Analysing the Effect of Clarifying Questions
on Document Ranking in Conversational Search

Antonios Minas Krasakis   Mohammad Aliannejadi   Nikos Voskarides   Evangelos Kanoulas

University of Amsterdam

{a.m.krasakis,m.aliannejadi,n.voskarides,e.kanoulas}@uva.nl

ABSTRACT

Recent research on conversational search highlights the importance of mixed-initiative in conversations. To enable mixed-initiative, the system should be able to ask clarifying questions to the user. However, the ability of the underlying ranking models (which support conversational search) to account for these clarifying questions and answers has not been analysed when ranking documents, at large. To this end, we analyse the performance of a lexical ranking model on a conversational search dataset with clarifying questions. We investigate, both quantitatively and qualitatively, how different aspects of clarifying questions and user answers affect the quality of ranking. We argue that there needs to be some fine-grained treatment of the entire conversational round of clarification, based on the explicit feedback which is present in such mixed-initiative settings. Informed by our findings, we introduce a simple heuristic-based lexical baseline, that significantly outperforms the existing naive baselines. Our work aims to enhance our understanding of the challenges present in this particular task and inform the design of more appropriate conversational ranking models.

1 INTRODUCTION

The rise of voice-based digital assistants such as Amazon Alexa and Google Assistant, has intensified the need for agents that can hold meaningful conversations with users. Towards this direction, researchers have developed conversational systems that support question-answering and task-oriented dialogue, among others [3, 4]. However, it is often the case that in such information-seeking conversations, users fail to express their information need adequately. This makes the ability of a conversational search system to support mixed-initiative interactions imperative [6, 8]. Such a system can assist users to refine their information need, i.e., by disclosing new information to them [8], or posing clarifying questions [1].

Clarifying questions trigger users' explicit feedback in the form of an answer, and have been shown to improve user experience [1, 2, 6, 9]. In Figure 1, we demonstrate examples of clarification-based conversations appearing in the conversational search dataset Qulac [1]. Specifically, we plot the most frequent user responses ($4 - grams$) to clarifying questions. Responses can be read starting from the circle centre ($START$) and moving outwards (e.g., "no I am looking..."). We observe that user responses often start with a "yes" or a "no", but frequently provide additional information (e.g., "No, I want..."). We hypothesise that this explicit feedback can be used to improve ranking, even when the question asked ranges from being partially relevant to completely irrelevant w.r.t. the information need.

However, using this feedback effectively is challenging due to: (i) the noisy nature of the natural language used in mixed-initiative conversations, (ii) the presence of complex and mixed (both positive & negative) signals that answers often convey (e.g., user: "Yes, but I would like ..."), and (iii) the presence of partially relevant information, in clarification questions that have received negative user feedback (e.g., [$inf.need$]: "Drafting of Declaration of Independence", [$system$]: "Would you like to learn more about Thomas Jefferson?" [$user$]: "No.").

In this paper, we study the effect of the user's feedback in mixed-initiative conversations. We categorise answers w.r.t. their polarity and length. Answer polarity indicates whether the question points to a relevant direction or not, while answer length enables us to (noisily) identify the presence of additional information in the response. We conduct our analysis on the Qulac dataset [1], using a query likelihood model chosen because of its simplicity and transparency [7].

In particular, we aim to answer the following research questions: **RQ1** How does the polarity and informativeness of the user's response to a system's clarifying question affect the performance of a term-matching ranking model in conversational search? **RQ2** How does the length of the clarifying question or the user's response affect the performance of a term-matching ranking model in conversational search? **RQ3** How well can a simple rule-based ranking model perform, compared to existing baselines?

Our contributions are as follows: (i) we study the answers that users provide in a conversation, and categorise them based on polarity and length using simple yet effective heuristics; (ii) we conduct an in-depth analysis of the performance of the QL model, for those answer types; (iii) we study the effect of the clarification question and answer on ranking performance, when used in isolation, combined or ignored, and (iv) we design a simple heuristic ranker, which outperforms the baseline lexical model significantly, solely by incorporating information about the answer type.
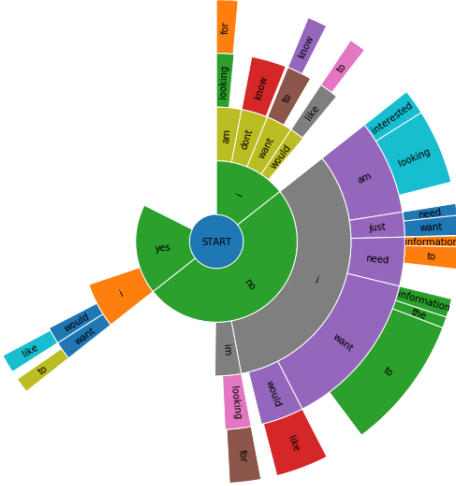
Figure 1: Most frequent 4-grams to answers of clarifying questions. Each word covers an arc proportional to its frequency.

