Inclusion vs exclusion oriented clarifying questions for conversational product search

ROBERT IGNAT, University of Twente, The Netherlands

Standard product search techniques such as faceted search are commonly employed in today's systems. Even though widely used, they come with several limitations, such as not being able to understand users' natural language. Therefore, other search methods like conversational search became a topic of interest to the scientific community. As conversational product search systems such as product advisors become more prevalent, the question of what the benefits and drawbacks of different approaches are arises. There is existing work discussing different types of clarifying questions or various frameworks such as critiquing. Typically, these questions aim to find an acceptable range of values. However, research on the benefits and drawbacks of asking clarifying questions that seek to exclude rather than include certain values is scarce. This research employed a qualitative user study in which we explored the advantages and disadvantages of inclusion-oriented and exclusion-oriented clarifying questions. Our findings suggest that exclusion-based questions are found to be harder to answer and are perceived more negatively, but can be useful for asking categorical questions in situations where it is likely for users to have "dealbreakers".

Additional Key Words and Phrases: clarifying questions, product search, conversational search, exclusion

1 INTRODUCTION

In E-commerce, users often need to search for products. Traditional search techniques offer modest performance - limitiations of classic search systems are outlined by ter Hofstede et al. [11] and Balfe and Smyth [2]. Similarly, Aliannejadi et al. [1] point out that most users fail to specify their complex needs in a single query when using traditional search techniques.

Experts have been looking for a way to solve these shortcomings, and one approach that is increasingly popular is the use of conversational search. Conversational search refers to the practice of having users communicate with a virtual agent through natural language, with the goal of the agent being to understand users' information or product needs and to filter out irrelevant items. Papenmeier et al. [8] demonstrate the need for natural language search systems for product search.

When the user's input is not clear enough for the system to confidently retrieve relevant items, conversational systems can use clarifying questions (CQs). CQs are questions that aim to elicit a more specific answer in the case of the user providing preferences that are too vague. As Zou et al. [14] mention, asking users good clarifying questions (CQs) is one of the main requirements for a good conversational search system.

Zou et al. [14] also conclude that the majority of users (66-84%) find conversational search particularly useful. However, while Aliannejadi et al. [1] claim that asking even one relevant CQ significantly boosts the performance of a system, Zou et al. [13] show that less relevant CQs can actually decrease a system's performance.

Therefore, many CQ styles and approaches such as "System Ask, User Respond" [12], negative feedback [4] and critiquing [9] have been designed and evaluated by the research community. However, the research around the different types of CQs and their formulation is limited.

In a study that analyzed how experts aid users in finding online recipes, Papenmeier et al. [7] identified a behaviour that is unusual for filtering mechanisms - instead of asking users what ingredients should be included in the meal, experts asked about which ones to exclude. In another research, Kern et al. [6] implemented a search system in which users could specify "must-not-haves", a feature that several participants found useful.

These situations in which exclusion is used raise the following **research question:**

"What are the benefits and drawbacks perceived by users of such exclusion-focused CQs, compared to standard inclusion-focused CQs?"

In this paper, we define *exclusion-focused CQs* to refer to questions that aim to exclude certain values, such as "Are there any brands that you would definitely not buy?", and use *inclusion-focused CQs* as a name for questions that ask for a range of acceptable values, like "What brands to you prefer?"

While the above cases spotted by Papenmeier et al. [7] and Kern et al. [6] show that there are situations where exclusion-oriented CQs are useful, research about the use of exclusion-oriented CQs is limited.

We developed two variants of a conversational search prototype system that implements exclusion-focused and inclusion-focused CQs respectively. We then explored how users perceive those two types of CQs in a qualitative user study.

The results of the study give insight into how exclusion-oriented CQs are perceived, and what their likely use cases are. Our user data suggests that using exclusion CQs as a default is not optimal, because they are generally perceived as being less relevant, harder to answer, and can sometimes fail to properly capture users' needs.

