annotations and transcriptions were created in ELAN.[1] Recordings were analyzed using Thematic analysis to generate themes [3]; search and interaction behaviours were analyzed in terms of words used and time on task. The transcriptions were changed to lower case, and punctuation and extra spacing was removed. The fill-word "uhm" was also removed for analysis purposes. However, we deliberately did not remove any errors, false starts or confirmations since these will likely occur in real case voice search scenarios.[2]

In the context of mixed initiative information retrieval dialogues, researchers have used the *control* and *initiative* terms interchangeably. However, we use the approach of *taking the initiative equals taking the turn* [5]. This means that one turn can consist of multiple moves or communication goals. An annotation schema relevant to our research aims was designed after inspecting the data; the complete data set was then coded using these data-derived *codes* [3]. Thus codes were applied to each turn taken by either User or Retriever and these codes were collated to create *themes*. Themes can consist of *subthemes* which capture specific concepts of that theme [3]. Themes were created independently of the previous turn meaning that each turn may consist of similar themes or subthemes. The first and second authors coded the data set independently. Inter-rater reliability was very high (Cohen's κ =0.88). Thirteen pairs provides data saturation, given that in thematic analysis saturation has been found to occur within 12 interviews [4]. We report our early findings of this inductive analysis of themes and patterns using the identified codes.

[1] http://tla.mpi.nl/tools/tla-tools/elan/

[2] Please contact the first author regarding access to the data.

3. RESULTS

Figure 1 shows the themes identified by the analysis of the coded videos. Turn 1 consists of the initial Information Request from the Users. We have grouped the response generating strategies of Turn 2 into the following themes:

Meta-communication Theme: In this theme Retrievers engaged in communication about the Information Request with Users before accessing the SERP, representing 59% of Turn 2 interactions. This *Meta-communication theme* included the following subthemes: Retrievers ask Users to repeat the Information Request (*Asks to repeat*, 28%); Retrievers formulate the desire for a refinement of the Information Request (*Query Refinement Offer*, 23%); or Retrievers confirm their action (*Confirms*, 8%). The *Query Refinement Offer* subtheme consisted of five different codes: multiple/single explicit or multiple/single implicit query suggestions or Information Request paraphrasing. The majority of *Query Refinement Offer*s (75%) were asked before Retrievers accessed the SERP. The *Ask to Repeat* subtheme consisted of utterances where Retrievers specified their desire for an Information Request repeat explicitly ("sorry say that again" or "can you repeat that please") or implicitly (hesitantly started to repeat the Information Request back to the User or informed the User about their predicament without explicitly asking for a repetition of the Information Request). This subtheme also captured the Retrievers' "yes" utterances (*Confirms*).

Search Engine Result Page (SERP) Theme: This theme presented 38.5% of Turn 2 interactions and included three subthemes: Retrievers presented the SERP with/without the source (*SERP Presentation without Modification*, 20.5%); the SERP was synthesized/an overview was given (*SERP Presentation with Modification*, 13%); and SERP overview was given with further search suggestions (*SERP Presentation with Modification and Suggestion*, 5%).

Scanning Document Theme: This theme represented 2.5% of Turn 2 and consists of Retrievers neither initiating meta-communication nor engaging in any SERP presentation activities; instead Retrievers directly accessed a document and presented this to the user.