2 EXPERIMENTAL SETUP

Data. We run our experiments on the task of document ranking. Different from typical ad-hoc search, here we assume that after the initial query (Q_0) posed by a user, a clarifying question (Q) is asked by the system, followed by a user's answer (A). We use the document corpus and single-round conversations provided by Qulac, the only large-scale conversational search dataset with clarifying questions we are aware of [1]. Qulac is built on top of the TREC Web Track 2009-2012 data and consists of 10K query-question-answer tuples for 198 TREC topics with 762 facets, with each Q&A corresponding to a facet. We randomly keep 40 topics for testing the performance of the heuristic ranker (section 3.3) and perform all our analysis on the rest of the topics.

Retrieval Model. We use a KL divergence query-likelihood model with Dirichlet prior smoothing [7]. To adapt this model in a conversational setting, we initially experimented with expanding Q_0 with the rest of the clarification round (Q and A). However, preliminary results indicated poor ranking performance, since the importance of the topic (expressed by Q_0) is largely underestimated. To mitigate this issue, we follow [1] and linearly interpolate the original query (Q_0) with the clarification round (Q and A, concatenated) using an equal score interpolation weight of 0.5. We use $NDCG@20$ for evaluation.

3 RESULTS

In this section, we present our experimental results w.r.t. the RQs.

3.1 Answer polarity and informativeness

In this section we aim to answer **RQ1** by measuring the performance of the retrieval model described in the previous section per answer type, using different parts of the conversation round. To do so, we identify two important characteristics of the user's response to a system's clarifying question, namely: (i) the polarity of the answer, and (ii) the informativeness of the answer. Answer polarity gives us a strong signal about the question being asked, and whether it was relevant to the user's information need. We

define four polarity classes: Positive P, Negative (N), "I don't know" (idk), and Other (O).[1] Motivated by Figure 1, we use a simple yet effective heuristic to annotate positive and negative polarity in an answer. More specifically, we tag an answer with P if it contains the term "yes", and with N if it contains the term "no".

3.1.1 Ranking with full conversation. Here we discuss the ranker's performance when using the whole clarification round ($Q_0 + Q + A$). We discuss the results shown in Table 1 per answer type below, focusing on the relative improvement or decrease w.r.t. when only using Q_0.

Positive answers (P). We observe that performance is improved when the answer is positive for both single-word and multi-word answers. This suggests that both the question and answer contain terms that complement the description of the information need and help the QL model rank relevant documents higher.

Negative answers (N). We observe that the performance drops substantially when the user answer is single-word ("no"). In contrast, the user answer is multi-word, the clarification round significantly improves the ranking performance. This indicates that even in the presence of conflicting or misleading clarifying questions, a simple term-matching model can benefit from the user's answer.

Other answers (O). We observe an improvement when the user answer is multi-word, similarly to what we observed in the negative answers.

"I don't know" answers (idk). We observe that it is preferable to ignore the clarification round altogether. This is likely related to the data collection strategy followed in [1], where the crowd-workers were instructed to respond "I don't know" to questions that cannot be answered in the context of the annotation task. Such questions include personal questions (e.g., "Do you have diabetes?") or questions irrelevant to the information need.

3.1.2 Ranking with part of the conversation. In this experiment, we aim to investigate how the performance of the QL ranker is affected when we take into account the clarification round *partially*, i.e., when we ignore the answer ($Q_0 + Q$) or the question ($Q_0 + A$). To this end, we first study how the performance is affected when we have positive multi-word answers, followed by cases where the answer polarity changes.

Positive multi-word answers. In Table 1, we see that ignoring part of the clarification round ($Q_0 + Q$ or $Q_0 + A$) seems to diminish any improvements that would occur in the presence of the entire round ($Q_0 + Q + A$). This indicates that in those cases the value of the clarification does not exist in isolation in questions or answers, but in their combination, implying that the question and the answer could contain complementary information. To further investigate this, in Figure 2 we plot the differences of the QL ranker when using the full clarification round ($\Delta NDCG(+Q + A)$) w.r.t. Q_0. Notice that we also plot the differences (of the same conversation) when using only the question in the x-axis ($\Delta NDCG + Q$), and only the answer in the y-axis ($\Delta NDCG + A$). This enables us to better understand how the performance of the overall clarification is affected by the question and the answer. We see that when $+Q$

---

[1]The *Other* category refers to the answers that did not fit the other three categories.

Table 1: QL performance per answer type and length (NDCG@20). Q_0 is the original query, Q is the clarification question and A is the answer to the clarifying question. Single refers to answers of length 1 and multi refers to answers of length greater than 1. We measure Δ w.r.t. Q_0 and indicate significant differences (2-sided $t - test$, $p - value < 0.05$) with †.

