Clarifying False Memories in Voice-based Search

Johannes Kiesel
Bauhaus-Universität Weimar
Weimar, Germany
johannes.kiesel@uni-weimar.de

Arefeh Bahrami
Bauhaus-Universität Weimar
Weimar, Germany
arefeh.bahrami@uni-weimar.de

Benno Stein
Bauhaus-Universität Weimar
Weimar, Germany
benno.stein@uni-weimar.de

Avishek Anand
L3S Research Center
Hannover, Germany
anand@l3s.de

Matthias Hagen
Martin-Luther-Universität
Halle-Wittenberg
Halle, Germany
matthias.hagen@informatik.
uni-halle.de

ABSTRACT

Queries containing false memories (i.e., attributes the user misremembered about a searched item) represent a challenge for search systems. A query with a false memory will match inadequate results or even no result, and an automatic query correction is necessary to satisfy the user expectations. For voice-based search interfaces, which aim at a natural, dialog-based search experience, a sensible answer to this kind of unintentionally ill-posed queries is even more crucial. However, the usual solutions in display-based interfaces for queries without matches (e.g., suggesting to drop some query terms) cannot really be transferred to the voice-based setting. Based on the assumption that false memory queries *could* be identified—a research problem in its own right—, we present the first user study on how voice-based search systems may communicate the respective corrections to a user. Our study compares the user satisfaction in a voice-based search setting for three kinds of false memory clarifications and a baseline case where the system just answers "I don't know." Our findings suggest that (1) users are more satisfied when they receive a clarification that and how the system corrected a false memory, (2) users even prefer failed correction attempts over no such attempt, and (3) the tone of the clarification has to be considered for the best possible user satisfaction as well.

ACM Reference Format:
Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2019. Clarifying False Memories in Voice-based Search. https://doi.org/10.1145/3295750.3298961

1 INTRODUCTION

Re-finding and known-item search are common retrieval tasks. But the longer an item was not accessed, the more the memory declines [4, 6, 7]. This decline may even lead to *false memories* (i.e., misremembered "properties" the desired item actually does not have). Queries containing such false memories will match no or only inadequate results, leaving the user at a loss when the retrieval system does not detect and correct the false memory.

When systems correct a user request, the user should be made aware of this to avoid confusion—a particular challenge for the new and promising technology of personal voice assistants. While such assistants are specifically useful for re-finding and known-item searches when not sitting at a computer (e.g., at a dinner table, what was that movie with actor X?), every additional information the system wants to transmit means the user has to listen for their answer for a longer time. Given a retrieval system that could identify and correct false memories, we carry out a user-centric study to broaden our understanding of how to best clarify false memories in voice-search setups (e.g., it was a movie with actor Y, not X!).

With this first study on the topic of false memory clarifications in voice-based search, we address the following research questions.

RQ I Does language fluency affect user satisfaction?
RQ II Do wrong clarifications degrade user satisfaction?
RQ III How to best clarify false memories?

To answer these questions, we conducted a user study in which the 12 participants had to find answers for various information needs using a voice assistant despite false "memories" in the need descriptions. We tested different ways in which a system could respond after it detected a false memory and had the users rate on 5-point Likert scales whether the assistant was helpful, behaved as expected, was easy to understand, and was pleasant to use. As data, we collected 672 judgments as well as background information and general comments from the participants (cf. Section 3), which we use for an empirical analysis of the research questions (cf. Section 4).

Our key findings are, among others, that for voice-based search (1) automatic corrections of false memories should be made clear to the user to increase user satisfaction, (2) users even prefer a failed attempt to correct a false memory over no such attempt under certain circumstances, and (3) the tone in which the correction is clarified impacts the user satisfaction as well. The results of our study can directly be employed in the design and evaluation of search-related voice-based interfaces. In addition, they open the door for more focused research on query corrections—not just false memories—in the future.

2 RELATED WORK

We briefly review the related work on clarifying queries and re-finding / known-item search (the user has previously accessed the item in case of re-finding but not necessarily in known-item search).

Query Clarification. Clarifying a user's query (intent) has been studied extensively in the context of web search interfaces [3, 18] but voice-based systems lead to different interaction patterns [20]. Voice-based search systems are characterized as mixed-initiative systems between users and memory-equipped agents [16, 17, 22],

Scenario: You try to remember the title of a controversial book that came out in the 1990s and claimed scientific evidence that whites are genetically superior to blacks. You think its title was like "The *something* Factor."

Interaction start: Alexa. Explore!
What is the title of the book from the 1990s that claimed superiority of Whites and is called "The *something* Factor" ?