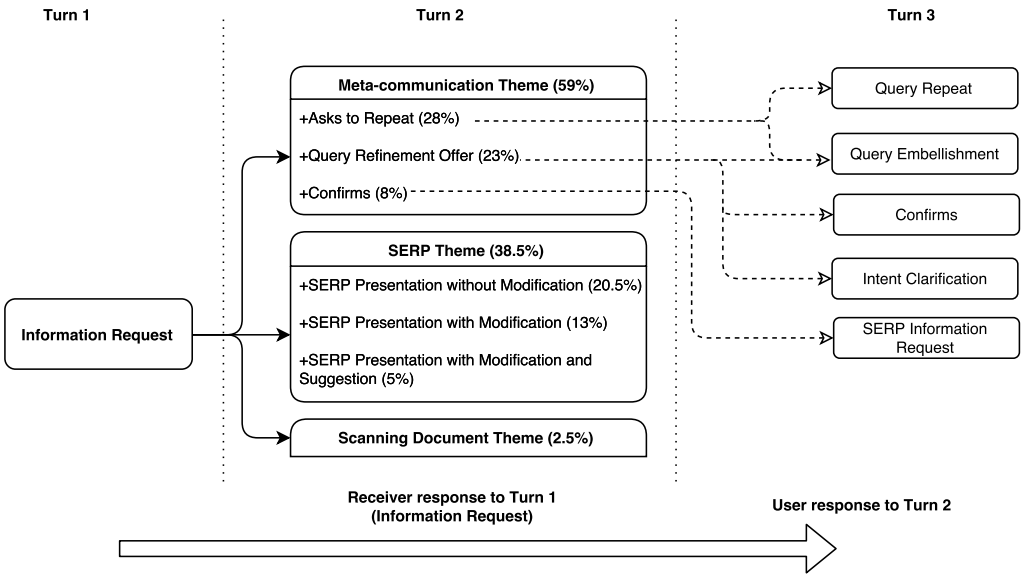

Figure 1: Interaction Theme Map (First three turns).

3.1 Search and Interaction Behaviours

The average length of Information Requests was 14.97 words. However, this length differed depending on the task complexity. As shown in Table 2, the average length of Remember Information Requests was shorter than for the Understand and Analyze categories. However, when non-query related words were stripped from the Information Request there was a higher number of query words in the Remember category. Simultaneously, the average time on search tasks for Remember tasks was nearly half the time as for Understand and Analyze tasks. The table shows a trend to longer Information Requests and time on the search task as the task complexity increases.

Finally, Users spent an average of 7.7 seconds in one turn and Retrievers 14.3 seconds. Table 2 indicates that Retrievers spent less time on Remember tasks than Understand and Analyze tasks. Even though participants received no instructions as to which search engines could be used, several participants were observed accessing Google Scholar. These search engine switches were mostly observed in Understand and Analyze search tasks.

3.1.1 Information Request Length

The average number of query terms in all unique queries issued for a task differed depending on the level of task difficulty. We observed two techniques which contributed to the length of the Information Requests. Firstly, we observed Users adopting a technique whereby they did not specify their query but instead discussed their information need. An example query is shown in Table 1 with the ▲notation. Secondly, some Information Requests were submitted using a very detailed and carefully crafted structure in order to capture the complete information request in one turn. An example is shown in Table 1 with the ▼notation. Users who submitted such Information Requests typically received a full SERP list from the Retrievers as search result. Users who started the Information Request with either of these two approaches were more likely to continue with this approach throughout that particular search task.

3.1.2 Information Request Similarity

Information Request similarity was measured after stopword removal and stemming between the overlapping Users' Information Requests words (i.e. query words) and backstory titles. The micro-averaged cosine similarity measure was used and these scores were aggregated according to the task complexity of Remember, Understand and Analyze. The similarity test scores for three types of queries were analyzed using one-way analysis of variance (ANOVA). Test assumptions of normality and homogeneity of variance were satisfactory, and the result was not statistically significant.

Table 2: Search and Interaction Behaviours

Measure	Task complexity		
	Remember	Understand	Analyze
Inf. Req. Length in Words	11.92	16.25	16.77
Avg. Amount of Query Words	3.15	0.84	2.2
Avg. Query Word Similarity	0.88	0.817	0.819
Avg. Time on Task	3.57 min.	7.09 min.	7.34 min.
Avg. User Time on Task	7.20 sec.	7.93 sec.	8.01 sec.
Avg. Retriever Time on Task	9.14 sec.	17.16 sec.	16.69 sec.
Meta-communication Theme	6	8	9
SERP Theme	7	5	3
Scanning Document Theme	0	0	1
Query Refinement Offer Subtheme	1	4	4
Task Stopped by User	11	7	6

3.2 Meta-communication User Responses

As mentioned in Section 3, Retrievers deployed three different response generating strategies on the initial Information Request which we divided into the following themes: *Meta-communication*,

SERP, and *Scanning Document*. The *Meta-communication* theme was most frequently used and therefore is discussed in more detail in this preliminary analysis. Retrievers refined the Information Request within the *Query Refinement Offer* subtheme. Users replied to the *Query Refinement Offer* with one of the following three types of responses: *Confirms* (confirming the proposed *Query Refinement Offer*); *Intent Clarification* (replying to clarify the search intent); or *Query Embellishment* (enrichment of the given query, for example by adding query words).

The *Ask to Repeat* subtheme shared the *Query Embellishment* response with the *Query Refinement Offer* subtheme. However, when Retrievers asked users to repeat their Information Request we noted that Users mostly replied with a *Query Repeat* (a repetition of their Information Request without enrichments).

The *Confirms* subtheme received responses from users such as "so what's the first result that you get..." and "what kind of results come up" and were recorded as *SERP Information Request*.