Answer Polarity	Answer Length	# samples	Q_0	$Q_0 + Q$	Δ (%)	$Q_0 + A$	Δ (%)	$Q_0 + Q + A$	Δ (%)
P	single	364	0.191	**0.206**	$+7.9^{\dagger}$	0.191	$+0.0$	**0.206**	$+7.9^{\dagger}$
	multi	1275	0.162	0.162	$+0.0$	0.163	$+0.6$	**0.188**	$+16.0^{\dagger}$
N	single	580	**0.130**	0.106	$-18.5^{\dagger}$	0.129	-0.8	0.106	$-18.5^{\dagger}$
	multi	3791	0.132	0.117	$-11.4^{\dagger}$	**0.162**	$+22.7^{\dagger}$	0.159	$+20.5^{\dagger}$
O	single	47	**0.177**	0.147	-16.9	0.085	$-52.0^{\dagger}$	0.170	-4.0
	multi	1729	0.153	0.132	$-13.7^{\dagger}$	0.169	$+10.5^{\dagger}$	**0.171**	$+11.8^{\dagger}$
idk	multi	346	**0.162**	0.141	$-13.0^{\dagger}$	0.109	$-32.7^{\dagger}$	0.143	$-11.7^{\dagger}$

and $+A$ follow the same trend – either improve (see quarter 2) or harm the performance (see quarter 4), their combination follows with a few exceptions. However, when questions harm but answers improve (see quarter 1), combining them in a conversational round can either improve or harm the performance. In the reverse case (see quarter 3), we observe a more robust positive impact in the combination of question and answer ($+Q + A$). After examining a few examples, we observe that the main cause of this is that questions often contain crucial information, without which the answer is incomplete in isolation (e.g., [Q_0] pork tenderloin [Q]"Would you like to make a **rub** for pork tenderloin?" [A]"Yes, but I need to find a **recipe**").

Q_0 **in positive vs. negative answers.** Another interesting observation deriving from Table 1 is that the performance when only using the original query (Q_0) is quite low for negative answers (second group) compared to positive answers (first group). It is important to highlight here that this performance is completely unaffected by the questions asked and answers received, since it only takes into account the original query. Our hypothesis is that this is a side-effect of the process through which the clarification questions were collected in [1]. Specifically, the crowd-workers were instructed to read the first two result pages of a commercial search engine using the original query (topic) and compose a clarification question based on those. Therefore, this created some bias in the question generation towards more *popular* information needs. Let us note here, that this bias is defined by the search engine used by each crowd-worker and could even be a desirable characteristic of the dataset.

To validate this hypothesis, in Table 2 we analyse the percentage of positive and negative answers received per facet, w.r.t. the performance of the facet using only the original query (Q_0), while ignoring the clarification round (Q and A). We see that there is a significant correlation between facet performance and the percentage of clarification questions pointed towards the correct direction. This suggests that identifying *easier* facets (higher $NDCG@20$) is an easier task for the crowd-workers. This is mainly because naturally the facets with higher performance have more relevant documents at the top of the ranked list, and hence it is more likely for a crowd-worker to observe a facet in a relevant document and ask a question about it.
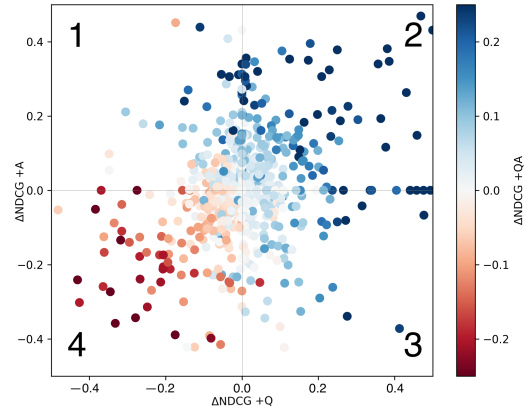

Figure 2: Scatter plot of $\Delta NDCG@20$ **(w.r.t. Q_0) on long, positive answers.**

Table 2: Correlation between $NDCG@20$ **when using Q_0 and percentage of positive, negative answers per facet.**

	Pearson's r	p-value
P	0.173	1.5^{-6}
N	-0.197	4.3^{-8}
rest	0.208	6.8^{-9}

3.2 Clarifying Question and Answer Length

Here, we aim to study the effect of the length of questions and answers on the performance (**RQ2**), as it is unclear whether longer conversations would result in more informative queries and better results. Traditional IR models have been thoroughly tested with keyword-based queries. However, in the mixed-initiative setting many irrelevant terms would appear, even after stop-word removal.