However, exclusion CQs can be beneficial, mainly in the case of categorical product aspects with a reasonable chance of users having "dealbreakers" (e.g., to exclude a specific brand).

This qualitative research contributes to the body of knowledge about clarifying questions in conversational product search by investigating the viability of exclusion-oriented CQs.

2 RELATED WORK

Conversational product search is a relatively new topic in computer science. Despite that, there are several sources mentioning faults of

TScIT 39, July 7, 2023, Enschede, The Netherlands

classical search systems or the need for a natural language search system, and literature about different approaches and their performance can be found. This section aims to give an overview of this literature.

In their work, Balfe and Smyth [2] have identified a mismatch between search engines' interfaces and users' vocabulary. They also spotted the problem of vague language, which is that people sometimes use language that is not clear enough to be used for filtering results, and systems do not have a way of requesting a more specific description. They also recognized how standard search engines are not capable of understanding context. Similarily, ter Hofstede et al. [11] point out how users often fail to properly filter items. In addition, Papenmeier et al. [8] explain that users are currently required to adapt to the requirements of search systems, which both increases the burden on users and decreases the system's performance.

Zhang et al. [12] have proposed a framework for conversational search - System Ask User Respond (SAUR), where the system asks the user several questions about different aspects of the product. In their paper, they formalized the search problem into mathematical terms. They have also expressed how they view the ability to ask CQs as one of the main advantages of conversational systems.

The benefit of asking clarifying questions was also observed by Aliannejadi et al. [1], who concluded that even one good CQ has a significant boost on performance. Zou et al. [13] came to the same conclusion, but they also noted a decrease in performance when lower quality CQs (questions that fail to stay on topic or solicit disambiguation) were used.

Aside from the SAUR framework, other conversation flows were also proposed. Bi et al. [4] demonstrated that a framework where the system gives users recommendations and then asks CQs regarding aspects of rejected items performs well, while Ricci and Nguyen [9] proposed a similar critique based paradigm, where users are shown product variants, and are requested to give critiques relative to the displayed options, until they reach a desirable item.

Although these approaches showed positive results, they make use of relative feedback, meaning feedback that only gives a value range in relation to a reference item (e.g. "Cheaper than X") which is found to be inferior to absolute feedback by Christakopoulou et al. [5]. These approaches also differ from simply asking CQs in that they rely on repeatedly showing the user results and letting them comment on those, rather than just asking for preferences.

Regardless of the exact approach to the conversational search system, an empirical study conducted by Zou et al. [14] confirmed that the majority of users find the conversational search paradigm useful. They also found that users are willing to answer around 11 to 21 questions, and that in 12-17% of cases, users would give erroneous answers to questions.

Regarding what types of CQs give the best results, Bi et al. [3] found out that closed CQs (yes/no answer) tend to have a lower mental load than open ones, due to closed CQs requiring less effort for users to answer.

While most of the above studies employed models that calculate the best CQ to ask from a predefined set of possible questions, Sekulić et al. [10] succeeded in employing a GPT2-based model in a system for generating CQs resulting in satisfying performance.

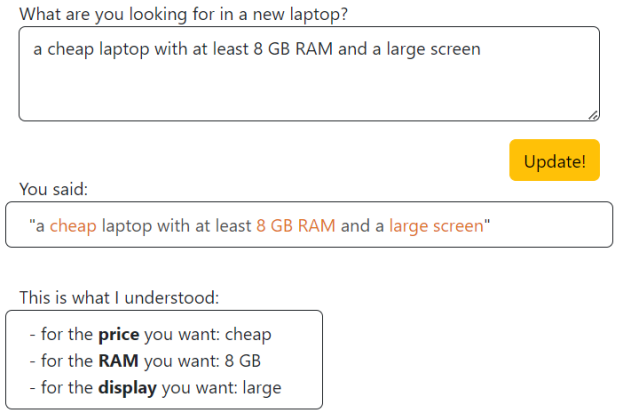

Fig. 1. The initial prototype system for processing natural language

Lastly, while negative feedback and critique based models were based on the assumption that users do not always know exactly what they want, Ricci and Nguyen [9] conclude that users often already have an idea of what they need, and suggest that letting users first list their preferences is desirable.