An example of a *Query Refinement Offer* with *Query Clarification* is shown in Figure 2.

Turn 1: USER	(Information Request)	Effectiveness of new security measures at airports
Turn 2: RETRIEVER	(Query Refinement Offer)	Australia or at the airport
Turn 3: USER	(Intent Clarification)	Put uhm... put international airports

Figure 2: Initial Conversation Example.

4. DISCUSSION & CONCLUSION

We presented on a study designed to explore techniques used in search over a speech-only communication channel. An annotation schema was designed and the observed techniques were classified into three themes: Meta-communication, SERP, and Scanning Document.

Retrievers accessed the *Meta-communication* theme more often as task complexity increased; they specifically asked for more query clarifications. Retrievers presented the SERP without engaging in meta-communication more often as task complexity decreased. We are currently investigating whether differences exist between the Information Requests which received a response from the *Meta-communication* or *SERP* theme. This would allow us to understand when it would be beneficial to engage the user in either of these themes.

The *Query Refinement Offer* subtheme from the *Meta-communication* theme consisted of five response generating techniques: multiple/single explicit or multiple/single implicit Information Request refinements, and paraphrasing of Information Requests. These response generating techniques may be comparable to query disambiguation and refinement techniques such as query auto-completion or expansion techniques in a textual setting.

The subtheme *SERP Presentation with Modification and Suggestion* allowed Retrievers to scan a SERP and synthesize multiple results presenting an overview of that SERP. This technique may prevent users from information overload by filtering out the important facts of multiple results. A similar technique was used by Retrievers synthesizing and aggregating information from documents in order to group and present common information.

Overall, Information Requests became longer in words as the task complexity increased. However, even though Remember Information Requests were the shortest in length, they contained more query words relevant to the backstory title. Nevertheless, the overall similarity of the Information Request query words was similar to the backstory title for all three levels of task complexity. The increase in query words in the Remember tasks may be due to the tasks requiring factoid-style queries [2].

Two techniques which contributed to the Information Request length were observed. Firstly, Users submitted Information Requests by talking about their information need. This technique is similar to *querying by babbling* [6]. Secondly, users submitted very detailed and carefully crafted Information Requests and resembles *teleporting*, i.e. trying to jump directly to the information target [8]. Interestingly, teleporting is also observed among people who use screen readers in order to directly access relevant information without having to go through search results [7].

The average time on task became longer as the task complexity increased. More tasks were stopped by Users in Remember tasks while Understand and Analyze tasks were more likely to continue until the time limit was met.

There are no clear signs that Understand and Analyze tasks resulted in different patterns to each other. However, Remember tasks received enough information to satisfy the Information Request in a much shorter time frame than the Understand and Analyze tasks. The Remember tasks also involved less meta-communication in the first two interactions about the task in order to complete the task.

Independently of task complexity, User–Retriever pairs did not utilize the same response generating themes in each task; that is, pairs utilized a combination of *Meta-communication* or *SERP* themes. No observations of Information Request abandonment were made; however it would be interesting to see whether this is also the case in a non-laboratory setting. We plan to investigate if the guidance of Retrievers or system may lower the abandonment rate.

This initial analysis of search conversations between two people was performed in order to explore conversational search patterns in a speech-only setting. The interaction themes and coding scheme for a conversational speech-only search were identified uncovering two main response generating themes, *Meta-communication* and *SERP*.

As future work, we will test and validate the response generating techniques observed in these themes with Wizard of Oz and crowdsourcing experiments. We will also continue coding the interactions and extend this study. The coding scheme could be used for future crowdsourced classification and for validation of similar experiments. We are ultimately interested in understanding conversation designing effective strategies in a spoken IIR setting.

5. REFERENCES

[1] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, New York, 2001.

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and IR system evaluation. In *Proc. of SIGIR'15*, pages 625–634, 2015.

[3] V. Braun and V. Clarke. *Successful Qualitative Research: A Practical Guide for Beginners*. Sage, 2013.

[4] G. Guest, A. Bunce, and L. Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field Methods*, 18(1):59–82, 2006.

[5] E. Hagen. An approach to mixed initiative spoken information retrieval dialogue. In *Computational Models of Mixed-Initiative Interaction*, pages 351–397. Springer, 1999.

[6] D. W. Oard. Query by babbling: A research agenda. In *Proc. of CIKM'12*, pages 17–21, 2012.

[7] N. G. Sahib, A. Tombros, and T. Stockman. A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *JASIST*, 63(2):377–391, 2012.

[8] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proc. of SIGCHI'04*, pages 415–422, 2004.