Post-interaction questions:

The system…	Agree				Neutral / Disagree	Don't know
…was helpful	□	□	□	□	□	□
…behaved as I expected	□	□	□	□	□	□
…was easy to hear/understand	□	□	□	□	□	□
…was pleasant to use	□	□	□	□	□	□

Figure 1: Example task. Participants started with "Alexa, Explore!" and read the question. The highlight shows the seeked item's attributes—one of which is misremembered. After the interaction, the participants rated the system.

but focused studies on the respective user interaction were lacking for quite some time [1, 15, 16]. Trippas et al. [19] were among the first who found that query refinement / clarification plays a crucial role in voice-based search (at least 25% of the interactions in their study are clarifications) while Braslavski et al. [5] studied human strategies to clarify questions on a question answering platform.

In a previous study [12], we analyzed clarifications in voice-based search for ambiguous queries (where several interpretations of the query seem plausible, in contrast to false memories where no interpretation matches the query without assuming an error). We found that clarifications do not decrease the user satisfaction and that the level of English proficiency of the participants played a major role in their perception of the (English) system.

Re-finding and Known Items. Blanc-Brude and Scapin [4] found that re-finding personal documents comes with the problem of memory decline and false memories (e.g., wrongly associated keywords after some weeks). Similar memory decline can be observed for email re-finding [6, 7] while re-finding in web search is often used to continue a previous, not too long ago session [21].

To study re-finding / known-item search without large-scale logs, different approaches tried to generate known-item queries automatically [2, 8, 13] or in human-computation games [14]. But these approaches fail to capture realistic known-item searches with false memories [10, 11]. Hagen et al. [9] thus crawled a set of 2,755 known-item intents (movies, books, etc.) from Yahoo! Answers and found that about 10% contain false memories. The information needs in our user study are inspired by these cases.

3 USER STUDY

In order to analyze clarifications of false memories in voice queries, we conducted a user study in which participants had to query a voice assistant in short information-seeking tasks. The task descriptions we provided, however, contained a wrong information to simulate the participant having a false memory (cf. Figure 1).

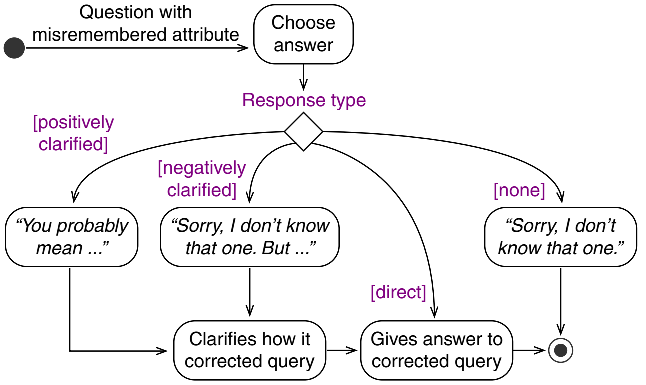
Question with misremembered attribute
Choose answer
Response type
[positively clarified]
[negatively clarified]
[none]
"You probably mean ..."
"Sorry, I don't know that one. But ..."
"Sorry, I don't know that one."
[direct]
Clarifies how it corrected query
Gives answer to corrected query

Figure 2: The four response types of our study with the corresponding Alexa reaction.

3.1 Setup

For the study, each participant had to (1) fill out a privacy-related consent form, (2) provide basic and study-related background information, (3) read the provided instructions, (4) complete two small "tutorial" tasks (off the record), (5) complete the 14 main tasks, and (6) give comments and suggestions (optional).

For each of the 14 tasks, a participant received a print-out as shown in Figure 1 and was seated in front of an Amazon Echo voice assistant. Each task contains a question that the user should read aloud to the voice assistant. The question fits the information need that is described in the scenario and always contains three attributes that specify the searched item. The tasks are based on real information needs from the false memory needs collected by Hagen et al. [9]. We rewrote the needs to short questions that are easy to speak aloud. We made sure that, due to the misremembered attribute, no answer exists that matches all three attributes. Note that the participants did not know which attribute is wrong.

To analyze how to cope with false memories in voice-based search, the voice assistant—controlled by us—responded in the following different ways to a participant's question (also cf. Figure 2):

none (2 tasks with this type in the study) It does not try to provide an answer but responds "Sorry, I don't know that one"—a standard line of Amazon's voice assistant in case no answer is found.

direct (4 tasks) It responds with an answer to a question where one attribute is modified. In accordance with best practices for voice-interface design, the response includes the attributes of the question.[1] We chose the system to mention the modified instead of the original attribute to hint at the modification.

negativeley clarified (4 tasks) It responds like for *none*, but then continues with a clarification how it modified one attribute to get to a question where it found an answer, and then answers this modified question like for *direct.*

positively clarified (4 tasks) It responds like for *negatively clarified*, but starts with the more positive suggestion "You probably mean … " instead of the failure-implying "I don't know."