3 METHODOLOGY

For the purpose of this research, we considered the scenario of users needing to buy a laptop, similar to the one used in the paper of Papenmeier et al. [7]. To aid them with this task, users were asked to use a conversational search system, which employs CQs. We compared two versions of CQs:

- Inclusion-focused CQs, e.g., "What brands of laptop would you prefer?" or "What is the price range that you would be willing to pay?"
- Exclusion-focused CQs, e.g., "What brands of laptop would you not buy?" or "What is a price that you would not be willing to pay?"

This was an in-person experiment that had a within-subjects design. The user study's design and procedures were cleared by the ethics committee of the University of Twente.

3.1 System implementation

For the user study, we used an already existing prototype (Figure 1) that is capable of interpreting natural language queries in laptop search. It maps values to product aspects and flags vague responses. We extended the system so that it first queries the user's initial preferences, and then acts as a chatbot that asks CQs about missing or vague aspects (Figure 2). Based on a toggle, the system asks either inclusion-focused CQs or exclusion-focused CQs (see above examples). After answers for all product features are gathered, the system announces that it is ready to display recommendations. However, the actual recommendation results are outside of the scope of this research.

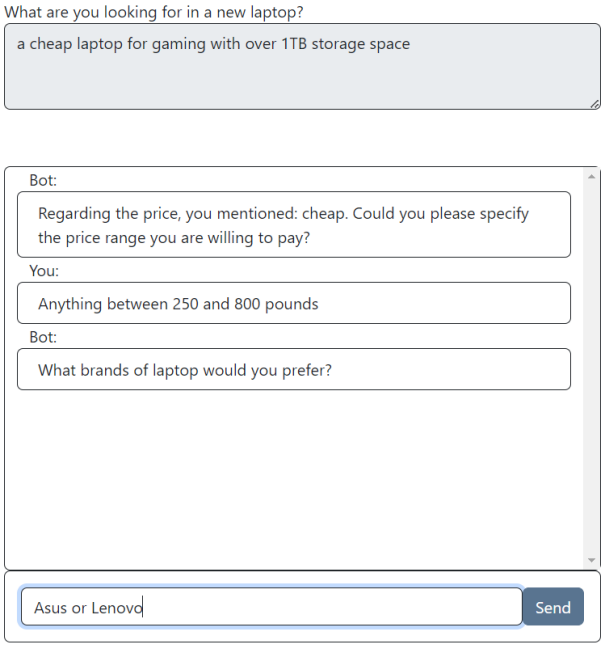

Fig. 2. The final system used in the user study

3.2 Use case

The laptop scenario is appropriate for this research because buying a laptop is a scenario that is easy for users to imagine, as they likely use laptops on a daily basis. Furthermore, it allows for a wide range of questions, of varying technical degrees and of different types - categorical and numeric. Specifically, we ask about the laptop's purpose, brand, color (categorical) and about price, screen size, storage space, RAM and battery life (numerical).

Before interacting with the chatbot itself, we first let users specify their preferences for a laptop in a text box, which aims to improve the interaction as argued by Ricci ang Nguyen [9]. This also lets the rest of the interaction consist of purely CQs and answers.

3.3 Participants

During this research, we interviewed 9 participants of ages between 21 and 27, 3 of which were females while the other 6 were males. All participants were students at the University of Twente, and were fluent in English. When asked to self-assess their knowledge about laptops on a scale from 1 to 5, all of them chose an answer between 2 and 4.

The interviewees were recruited via word-of-mouth on the university campus and via the author's personal network.

Participants were told that the study concerns the language of conversational search systems, but were not given information about what specific aspects of language are being studied, or about what the difference between the two interactions will be.