Table 4 shows that significant correlations exist between the length of questions or answers and improvement in $NDCG$ when those are added to the query. This correlation is strongest for answers, which is expected, since users have concrete knowledge of the information need in our setup. For questions, we observe the same trend. This is important, as most answers are negative (see Table 1) and highlights that even questions that receive negative answers contain helpful information for ranking, since they reduce the term mismatch. To further clarify why this happens, we provide examples where questions received negative and single-word

Table 3: Qualitative analysis of clarification rounds with single-word negative answers.

		$\Delta NDCG$
Inf. need	*Drafting of Declaration of Independence*	
Q_0	all men are created equal	
Q	- Would you like to learn more about **Thomas Jefferson**?	+18.53
A	- No	
Inf. need	*Information about Atari arcade games*	
Q_0	atari	
Q	- Would you like to play atari **arcade games** online?	+15.97
A	- No	
Inf. need	*How is a credit score determined?*	
Q_0	credit report	
Q	- Would you like to know about the process of **disputing** credit report?	+13.57
A	- No	

Table 4: Pearson correlation between the length of the question and answer w.r.t. $\Delta NDCG$ from the original query (Q_0).

X	Y	Pearson's r	p-value
# tokens in Q	$\Delta NDCG(+Q)$	0.071	1.1^{-11}
	$\Delta NDCG(+Q+A)$	0.006	5.6^{-1}
# tokens in A	$\Delta NDCG(+A)$	0.130	4.4^{-34}
	$\Delta NDCG(+Q+A)$	0.056	1.1^{-7}
# tokens in $Q+A$	$\Delta NDCG(+Q+A)$	0.049	4.1^{-6}

responses (implying that no meaningful information was provided in the clarification round), but resulted in a big increase in $\Delta NDCG$ (Table 3). We observe that while asking clarifying questions, a number of contextually relevant words appear. For instance, **Thomas Jefferson** is a very relevant entity w.r.t. the drafting of the declaration of independence and their name regularly appears within the relevant results, although the query is not strictly connected to him. Likewise, **arcade games** are relevant words which were previously not mentioned and help improve the ranking. Even further, words such as **disputing** are much less relevant to the information need, but they still appear in the relevant documents because they do contain information about those too (e.g., the disputing credit reports). This indicates that taking into account such conversations has, to a certain extent, similar effects to query expansion and can prove helpful, even when the added terms are not strictly relevant to the information need.

3.3 Heuristic Ranker

Based on the insights from the results presented so far, we design a ranker that heuristically classifies the user responses using polarity and length, and chooses which parts of the clarification round (Q, A) to use (**RQ3**). Following the observations of Table 1, the ranker interpolates the original query (Q_o) with: (i) only A when answers are multi-word negative N, (ii) $Q+A$ for all answers that are positive, or "Other" (O) and multi-word, and (iii) ignores both question and answer (only uses Q_o) elsewhere, where clarifications do not seem to help. We evaluate our model on a held-out test set and report the results in Table 5. We observe improvements compared

Table 5: Comparison of the heuristic ranker with variations of the QL model.

Ranker	NDCG@20
$Q0$	0.148
$+Q$	0.134
$+A$	0.163
$+Q+A$	0.166
Heuristic ranker	**0.171**

to all of the baselines, despite the simplicity of the proposed model (the improvement w.r.t. the best compared model, $Q0+Q+A$, are significant at $p < 0.001$). This suggests that using more advanced techniques to understand and effectively incorporate conversational feedback is likely improve ranking in this task.

4 CONCLUSIONS AND FUTURE WORK

In this work, we provided insights on the task of document ranking with clarification-based conversations. We highlighted the importance of effectively understanding and incorporating explicit conversational feedback, and demonstrated challenges by quantitative and qualitative means. We argued that a more fine-grained treatment of the conversations is crucial to the success of conversational search and propose a heuristic ranking model, which addresses part of the problem despite its simplicity. As future work, we plan to expand our study to account for recently developed neural ranking models for this task [5]. Also, we aim to explore more sophisticated methods for classifying answers and develop ranking models that can better incorporate explicit conversational feedback.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *SIGIR*.

[2] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What Do You Mean Exactly? Analyzing Clarification Questions in CQA. In *CHIIR*.

[3] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*.

[4] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *EMNLP*.

[5] Helia Hashemi, Hamed Zamani, and W Bruce Croft. 2020. Guided Transformer: Leveraging Multiple External Sources for Representation Learning in Conversational Search. In *SIGIR*.

[6] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward Voice Query Clarification. In *SIGIR*.

[7] Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *SIGIR*.

[8] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *ICTIR*.

[9] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N Bennett, Nick Craswell, and Susan T Dumais. 2020. Analyzing and Learning from User Interactions for Search Clarification. In *SIGIR*.