Zero-shot Clarifying Question Generation for Conversational Search

Zhenduo Wang
zhenduow@cs.utah.edu
University of Utah

Yuancheng Tu
yuantu@microsoft.com
Microsoft

Corby Rosset
corbyrosset@microsoft.com
Microsoft

Nick Craswell
nickcr@microsoft.com
Microsoft

Ming Wu
mingwu@microsoft.com
Microsoft

Qingyao Ai
aiqy@tsinghua.edu.cn
Tsinghua University

Abstract

A long-standing challenge for search and conversational assistants is query intention detection in ambiguous queries. Asking clarifying questions in conversational search has been widely studied and considered an effective solution to resolve query ambiguity. Existing work have explored various approaches for clarifying question ranking and generation. However, due to the lack of real conversational search data, they have to use artificial datasets for training, which limits their generalizability to real-world search scenarios. As a result, the industry has shown reluctance to implement them in reality, further suspending the availability of real conversational search interaction data. The above dilemma can be formulated as a cold start problem of clarifying question generation and conversational search in general. Furthermore, even if we do have large-scale conversational logs, it is not realistic to gather training data that can comprehensively cover all possible queries and topics in open-domain search scenarios. The risk of fitting bias when training a clarifying question retrieval/generation model on incomprehensive dataset is thus another important challenge.

In this work, we innovatively explore generating clarifying questions in a zero-shot setting to overcome the cold start problem and we propose a constrained clarifying question generation system which uses both question templates and query facets to guide the effective and precise question generation. The experiment results show that our method outperforms existing state-of-the-art zero-shot baselines by a large margin. Human annotations to our model outputs also indicate our method generates 25.2% more natural questions, 18.1% more useful questions, 6.1% less unnatural and 4% less useless questions.

CCS Concepts

• **Information systems → Users and interactive retrieval**.

Keywords

conversational search, asking clarifying question, natural language generation

WWW '23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583420

1 Introduction

One common cause of search failure is ambiguity in queries, which refers to the queries with multiple relevant information needs or unclear intent. Ambiguous queries are often the result of users not knowing how to formulate their needs. For example, a user looking for "the Discovery Channel's dinosaur site with pictures and games of dinosaurs" and a user looking for "different kinds of dinosaurs" can both search with the query "dinosaur". Ambiguous queries can also indicate the user is conducting an exploratory search, such as learning or investigating searches [30]. A popular system feature for query ambiguity is search result page diversification [21, 41]. However, it can hardly be applied to searches on devices with small screens or devices with only speech functions by design.

In fact, both scenarios of ambiguous query incentivize the search system to have multi-turn user-system interaction capabilities, i.e., Conversational Search, which has recently become a growing research frontier in the Information Retrieval (IR) community. Conversational Search addresses query ambiguity by arguably its most characterizing feature, mix-initiative interactions. It means that not only the user but also the system can proactively lead the conversation by asking clarifying questions about the user's search intent. These clarifying questions chiefly determine the quality of conversational search. Therefore, existing works have extensively explored various approaches to selecting or generating high-quality clarifying questions.

However, there are two challenges that limits the application of conversational search systems in real-world. First, as a new retrieval paradigm, there isn't any mature online service for open-domain conversational search. The cost of collecting large-scale conversational search logs is still prohibitive, and the building and evaluation of reliable conversational systems is thus difficult, which further increase the difficulty of collecting conversational logs in practice. We refer to this dilemma as the cold-start problem for clarifying question generation. Second, traditional clarifying question generation methods [44, 54] often rely on supervised learning with labeled or artificial conversational logs. It is unrealistic to require such logs to cover all the topics of possible queries, and models trained with incomprehensive datasets could suffer from catastrophic forgetting [32] and inevitably be biased on unseen queries. We refer to this problem as the data bias in conversational search log collection.

Unlike previous studies [13, 39, 44, 47, 54], we explore a new task of clarifying question generation in zero-shot scenarios without the use of conversational search logs. The main idea is to learn a clarifying question generation model directly from large-scale

text data and search engine traffic without collecting or labeling conversational data from a conversational search system. In this way, we can avoid both the cold start and data bias problem from the very beginning. While there have been many studies on zero-shot text generation, as shown in this paper, applying these methods to clarifying question generation directly often produce unsatisfactory results because of two reasons: First, the conventional sequence-to-sequence language generation models cannot efficiently learn the needed correlations between the initial queries submitted by users and the clarifying questions generated by systems. Their generations tend to talk about general topics that are not relevant to the specific search need; Second, existing zero-shot language model generations are usually narratives instead of questions. How to guide the zero-shot model to generate text in question forms that are proper for each search query is still unknown.