3.4 Procedure

The setup of the experiment was offline, with the implemented system hosted on the researcher's laptop. First, participants were given the information letter to read, after which they gave informed consent. The data was gathered via a temporary audio recording of the entire session.

Participants were informed about the scenario, which is to imagine that they need a new laptop, and are using the conversational system in order to make their choice.

Each participant interacted with both variants of the system. The order of the interactions was alternated (5 people had inclusion as their first variant, and 4 had exclusion).

After each interaction, when the system claimed to have results, participants were asked about their experience. The goal was to understand what benefits and drawbacks the users perceive in relation to the given CQ style. Participants were given the opportunity to freely express their opinions and feelings about the system and its question style.

To facilitate the discussion, some additional questions were asked. Participants needed to rank each following aspect on a 7-point Likert scale, while also justifying their choices:

- The level of satisfaction with the interaction
- The ease of use of the system (and how easy it was to answer the questions)
- The relevancy of the questions asked by the system
- The likeliness of the system being able to make good recommendations, based on the questions asked

It is worth noting that the Likert scales were simply used as a catalyst for conversation - this qualitative study does not analyze the scales' results.

Following that, participants were asked what the most and least relevant question was, and whether there were any product requirements that they feel the system has failed to elicit.

As a last step, participants were asked which system they prefer and why, and whether there is any requirement that the not preferred version did a better job of eliciting.

3.5 Data processing & Analysis

Because of the qualitative nature of the study, no statistical analysis of the results was made. Instead, the feedback for each CQ style was collected, and analyzed qualitatively with the goal of providing a nuanced view of the advantages and drawbacks of each style. We followed a standard procedure of inductive coding.

First, for every interview, the audio recording was transcribed, after which the recording was deleted. The transcription was then processed, and every statement made by the participant was noted. Then, statements that help to answer the research question were differentiated from other statements regarding the interaction. Lastly, all the statements were iteratively clustered together until there was a coherent image of what the key findings are.

4 RESULTS

This research is an exploratory study that did not confirm or test statistical hypotheses. Because of the qualitative nature of the study, little attention will be paid in this section to exactly how many

users held a certain view. Instead, we aim to report all perceptions, regardless of how common they were within our user group.

4.1 General patterns

Aside from data concerning the research question during the interviews, participants revealed additional interesting data regarding conversational search, their desires, needs, and their level of knowledge. This subsection briefly outlines these findings.

Regarding how important each aspect about a laptop is, most users said that color is not very relevant for them, while price and RAM are very relevant. Purpose is also seen as quite relevant, as some participants claimed it also informs the system about other requirements. One interviewee said that the relevancy of some aspects depends on the person, which seems to be true, as users did not have the same opinions about the relevancy of battery life, storage space and brand.

Users also expressed what other aspects (that were not part of the interaction) are important to them, and the processor's performance was the most common answer. Other mentioned aspects were weight, jacks, keyboard lights, noise level, brand deals, and HDD vs SSD storage.

Aspects aside, a commonly desired feature was to have the chatbot give either reference values or suggestions for each of the questions. Besides that, some participants wanted the chatbot to act more proactively during the interaction, while others wished for some confirmation on the chatbot's side that their requirements were understood.

Finally, several interviewees did not have a complete, clear image of what they wanted, and some mentioned not being knowledgeable about certain aspects, such as RAM or brands. In addition, a participant claimed that they would ask an expert friend for recommendations.

The rest of this section details the results that helped with answering the research question. Specifically, statements that were relevant for the research question were assigned to one of the following themes: ranges vs. bounds, ideal vs. minimum product, perceived performance, effort of answering, unexpectedness of phrasing, and tonality.