As computationally resolving false memories is error-prone, we test the response types with attribute modifications (all but *none*)

[1] developer.amazon.com/docs/custom-skills/voice-design-best-practices-legacy.html

both for correct modifications of the misremembered attribute and for modifications of an actually correctly remembered attribute (2 tasks each). To signal the participant a correct modification, the instructor gave the participants a short thumb up gesture, which should simulate an “Oh yes!”-moment. However, if the instructor shows thumb down, the participant had to modify the question themselves until they got the desired answer. Participants where instructed before they started the tasks to modify questions in such a case by dropping a single attribute.

After each task, the participant had to rate their experience with the system on four metrics (cf. Figure 1). The whole setup was verified by one of our university’s privacy officers before execution.

For the study, we avoided voice recognition errors, which occur frequently in today’s voice interfaces, by using a tightly fit interaction model and pre-formulated questions. In detail, the assistant’s recognition model was trained to listen for the different questions of the known tasks, complete or with one attribute dropped. This setup allowed for an excellent speech recognition. This is important, as our study would otherwise be outdated quickly by the current rapid advancements in speech recognition quality. To avoid order-biases, we randomized the task-order for each participant.

3.2 Participants

After a small pilot study (not considered in the results), we recruited 12 different participants from our university for the main study. The participants were between 18–30 (6 participants) and 31–49 years old (6). Furthermore, 7 were male and 5 were female. Participants had an intermediate (5 participants) or proficient (7) English level. Finally, 2 participants stated to never use voice assistants, whereas 9 use them rarely and 1 uses them frequently.

3.3 Data

The 12 participants took between 20 and 26 minutes for completing the study. Each participant finished 14 tasks for a total of 168 interaction phases. For each phase, we collected 4 ratings in post-interaction questionnaires, totaling in 672 ratings.

4 RESULTS

The research questions we raised in the introduction all focus on user satisfaction, which we measure by participant ratings of the system’s effectiveness (did it answer the question?), predictability (did it behave as expected?), clarity (was it easy to hear and understand?), and pleasantness of use. In order to allow for a direct comparison with our previous study [12], we analyze the participants’ ratings using graphical plots and statistical testing like we did there. We always use Fisher’s exact test for statistical testing as it is most-suited for the small size of our study.

4.1 Does language fluency affect satisfaction?

Since query corrections may arise unexpectedly, it can be difficult to make it clear to the user what happened, especially in voice interfaces and (as one would assume) even more so when the user is less fluent in the spoken language. Figure 3 illustrates the ratings of the participants for the four study questions according to the participant’s fluency of English. Based on previous results [12], we expected that participants that are less fluent in English would find the system less pleasant to use. However, while there is a significant difference ($p < 0.01$), there is no correlation between fluency and pleasantness ratings (Pearson’s $r = 0.05$). Moreover, as the figure suggests, there is no significant difference in perceived effectiveness and predictability ($p > 0.05$). Therefore, for the response types we tested, English fluency seems to be not too much of an issue, even though we measured significant differences with respect to the perceived clarity ($p < 0.01$, $r = 0.28$). We believe that this difference to previous results stems from the fact that we did not use response types in which participants needed to formulate their own replies, which is much harder for users that are less fluent.

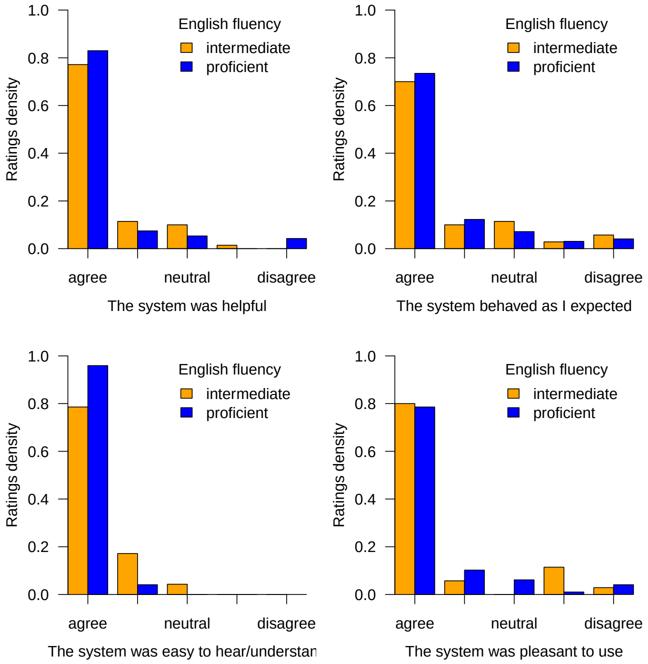

Figure 3: Overall ratings by English fluency.