To solve the above problem, in this paper, we propose to constrain the clarifying question decoding with search facets. Facets refer to possible subtopics of a query (e.g., "pictures", "map", "populations" are possible facets for the query "I am looking for information about South Africa") that can be effectively extracted from search result pages (SERP) [58], knowledge graphs [50], or other sources [39, 49] in unsupervised manners. Constrained language decoding refines the naive beam-search decoding with the ability to rank facet-containing generations higher, resulting in more questions about the facet. We also initialize the decoding with questioning prompts instead of generating the entire question sentence. Multiple question templates are used in this process and eventually ranked for the best generation.

To demonstrate the effectiveness of our zero-shot system, we compare with multiple existing non-facet and facet-driven baselines, including several state-of-the-art supervised learning methods. They will be finetuned on a training set, which is not accessible by our zero-shot system. Nonetheless, we show that our system significantly improves these baseline methods by a large margin, which implies our system is the best solution for zero-shot clarifying question generation.

During the evaluation phase, we compute automatic metrics [5, 24, 25, 35] and employ humans to provide quality labels for the generations of the compared systems. Our human annotators evaluate the generations from both their *naturalness* [44] from language perspective and *usefulness* [40, 44] from utility perspective. The automatic metric scores are our primary evaluation, from which we conclude that our system is the best. Human annotation results suggest our method generates 25.2% more natural questions, 18.1% more useful questions, 6.1% less unnatural and 4% less useless questions, which aligns with automatic evaluations and reinforces our conclusions with confidence.

We consider the following key contributions of our work:

- We are the first to propose a **zero-shot** clarifying question generation system, which attempts to address the cold-start challenge of asking clarifying questions in conversational search. The zero-shot setting also maximizes the generalizability of our system to serve different search scenarios.
- We are the first to cast clarifying question generation as a constrained language generation task and show the advantage of this configuration. We show that a simple constrained decoding algorithm, even under zero-shot setting, can guide clarifying question generation better than finetuning the model with limited training data. Our work is a compelling demonstration of how large deep models benefit from properly integrating human knowledge.
- We propose an auxiliary evaluation strategy for clarifying question generations, which removes the information-scarce question templates from both generations and references. Results computed this way expose the limitations of the existing default evaluation strategy and provide insights into the actual quality of generated questions.

2 Related Works

Conversational Search Conversational Search refers to the process of information-seeking involving natural language conversations with the search system [55]. It has been identified as one of the most important research area of IR [3, 11]. Recently, a plethora of seminars and tutorials has been given about Conversational Search from different standpoints such as [3, 15–18, 55]. Radlinski and Craswell [37] proposed a theoretical framework for Conversational Search and highlighted mix-initiative as one of its most desired perspective. Later, Zamani and Craswell [53] designed an abstract pipeline for a complete conversational search system. Conversational Search is also closely related to many other research areas such as Task-oriented Dialog System, Conversational Question Answering, Conversational Recommendation, and Chatbot. Anand et al. [3] provided one possible view to connect Conversational Search, Dialog System, and Chatbot. Recently, Zamani et al. [55] defined Conversational Search, Recommendation and Question Answering as subdomains of Conversational Information Seeking.

Resolving Query Ambiguity The query ambiguity problem is an important motivation to promote conversational search over conventional single-turn search. Ambiguous query generally refers to the queries for which the search system cannot confidently identify the user's information need and return search results [21]. Queries can be ambiguous for various reasons, such as containing multiple distinct interpretations or under-specified subtopics [10], anaphoric ambiguity, and syntactic ambiguities [43]. Approaches to clarifying query ambiguity can be roughly divided into three categories: (1) Query Reformulation such as [12, 14, 26, 52] iteratively refines the query; (2) Query Suggestion such as [40, 45, 51] offers related queries to the user; (3) Asking Clarifying Questions such as [8, 38, 39, 54, 57] proactively engages users to provide additional context. While the three approaches share many structural and functional similarities, they cannot be replaced by each another. Because none of them is the best in all scenarios, for example, asking clarifying questions could be exclusively helpful to clarify ambiguous queries without context. In contrast, query reformulation is more efficient in context-rich situations. Query suggestion is good for leading search topics, discovering user needs, etc.

Asking Clarifying Questions Among the approaches to resolving query ambiguity, asking clarifying questions (CQ) is the most studied [21], and is considered as more convenient because of its proactivity [37, 46, 58]. Existing studies about asking CQ can be divided into two main categories: (1) ranking/selecting CQ such as [1, 8, 38], and (2) generating CQ such as [13, 39, 44, 47, 54]. Rao

and Daumé III [39] applied generative adversarial learning in training sequence-to-sequence question generation model. Zamani et al. [54] proposed a rule-based template completion model and two neural question generation models to generate CQ given the query and its aspect. Later in [13, 47], the authors also demonstrated templates could guide CQ generation. Their solutions effectively convert the CQ generation problem to a selection task. Similar to using query aspect, Sekulić et al. [44] also proposed a query facet-driven approach. Recently, Zhao et al. [58] showed such query facets could be extracted from web search results and guide question generation.