4.2 Ranges vs. upper/lower bounds

For numerical attributes, the inclusion system asked for acceptable ranges, while the exclusion one requested lower / upper bounds. Many participants showed a need to set ranges for acceptable values (with both lower and upper bounds to express their needs. Some users perceived the exclusion questions as less relevant overall, mentioning the aspect of giving ranges with upper and lower bounds as being preferred. Specifically, a lot of feedback was targeted towards the question about price, where many participants voiced their preference for stating a price range (inclusion) over stating a price that would be too high (exclusion). Their main reason was that the exclusion variant does not allow them to specify a lower bound, and they had a quality expectation attached to the price.

However, this sentiment was not unanimous, as there were those which only cared about budgeting, and so preferred the exclusion variant. In a similar way, some stated that they liked the exclusion battery life question better, as they did not care about the upper bound. Some users complained about the case where a range with both the lower and the upper bound of their desired range was needed (such as for screen size, where the exclusion question was "What screen sizes would be too small or too large?"). In such cases, they found that they had to answer two things at the same time (too small and too large), which was confusing for them.

4.3 Ideal vs. minimum viable product

One of the main issues that several participants identified is that exclusion-based CQs do not elicit the user's ideal product, but rather the minimum acceptable one, and are likely to result in lower-end recommendations. As one user puts it, "Because of the way that the questions are phrased, I'm going to give an answer that is like, this is the bare minimum, but not what I'm looking for.". Similarly, a user said that inclusion numerical questions allow you to still specify a minimum, while also indicating you would prefer more.

4.4 Perceived performance

Participants mentioned several issues with the questions that have an impact on their perceived performance of the system.

Opinions were split, as some said that the algorithm behind the exclusion system would work better, while others speculated that the exclusion system might not lead to good recommendations. Other users claimed that the exclusion system has lower perceived performance than the inclusion one, that the exclusion questions were seen as overall "worse", and that the negative phrasing "makes the relevancy lower". In addition, some participants argued that exclusion opens up the possibility of the user being misunderstood.

For categorical questions, the feedback heavily depended on the aspect being queried. Participants claimed that the exclusion CQ about the laptop's purpose fails to elicit their requirements, as they found their answer unlikely to be informative, arguing that saying that you do not use the laptop for video editing does not mean you do not use it for gaming or other demanding tasks. Another mentioned example where the excluded values might not even give much information would be the exclusion RAM question ("How much RAM would be too little?"). While 4 GB might be the most representative answer for the user's needs, 2 or 1 GB are also technically correct answers.

Some users also raised the concern that one might still miss some undesired categories when responding to exclusion-oriented categorical questions. In addition, some interviewees felt that while exclusion successfully filters out unwanted values, it does not actually capture the user's needs, leaving ambiguity about what they actually want. To support this, they claimed that excluding a few values does not mean that all the other categories are actively desired.

One other view mentioned by some interviewees is that the inclusion questions give the user more control, as they do not demand as strict of a response as some of the exclusion CQs. To add to this, a participant felt that inclusion questions help the system understand which aspects are more important.

Excluding brands was also viewed as more relevant than listing acceptable ones by several users, as they claimed that, while they

had no clear preference or demand for a specific brand, they did have in mind brands that they would clearly not buy. One of the participants that held this position described it as "I don't want this one, but anything else is fine". At the same time, an interviewee had the opinion that "If I care about the brand at all, then I have a brand that I do want".

4.5 Effort of answering

The main finding for this aspect is that participants had a harder time answering exclusion CQs than inclusion ones, but their reasoning tended to focus on the categorical questions.

Most of the possible values were undesirable for participants, as they only wanted a few categories. Therefore, they had to list all the unwanted values, which was seen as "tedious".

As a response to this, some users resorted to flipping the exclusion question around - instead of listing what values they do not want, they answered "Anything **but** [their desired values]".

While the exclusion color question was harder to answer, it can have a positive impact on users' needs elicitation. For example, someone mentioned that having to exclude all the unwanted colors makes them actually think about all the colors, and what they actually want - "It makes me think about all the possibilities".