4.2 Do wrong corrections degrade satisfaction?

Query corrections can have a severe impact on a user’s perception of the system, and thus it is natural to ask under what circumstances their gain outweighs their cost. While query corrections in case of false memories may allow the system to give an answer at all, users might get annoyed by unmotivated and especially unhelpful modifications to their query. As Figure 4 shows, the participants were, for some response types, more pleased with a system that corrected their query even if it did not help them with their information need. For *positively clarified* corrections, this result is significant despite the small sample size for both the general case and when restricted to tasks where the misremembered attribute is modified (both $p < 0.05$). This result suggests that systems should try to correct queries, even at the risk of picking the wrong attribute.

4.3 How to best clarify corrections?

Once the system identifies a potential correction, the question arises how to clarify that and/or the fact that results were retrieved for

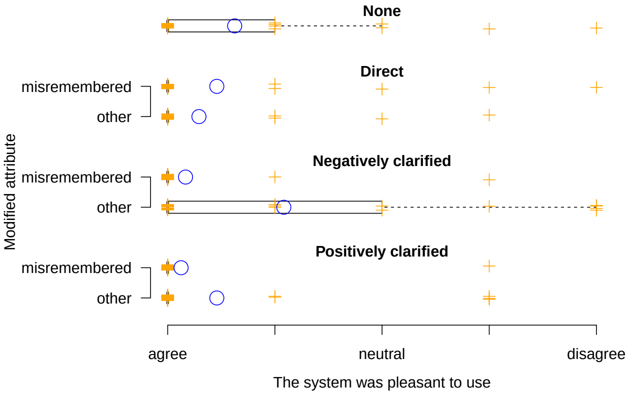

Figure 4: Box plot, single ratings (+), and mean (○) pleasantness by response method and modified attribute.

a correction to the user—or whether to clarify all. While much more ways of explanation are imaginable, we here focus on two of the most basic parameters to explanations: should corrections be clarified at all and does the tone of the correction matter?

For the case that the system modifies an attribute that was actually remembered correctly, Figure 4 shows a clear difference between the two response types that clarify the correction, where the positive one (which is worded more like a suggestion than a correction) was rated as more pleasant on average. This is intuitive as a wrong suggestion (as in *positively clarified*) is not as harsh as a wrong correction (as in *negatively clarified*). Moreover, while only three participants gave a rating between neutral and disagree for *positively clarified* (cf. Figure 4), seven participants did so for *negatively clarified*. In fact, four participants fully disagreed that the latter was pleasant to use, but none did so for *positively clarified*. However, due to the small sample size, this difference is not statistically significant ($p > 0.05$) and should be validated in future studies. Still, our results suggest that the tone of the correction should be taken into account to maximize user satisfaction.

For the case that the modified attribute is the misremembered one, Figure 5 displays the participants' ratings in detail. As can be seen, the response types that clarify the corrections receive better ratings in both predictability and pleasantness. Furthermore, two participants specifically made a comment that the explanations were useful and that they would wish for such a functionality in the real assistant. Even though this result is not statistically significant for our small sample size ($p > 0.05$), the distribution of ratings and informal feedback nevertheless suggests that explanations of the query corrections are well-received and to some degree expected, even if today's assistants do not yet contain this functionality.

5 CONCLUSION

We present the first user study on how voice-based search systems may communicate false memory corrections to their users. We identified three key research questions for false memory clarification regarding adaptations to the user's language level, possible costs of query corrections, and optimal clarification phrasing. Our findings are—among others—that (1) clarifications usually raise user satisfaction; (2) a failed correction, where the system assumes the false memory in the wrong attribute and modifies what was actually correct, is preferable over no correction, and (3) the tone of the clarifications impacts user satisfaction as well, with a positive tone achieving a higher user satisfaction. Unlike previous research [12], we did not find that the user's language fluency has a large impact on their experience. We attribute this to the fact that our participants did not really need to formulate own queries.