Constrained Natural Language Generation Our work applies constrained natural language generation to generating clarifying questions for conversational search. The task of constrained natural language generation was proposed in [4], where the problem was modeled as beam search over 2^C states representing all combinatorial satisfaction states of C constraints. This exponential complexity limited its applications. Hokamp and Liu [19] propose a grid beam search method that groups beams by the number of constraints already satisfied. Miao et al. [33] propose to edit generations using constraints with Metropolis Hastings sampling. Welleck et al. [48] develop a non-monotonic tree-based generation system which can generate texts given constraints at arbitrary positions. Zhang et al. [56] suggest a tree-enhanced Monte-Carlo approach for text generation via Combinatorial Constraint Satisfaction. More recently, Lu et al. [28, 29] propose NeuroLogic Decoding and A* search. Their decoding algorithms incorporate constraints as Conjunctive Normal Forms (CNF) and estimate the viability of each beam to satisfy constraints by sampling their future generations.

3 Zero-shot Facet-constrained Question Generation

This section gives detailed descriptions of our zero-shot clarifying question generation system, which addresses two challenges in naive models for zero-shot clarifying question generation. Our system is zero-shot, meaning that we do not train our system on any training data for clarifying question generation. The generation is also facet-constrained, which implies that we use the search facet in our question generation. A facet is one possible search direction for the ambiguous query; for example, *pictures, map, location, populations* are possible facets for the query "I am looking for information about South Africa". Facet has been considered useful [44, 54, 58] for clarifying question generation since it provides a relevant direction for inquiring about the user intent. Clarifying question generation can be challenging without facets because the generations are often too general and clueless. In [58], Zhao et al. proposes a facet extraction approach, which shows that these useful keywords can be easily extracted from web search results. Previous works also suggest that facets can also be extracted from various sources, including product reviews [39], images [49], or knowledge graphs [50].

The backbone of our system is a checkpoint of the public Generative Pretrained Transformers (GPT-2) [36] pretrained on a seperate large scale text corpus. Originally, the inference objective of GPT-2 is to predict the next token given all previous texts.

$$L = \sum_{t=1}^{T} \log P(x_t | x_{1:(t-1)}, \theta) \tag{1}$$

where $x_{1:(t-1)}$ is the generated sequence by step $(t-1)$, θ is GPT-2 parameters.

One naive method to adapt GPT-2 for clarifying question generation is to append the query q and facet f together as initial texts and let GPT-2 generate a continuation cq as the clarifying question. However, this method faces two challenges. The first challenge is it does not necessarily cover facets in the generation. Previous work [44] proposes a finetuning approach, which trains on a collection of f [SEP] q [BOS] cq [EOS] paragraphs. However, as reported in their work, this structure does not outperform simply using the query alone as input. We analyze the generations of their model and find that the coverage of facet words in these generations is only about 20%. This number implies that simply appending facet words to the input of GPT-2 is highly inefficient in informing the decoder.

The other challenge is that the generated sentences are not necessarily in the tone of clarifying questions. This is because clarifying questions makes up only a small portion of natural language usage. GPT-2 is pre-trained on web texts, most of which are narrative. Even for the questions, they are not necessarily for the purpose of clarifying. As a result, pre-trained GPT-2 often generates relevant factoids following the query and facet.

To explain our proposed system easier, we divide our system into two parts: (1) facet-constrained question generation and (2) multi-form question prompting and ranking. The two parts respectively address the above two challenges of zero-shot GPT-2.

3.1 Facet-constrained Question Generation

In the first abovementioned challenge, we find that existing works struggle to generate facet-related clarifying questions. Unlike these works, we believe that simply appending the facet to the input is inefficient. Instead, our model utilizes the facet words as decoding constraints. Specifically, we see the task of generating facet-related questions as a facet-constrained language generation problem, i.e., to use the facet words as constraints during generation decoding. To encourage the decoder to choose generations containing more facet words, we employ an algorithm called NeuroLogic Decoding [29].

NeuroLogic Decoding is based on beam-search decoding. In each decoding step t, assuming the already generated candidates in the beam are $C = \{c_{1:k}\}$, where k is the beam size, $c_i = x^i_{1:(t-1)}$ is the ith candidate, and $x^i_{1:(t-1)}$ are tokens generated from decoding step 1 to $(t-1)$, NeuroLogic Decoding works in the following steps:

(1) Generate the next token distributions $x^i_t \sim P(x^i_{1:(t-1)})$ for each candidate in the beam with GPT-2. Assume that the vocabulary size is $|V|$, then this will create $k \times |V|$ new candidates.
(2) **Pruning Step:** Among these candidates, discard all but candidates that are in both in the top-α tokens in terms of $p(x_{1:t})$ and the top-β in terms of number of facet words contained by $x_{1:t}$.