4.6 Unexpectedness of phrasing

The dominant sentiment among participants was that the way the exclusion questions were asked was "weird", unexpected and confusing for some. As one of the interviewees put it, "It wasn't the kind of questions I would expect [...] the wording was really weird", referring to the questions asking what they do not want, and to how they are worded in a negative way (e.g., "What is a price that you would not be willing to pay?"). A few users also mentioned how the wording is different from other chatbots, or from their envisioned idea of a chatbot.

One other aspect that participants brought up was that exclusion questions were more difficult to answer due to their unexpected phrasing. That being said, multiple participants expressed that, putting their sentiments aside, the questions themselves were clearly understandable.

4.7 Tonality

While some participants said that they liked the exclusion concept, others noted that they prefer the inclusion version, and that the inclusion questions felt better.

Several participants voiced a dislike of the way that exclusion CQs were phrased. One of these users also claimed that in their communication science background, they were advised against using negative phrasing.

Similarly, some perceived the inclusion interaction as being more conversational, the system being "More inviting to interact with", and the questions as being closer to how a human would ask about such things. Another piece of feedback was that exclusion wording negatively influences the interaction's feel/mood.

5 DISCUSSION

The outlined results seem to show that using exclusion as a default way of asking CQs is not recommended, as it generally tends to have a negative impact on the user's experience, compared to standard inclusion-oriented CQs. When used in the wrong situation, exclusion CQs have the risk of failure to elicit the user's actual requirements.

Besides that, exclusion-based CQs seem to be overall harder to answer, mostly due to users needing to exclude a lot of values, or due to the unexpected phrasing. This can be a problem since, as Zou et al. [14] concluded, users give erroneous answers around 17% of the time even when the CQs that are asked are not hard to answer. Therefore, using questions that are more difficult to answer would likely increase that error percentage, which would have a significant impact on the quality of recommendations.

On the other hand, there is a possibility of exclusion having a use in situations were the system wants to make people reconsider their preferences, and think about all the potential values for categorical questions - this was shown to be possible when a participant preferred the exclusion color question, saying that it forced them to "actually start considering all the other colors".

Based on the results, exclusion CQs could also work for numerical range questions, if users only care about one of the bounds (higher / lower) - users which only cared about their budget for the price questions, and those who only had a minimum requirement for battery life responded positively to exclusion questions about those aspects.

However, the main scenario in which exclusion seems to outperform inclusion is when addressing categorical questions, in cases where it is likely for users to have dealbreaker values. Such a situation occured in the case of the brand question, where multiple participants voiced their preference for the exclusion variant, mentioning the lack of a strong preference, but also the existence of brands that they would never buy. This behavior is similar to the one already identified by Papenmeier et al. [7], where recipe experts would ask participants what ingredients they should exclude, possibly because of allergies or intolerances.

An interesting finding was how participants were not expecting the phrasing of exclusion questions. The most likely cause of this is that conversational agents today seldom make use of exclusion, as it is not a thoroughly researched topic. This unexpectedness can cause the questions to be perceived in a more negative light, but if exclusion were to be widely used (in an appropriate manner), it is possible that exclusion CQs would be seen less negatively. Still, it did not seem that the lack of familiarity was the only cause of the perceived negative tonality. As one participant mentioned, negative phrasing is not generally recommended for communication.

It therefore seems that exclusion CQs should not be used without a proper reason. But, if it suspected that there are categories of a product aspect that would be a dealbreaker for users, then exclusion-oriented CQs can be a useful tool for filtering out unwanted items.

This finding is consistent with the conclusions drawn by Kern et al. [6], who found that, while using a non-conversational search interface, users would sometimes make use of exclusion when applying filters for finding their ideal product.

The question is, then, how does one identify cases in which most users have dealbreakers? Since trying to predict such situations is unlikely to be reliable, one possible way would be to have a dynamic system that uses inclusion by default, but if many users struggle with an inclusion question, then the system would switch to exclusion for that question. This could be achieved using a large language model system, similar to the one implemented by Sekulić et al. [10]. The findings of this research point to the opportunity of enhancing such a system by instructing it to use exclusion when it is deemed appropriate.

