# Asking Clarifying Questions for Conversational Search

**Ansh Bisht** and **Jordan Dickson** and **Jack Harrison** and **Ruben Lazell** and **Andrew Taison**
School of Computing, Engineering and the Built Environment,
Edinburgh Napier University
Matric Numbers: 40527530, 40545300, 40537035, 40679914, 40538519

## Abstract

Short abstract.

## 1 Introduction

Brief context of problem - why clarifying Q's matter. What is the goal. Structure of the paper.

## 2 Related Work

This project is motivated by recent work in conversational search, particularly where ambiguity in user queries is addressed by asking clarifying questions. Across the discussed literature it is commonly noted that users often struggle to express their intent clearly in a single query and find that even simple follow up questions can help search systems return more relevant results. Whilst our system simplifies many of the models used in previous work, it is inspired by several of their core ideas, adapting them to be lightweight for demonstration purposes and time constraints.

### 2.1 Clarification as a Ranking Problem and the Qulac Dataset

Aliannejadi et al. (2019) examine how clarifying questions can improve conversational search, showing that even a single, well targeted question can significantly improve retrieval performance. To support this, they introduce the Qulac dataset, which extends TREC Web Track topics with crowd sourced question and answer pairs organised by query facets (categories). This dataset has been widely used in related research. In addition to the dataset, they propose a retrieval framework in which both documents and candidate questions are retrieved and ranked based on the user's query and conversational context. Question ranking is performed using a BERT model pre-trained on Wikipedia and fine-tuned on Qulac.

Our system builds on this framework, making use of the Qulac dataset for its high quality question-answer pairs and Aliannejadi et al.'s multi-turn conversation history extension dataset. Like their model, we treat clarification as a ranking problem, selecting the most relevant question from a fixed set based on the current context, rather than generating questions from scratch.

### 2.2 User Behaviour and When to Clarify

Alongside ranking-based clarification frameworks, other studies have focused on how user behaviour can affect the system and effectiveness. Two notable examples are the CoSearcher system introduced by Salle et al. (2022) and the risk-aware decision model proposed by Wang and Ai (2022). Whilst these works differ in their objectives, both incorporate user simulators and address when and how clarification should take place in conversational search.

Salle et al. (2022) present CoSearcher, a user simulator designed to evaluate search intent refinement in conversational search by modelling different user behaviours. Their main focus is on testing how different facet-ranking strategies perform under varying conditions such as low user patience or reduced cooperativeness. The authors experiment with a range of ranking models, including LexVec, and use BERT to classify user responses and determine whether further clarification is necessary. Although Sentence-BERT (SBERT) was initially considered, it was not used as LexVec was found to slightly outperform it. However, the gap between LexVec and SBERT was small and they prioritised computational speed. Our system on the other hand is much smaller in scale and can make use of SBERT's sentence level matching and supervised nature, making implementation more straightforward.

Wang and Ai (2022) approach the clarification process as a decision making problem, where the system has to decide whether to ask a question or return a result based on how confident it is. To explore this, they train a reinforcement learning model using a reward function that balances retrieval performance with the length and usefulness of the conversation. Like CoSearcher, they implement a user simulator to represent different behaviours, including tolerance and patience. They use these two parameters to determine how many questions a user will put up with before giving up.

Both studies highlight the importance of choosing not just what to ask, but when to ask it. Whilst our system does not include a user simulator or learning-based decision model, we were influenced by how these papers handle ambiguity and user behaviour. In our case, the decision to clarify or move on is made using a straightforward confidence check from the BM25 module. If the score is high, we assume the summary is a good enough match for the user's intent. If it is low, the system switches to question selection instead. Whilst this is a much simpler method, it follows the same general idea that clarification should depend on context and confidence, rather than being the default.

### 2.3 Sentence-Level Semantic Similarity

Clarifying question selection requires more than just a list of candidates, it also needs a way to evaluate how well each one fits the current query. For that, we use SBERT, introduced by Reimers and Gurevych (2019), which adapts BERT to produce fixed size sentence embeddings that can be compared directly using cosine similarity. Furthermore, Reimers and Gurevych show that SBERT is much more efficient than standard BERT, as it allows sentences to be encoded independently. In tasks such as clustering of sentences, they report a reduction in computation time from hours to seconds. Although newer versions such as Sentence-T5 exist and have been found to perform better in some scenarios, SBERT worked well for our needs (Ni et al., 2021). SBERT's ability to compare semantic similarity efficiently, makes it well suited for ranking tasks.

