Asking Clarifying Questions for Conversational Search

Md Rayhan
40645696

Abstract

Human users tend to ask concise natural language queries when interacting with computer systems like conversational bots or search engines. These short and colloquial inquiries frequently present ambiguity, challenging both short-form and long-form answer generation algorithms to provide results with high confidence. While human users anticipate that systems should accommodate their colloquial queries, the act of posing clarifying questions for conversational search ambiguity is regarded as a more natural mode of interaction, rather than generating answers with insignificant confidence. This study explores the implementation challenges of various techniques from the literature for clarification question generation in conversational search setting. Methodologies include- (i) rule-based user intent detection & rule-based utterance, (ii) Intent classifier & generative transformers, and (iii) Large language based Retrieval Augmented Generation(RAG). Career FAQs dataset is chosen as one specific domain for the task. The performance is mostly evaluated by humans for the naturalness of the clarification queries along with Reference- free QUestion Generation Evaluation(RQUGE) as an automated metric. For academics and professionals, this study is significant as it employs various mode of clarification query generation with latest open-source tools and frameworks such as Rasa open-source, fine-tuned GPT etc.

Keywords— Conversational Search System, Ambiguous query, Conversational question answering, Clarification question

1 Evaluation

1.1 Reference- free QUestion Generation Evaluation(RQUGE)

Fabbri et al. (2022) fine-tuned a RoBERT model that utilises a bi-encoder extractive QA loss resulting in encoding information about questions that can be answered by each query term. In RQUGE architecture, (Mohammadshahi et al., 2022) then fine-tuned the model with MOCHA human ratings QA dataset (Chen et al., 2020) to achieve the downstream task of calculating acceptance score κ given the answer term a_c, the candidate question q_c, and the context D.

RQUGE differs from existing automated quality measures of generated question tasks such as BLEU, ROUGE, BERTScore, and BLEURT as the latter require gold references. Because human-annotated gold reference is expensive in this experimental study, we adopted RQUGE metric as it does not require reference questions. Furthermore, RQUGE's judgements shows significant alignment with hu- man judgments on SQuAD, MS- MARCO, and NQ dataset. Unlike QG metrics, RQUGE does not penalise a valid question that may not have high lexical or semantic similarity to the reference question.

Table 1: Average RQUGE score for clarification questions generation

Strategy	Average RQUGE score
(i) Rule-based user intent detection & Rule-based utterance	—
(ii) Intent Detection with classifier	—
(iii) Large language based Retrieval Augmented Generation(RAG)	1.7364

2 Software and Data

Feel free to request access of the research notebook.

References

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2022. Rquge: Reference-free metric for evaluating question generation by answering the question. *arXiv preprint arXiv:2211.01482*.