Another good reason for using exclusion might be if the system does not have a product that satisfies all the user's requirements. Several participants said that exclusion successfully filters out dealbreakers, but does not get a proper image of their ideal product. However, if their ideal product does not exist, a less optimal, but still viable product should be recommended. But this would mean not adhering to all the user's requirements, which can result in recommending products that have dealbreakers. Exclusion-oriented CQs could then be used to identify such dealbreakers and avoid them in recommendations.

Interestingly, some users decided to flip an exclusion question around, by answering "Anything **but** [what they want]". This shows that there could be value in letting users give both inclusion and exclusion oriented answers. While this was achieved successfully by Kern et al. [6] for non-conversational search by letting users add weights to their preferences (including "must-haves" and "must-not-haves"), implementing this for a conversational search agent that only uses natural language might be more challenging - while their system was able to the the exact weights of preferences from sliders in the interface, a conversational agent would have to deal with the vagueness present in the way people would likely describe how important each preference is.

That being said, our research is subject to several limitations. First of all, it is worth noting that all the interviewees were students at the University of Twente, of ages between 21 and 27, so young persons that pursue higher education. This is not a representative sample for the general population, and so the opinions of our participants might not be perfectly representative of those of the general public. This is relevant because many conversational search systems are designed for the public at large.

It is also worth mentioning that most of the participants did not have English as their native tongue. This is relevant, since the natural language processing software used for the prototype was trained on native speakers from England. However, all participants demonstrated good command of the English language, as they all followed an English-taught programme. Also, there seemed to be no point during any of the interviews where lack of English knowledge was a problem.

Additionally, letting users first list out their preferences came with a drawback - because of the different aspects mentioned by users, not all users were presented with all the chatbot's questions, and so some potential feedback could have been lost.

6 FUTURE WORK

This research pointed to the possiblity of exclusion CQs being superior in the case of categorical questions where many users have dealbreakers. However, due to the qualitative nature of the study, this should not be interpreted as a general, objective fact. Therefore, more information could be added to the body of knowledge about exclusion-oriented CQs by performing a large scale qualitative study that seeks to conclude whether in the above case exclusion is superior, and if so, to what degree it performs better than inclusion.

Similar to Papenmeier et al. [7], this research pointed to the fact that there are situations where exclusion can be useful. That being said, there is no comprehensive list of well-studied cases in which exclusion CQs are preferred. As a consequence, what the exact situations that benefit from exclusion are could be investigated.

In addition, the systems tested in this study used only inclusion or exclusion, respectively. The research community could benefit from studying whether questions that allow for both types of answers (e.g., "Are there any specific brands that you want, or that you do not like?") perform better than purely inclusion or exclusion questions in certain situations.

7 CONCLUSION

Usability issues of classical approaches created a need for research about conversational product search systems. The ability to ask clarification questions is one of the key features that many such systems ought to have. While there is some literature covering certain aspects of CQs, no study has been done to assess the viability of using exclusion-oriented questions.

This research investigated the benefits and drawbacks of exclusion-oriented CQs, compared to the widely-used standard inclusion CQs. After studying the information gathered in our user study, we conclude that our data suggests that exclusion CQs are perceived more negatively in general, but that they can be superior to inclusion ones in cases such as categorical questions where users are likely to have dealbreaker values. These findings add to the growing body of knowledge about clarifying questions, as there is no prior work dedicated to studying the viability of exclusion CQs.

It is hoped that by adding knowledge to this field, future users will have an easier time interacting with conversational systems, and so will benefit more from the enhancement that such technology brings.

ACKNOWLEDGMENTS

I would like to express my gratitude to Andrea Papenmeier for supervising this research, for offering guidance when needed and for providing timely, valuable feedback. I would also like to thank the track chair - Mariët Theune, and Anis Hasliza Abu Hashim for creating an environment where peers could share valuable feedback with each other. In addition, I would like to thank my colleagues, Liran Neta an Wybe Pieterse for sharing knowledge and for offering help with participant recruitment. Last but not least, I appreciate the participants of the user study for offering me their time and attention.