### 2.4 Dialogue Management with Rasa

Rasa is an open-source framework designed for building conversational AI applications, combining natural language understanding (NLU), dialogue management and a modular architecture that supports flexible system design (Rasa, 2025; Team, 2022). Dinesh et al. (2021) demonstrate how Rasa can be used in task oriented systems, such as answering student questions related to their academic schedules or syllabuses. Their work shows that even relatively simple applications can benefit from Rasa's modular setup, which allows different components to be added or swapped depending on the task. This made Rasa a good fit for our project, where a way to manage conversational flow was required whilst simultaneously relying on external modules for retrieval and clarification.

### 2.5 Question Generation and Generative Approaches

Whilst our system uses a fixed pool of pre-written clarifying questions, recent work has explored generating them from scratch. Wan (2023) address the challenge of generating clarifying questions in cold-start scenarios, where there is a lack of real-world conversational data. They propose a zero-shot method where question generation is guided using query facets and question templates rather than directly concatenating facets to query input as in previous models. They show that this approach produces more useful and natural questions than previous zero-shot baselines.

Zhao et al. (2024) introduce an intent-aware framework for generating clarifying questions. In their study, verbs are extracted from search results using the user's query based on the idea that these verbs represent the user's intent. These extracted verbs are then combined with templates to produce intent-aware questions, with the goal of helping to avoid vague or unhelpful questions. Their results show that tailoring questions to a user's intent can improve both accuracy and user satisfaction.

Although we do not attempt question generation in our project, these studies helped our understanding of how clarification might be handled in more advanced systems beyond retrieval-based approaches. In particular, they highlight the importance of considering intent and context, and have provided a basis for designing our simpler ranking-based method.

## 3 Methodology

System architecture. What models/algorithms used. How are the questions generated. What inputs/outputs are there.

(Aliannejadi et al., 2019)

## 4 Evaluation

How did we check the system works. Describe testing process.

## 5 Results and Discussion

Sample outputs. Summary of system behaviour. What worked well, what didn't? Possible improvements.

## 6 Conclusion

start here.

## Limitations

Required by ACL format, and should be AFTER conclusion. Discuss honest limitations of the work.

## Ethics Statement

Required by ACL format. Could just be a sentence or two. Explicit ethics statement on the broader impact of the work, or other ethical considerations.

## References

2023. Zero-shot clarifying question generation for conversational search. Proposes a <b>zero-shot method</b> to generate clarifying questions <b>without requiring conversational search logs</b> or labeled training data.<br/><br/>The method extracts <b>query facets</b> (e.g., "pictures", "maps", "population" for "South Africa") and <b>guides the question generation process</b> so that the generated clarifying question is <b>contextually relevant</b>.<br/><br/>We could use <b>a predefined set of facets</b> or <b>identify key terms in the query</b> to generate clarifying questions.<br/><br/>Unlike other papers that <b>evaluate question quality based on search performance</b>, this paper <b>evaluates question usefulness independently</b>. Uses <b>automatic metrics (BLEU, ROUGE, METEOR) and human judgments</b> to assess how <b>natural and useful</b> the generated questions are.

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

Tanya Dinesh, Anala M R, T. Terry Newton, and Smitha G R. 2021. Ai bot for academic schedules using rasa. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–6. IEEE.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models.

Rasa. 2025. Welcome to the rasa docs.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Alexandre Salle, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2022. Cosearcher: studying the effectiveness of conversational search refinement and clarification through user simulation. *Information Retrieval Journal*, 25:209–238.

Xaqt Team. 2022. Introduction to rasa.

Zhenduo Wang and Qingyao Ai. 2022. Simulating and modeling the risk of conversational search. *ACM Transactions on Information Systems*, 40:1–33.

Ziliang Zhao, Zhicheng Dou, and Yujia Zhou. 2024. Generating intent-aware clarifying questions in conversational information retrieval systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3384–3394. ACM.

## A Appendix

Possibly not needed.