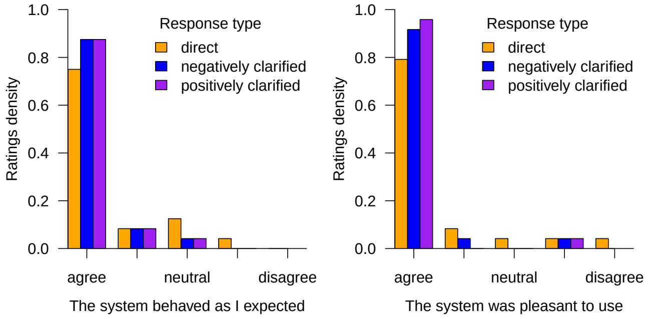

Figure 5: Ratings for predictability and pleasantness when the misremembered attribute is corrected.

All these findings speak in favor of more conversation-oriented interfaces. Participants were the most satisfied with the system that kept the conversation going, by offering a clarification in the positive tone of a polite suggestion. Moreover, in this case participants where also more forgiving to failed corrections, which is especially important as the correction of false memory queries is a difficult task and current systems fail to achieve satisfiable performance yet.

However, the limitations of the study have to be considered for further interpretation. Our participants do not constitute a representative sample of voice assistant users. Therefore a large-scale study (e.g., using crowdsourcing) should be conducted to improve the confidence in the results of our preliminary study. Furthermore, while the false memories used in the study are based on "real" false memories of information seekers, they are not the false memories of the participants. Therefore, real users with false memories may react differently. Especially, they may not accept so easily that they misremembered something. In this regard, our setup mirrors the case where all users directly accept their error.

Our findings open the door for more focused research on query corrections and false memory clarifications in voice-based search. Specifically, more ways of clarification should be investigated in the future, with "positively clarified" as the new baseline. For example, the system might ask the user whether they may have erred on some attribute in the question. Finally, to maximize user satisfaction, experiments with several positive response types could be conducted to find a set of suitable response phrases for alternation.

REFERENCES

[1] James Allan, W. Bruce Croft, Alistair Moffat, and Mark Sanderson. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012. *SIGIR Forum* 46, 1 (2012), 2–32.

[2] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building simulated queries for known-item topics: An analysis using six european languages. In *Proceeding of SIGIR 2007*, pages 455–462.

[3] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. 2011. Query suggestions in the absence of query logs. In *Proceedings of SIGIR 2011*, pages 795–804.

[4] Tristan Blanc-Brude and Dominique L. Scapin. 2007. What do people recall about their documents?: Implications for desktop search tools. In *Proceedings of IUI 2007*, pages 102–111.

[5] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What do you mean exactly?: Analyzing clarification questions in CQA. In *Proceedings of CHIIR 2017*, pages 345–348.

[6] David Elsweiler, Mark Baillie, and Ian Ruthven. 2008. Exploring memory in email refinding. *ACM Transactions on Information Systems.* 26, 4 (2008), 1–36.

[7] David Elsweiler, Mark Baillie, and Ian Ruthven. 2011. What makes re-finding information difficult? A study of email re-finding. In *Proceedings of ECIR 2011*, pages 568–579.

[8] David Elsweiler, David E. Losada, José Carlos Toucedo, and Ronald T. Fernández. 2011. Seeding simulated queries with user-study data for personal search evaluation. In *Proceedings of SIGIR 2011*, pages 25–34.

[9] Matthias Hagen, Daniel Wägner, and Benno Stein. 2015. A corpus of realistic known-item topics with associated web pages in the ClueWeb09. In *Proceedings of ECIR 2015*, pages 513–525.

[10] Claudia Hauff, Matthias Hagen, Anna Beyer, and Benno Stein. 2012. Towards realistic known-item topics for the ClueWeb. In *Proceedings of IIiX 2012*, pages 274–277.

[11] Claudia Hauff and Geert-Jan Houben. 2011. Cognitive processes in query generation. In *Proceedings of ICTIR 2011*, pages 176–187.

[12] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward voice query clarification. In *Proceesings of SIGIR 2018*, 1257–1260.

[13] Jinyoung Kim and W. Bruce Croft. 2009. Retrieval experiments using pseudo-desktop collections. In *Proceedings of CIKM 2009*, pages 1297–1306.

[14] Jinyoung Kim and W. Bruce Croft. 2010. Ranking using multiple document types in desktop search. In *Proceedings of SIGIR 2010*, pages 50–57.

[15] J Lai and N Yankelovich. 2006. Speech Interface Design. (12 2006).

[16] 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 CHI 2016*, pages 5286–5297.

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

[18] Fabrizio Silvestri. 2009. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in IR* 4, 1–2 (2009), 1–174.

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

[20] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search. In *Proceedings of CHIIR 2018*, pages 32-41.

[21] Sarah K. Tyler and Jaime Teevan. 2010. Large scale query log analysis of re-finding. In *Proceedings of WSDM 2010*, pages 191–200.

[22] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of CHI 2017*, pages 2187–2193.