REFERENCES

[1] Mohammad Aliannejadi, Fabio Crestani, Hamed Zamani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Inc, (July 2019), 475–484. ISBN: 9781450361729. DOI: 10.1145/3331184.3331265.

[2] Evelyn Balfe and Barry Smyth. [n. d.] Improving Web Search Through Collaborative Query Recommendation. Tech. rep. www.friendster.com.

[3] Keping Bi, Qingyao Ai, and W. Bruce Croft. 2021. Asking Clarifying Questions Based on Negative Feedback in Conversational Search. In *ICTIR 2021 - Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. Association for Computing Machinery, Inc, (July 2021), 157–166. ISBN: 9781450386111. DOI: 10.1145/3471158.3472232.

[4] Keping Bi, Qingyao Ai, Yongfeng Zhang, and W. Bruce Croft. 2019. Conversational product search based on negative feedback. In *International Conference on Information and Knowledge Management, Proceedings*. Association for Computing Machinery, (Nov. 2019), 359–368. ISBN: 9781450369763. DOI: 10.1145/3357384.3357939.

[5] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 13-17-August-2016. Association for Computing Machinery, (Aug. 2016), 815–824. ISBN: 9781450342322. DOI: 10.1145/2939672.2939746.

[6] Dagmar Kern, Wilko Van Hoek, and Daniel Hienert. 2018. Evaluation of a search interface for preference-based ranking - Measuring user satisfaction and system performance. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, (Sept. 2018), 184–194. ISBN: 9781450364379. DOI: 10.1145/3240167.3240170.

[7] Andrea Papenmeier, Alexander Frummet, and Dagmar Kern. 2022. "mhm.." conversational strategies for product search assistants. In *CHIIR 2022 - Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. Association for Computing Machinery, Inc, (Mar. 2022), 36–46. ISBN: 9781450391863. DOI: 10.1145/3498366.3505809.

[8] Andrea Papenmeier, Alfred Sliwa, Dagmar Kern, Daniel Hienert, Ahmet Aker, and Norbert Fuhr. 2020. 'A Modern Up-To-Date Laptop' – Vagueness in Natural Language Queries for Product Search, (Aug. 2020). DOI: 10.1145/3357236.3395489.

[9] Francesco Ricci and Quang Nhat Nguyen. 2007. IEEE INTELLIGENT SYSTEMS Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System. Tech. rep. www.computer.org/intelligent.

[10] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2021. Towards Facet-Driven Generation of Clarifying Questions for Conversational Search. In *ICTIR 2021 - Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. Association for Computing Machinery, Inc, (July 2021), 167–175. ISBN: 9781450386111. DOI: 10.1145/3471158.3472257.

[11] A. H.M. Ter Hofstede, H. A. Proper, and T. H.P. Van Der Weide. 1996. Query formulation as an information retrieval problem. *Computer Journal*, 39, 4. DOI: 10.1093/comjnl/39.4.255.

[12] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards conversational search and recommendation: System Ask, user respond. In *International Conference on Information and Knowledge Management, Proceedings*. Association for Computing Machinery, (Oct. 2018), 177–186. ISBN: 9781450360142. DOI: 10.1145/3269206.3271776.

[13] Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Maria Soledad Pera, and Yiqun Liu. 2023. Users Meet Clarifying Questions: Toward a Better Understanding of User Interactions for Search Clarification. *ACM Transactions on Information Systems*, 41, 1, (Jan. 2023). DOI: 10.1145/3524110.

[14] Jie Zou, Evangelos Kanoulas, and Yiqun Liu. 2020. An Empirical Study of Clarifying Question-Based Systems, (Aug. 2020). http://arxiv.org/abs/2008.00279.