

Ziliang Zhao Gaoling School of Artificial Intelligence Renmin University of China Beijing, China zhaoziliang@ruc.edu.cn

ABSTRACT

Asking multi-turn clarifying questions has been applied in various conversational search systems to help recommend people, commodities, and images to users. However, its importance is still not emphasized in the Web search. In this paper, we make a step to extend the multi-turn clarification generation to Web search for clarifying users' ambiguous or faceted intents. Compared with other conversational search scenarios, Web search queries are more complicated, so clarification should be generated instead of being selected which is commonly applied in current studies. To this end, we first define the whole process of multi-turn Web search clarification composed of clarification candidate generation, optimal clarification selection, and document retrieval. Due to the lack of multi-turn open-domain clarification data, we first design a simple yet effective rule-based method to fit the above three components. After that, by utilizing the in-context learning and zero-shot instruction ability of large language models (LLMs), we implement clarification generation and selection by prompting LLMs with demonstrations and declarations, further improving the clarification effectiveness. To evaluate our proposed methods, we first measure whether our methods can improve the ability to retrieve documents. We also evaluate the quality of generated candidate facets. Experimental results show that, compared with existing single-turn methods for Web search clarification, our proposed framework is more suitable for open-domain Web search systems in asking multi-turn clarification questions to clarify users' ambiguous or faceted intents.

CCS CONCEPTS

• Information systems \rightarrow Search interfaces.

KEYWORDS

Search Clarification, Clarifying Question, Conversational Search

ACM Reference Format:

Ziliang Zhao and Zhicheng Dou. 2024. Generating Multi-turn Clarification for Web Information Seeking. In *Proceedings of the ACM Web Conference* 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589334.3645712

WWW '24, May 13-17, 2024, Singapore, Singapore

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0171-9/24/05...\$15.00 https://doi.org/10.1145/3589334.3645712 Zhicheng Dou* Gaoling School of Artificial Intelligence Renmin University of China Beijing, China dou@ruc.edu.cn



Figure 1: A process of asking multi-turn clarifying questions.

1 INTRODUCTION

Search clarification has become an important part of conversational Web search [4, 44, 45]. When a user issues an ambiguous or faceted Web query, the system delivers a clarification pane composed of a *clarifying question* and several *candidate facets* representing potential intents for the user to select [43]. The query can be refined according to the user's selection to retrieve a new set of documents, and the system can continue to clarify. A typical clarification process is shown in Figure 1. It can be seen that the process is essentially a *multi-turn* interaction or conversation between the user and the system. The multi-turn mechanism is especially emphasized when the user's search intent is complicated or less specific, while single-turn clarification cannot satisfy the user's need [5, 23, 36].

Nowadays, multi-turn clarification has been applied in many Information Retrieval (IR) scenarios [26]. For example, in Conversational Recommender Systems (CRS), the system asks the user about the attributes of commodities turn-by-turn for recommendation [6, 18, 19, 50]. In conversational search systems, the system asks the user to deliver more information about her needs [1, 10, 15]. Besides, multi-turn clarification has also been applied in other closedomain applications like interactive classification [42] and twentyquestions task for picture guessing [40]. These studies discuss the

^{*}Zhicheng Dou is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

importance of multi-turn clarification in IR systems, and inspire us that open-domain Web search should also include such a *multi-turn* process to help users better find the information they want.

However, existing Web search clarification studies [31, 37, 43-45, 48] focus on generating a single pane [11, 12, 27], and continuing the process by restarting a clarification based on the updated query to achieve a *pseudo* multi-turn process. Although this approach can generate reasonable clarification panes, it is not the optimal choice in a multi-turn scenario, in which does not cater to the user's intent, or requires lots of turns to find the user's intent. This is because existing methods only adopt greedy strategies and cannot consider the global relation of multi-dimensional potential panes in each turn. For example in Figure 2, we assume that the user intent "download google chrome exe 64 bit" can be achieved by different clarification paths. Since existing methods often ignore the potentially multi-dimensional essence at each turn, they can easily go through some wrong or long paths (the red one). In this paper, we make the first step to try to extend existing multi-turn clarification scenario into the Web search by generating several potential clarification panes at each turn and select the optimal path (the blue one in Figure 2) to achieve multi-turn Web clarification.

To formally describe the problem we are studying, we refer to existing close-domain clarification scenarios, especially CRS, and identify a main goal and a framework composed of three main components. Our main goal is to help users search for satisfying documents with as few clarification turns as possible. To achieve this, we formalize the multi-turn process, and then define a framework including three components: (1) Clarification candidate generation: Since the target (documents) to be recommended in Web search is dynamically updated with each turn of user selection, it is necessary to dynamically generate potential clarification panes in each turn. (2) Optimal clarification selection: Similar to CRS, each query may have multiple potential clarification panes. After the clarification candidate generation, a certain strategy needs to be applied to select the optimal clarification pane composed of a question and several clickable facets. (3) Document retrieval: After the user submits the query or clicks one facet for clarification at each turn, a new document list should be retrieved.

Due to the lack of corresponding studies and datasets, we first design a rule-based method MulClari-Rule to fit our proposed framework. The method first generates independent candidate facets with the query and search result pages, clusters them to construct the candidate facets set, and then generates a question based on existing question generation algorithms [43, 48]. After that, it selects the optimal clarification at each turn by applying an entropy-based strategy. However, the rule-based method is weak at obtaining abundant contextual information between different turns. Nowadays, large language models (LLMs) have performed well in many Natural Language Processing (NLP) tasks. Their zero-shot instruction and in-context learning ability lead to strong ability in multi-turn conversation modeling. With their abilities, in this paper, we further propose an LLM-based method MulClari-LLM prompted by human-designed demonstrations, to implement the components above and achieve multi-turn Web search clarification.

It is challenging to evaluate multi-turn Web search clarification due to the diversity of user queries. In this paper, we propose evaluating the multi-turn clarification from two perspectives: (1) Since our goal is to provide users with documents in as-few-as-possible clarification turns, we first evaluate the ability to retrieve satisfying documents for ambiguous or faceted queries. To achieve this, we rely on the Qulac dataset [1] together with its relevance judgments and evaluate the document ranking results after clarification by MRR, NDCG, and P@1. (2) We also evaluate the quality of the first-turn generated clarification panes using single-turn evaluation metrics based on the MIMICS dataset [44]. The experimental results demonstrate that, first, compared with single-turn baseline models, our proposed multi-turn strategy can find the information the user wants in as few turns as possible. Second, the goals of multi-turn scenarios are significantly different from those of single-turn scenarios, and the definition of high quality in single-turn scenarios may not necessarily meet the goals of multi-turn scenarios.

The main contributions of this paper include:

- To our best knowledge, we are the first to try to extend the single-turn Web search clarification to multi-turn, enriching the existing conversational search scenarios.
- We define the goal of Web search clarification and the generation process including clarification candidate generation, optimal clarification selection, and document retrieval. We further design a rule-based and an LLM-based method to implement our proposed framework.
- We design two approaches to evaluate our proposed methods. The experimental results show that our methods can retrieve satisfying documents in as few turns as possible.

2 RELATED WORK

Open-domain Search Clarification. Aliannejadi et al. [1] first proposed search clarification in conversational search systems. However, it could only retrieve and select questions and let the user respond by natural language, which was not suitable for Web search. In Web search systems, the user's query is very complex, so the question selection is extremely difficult because the question in the limited dataset can not satisfy the large-scale Web query. More recently, Zamani et al [43, 45] first emphasized the importance of clarification in the field of Web information retrieval. They proposed the MIMICS [31, 44] for the Web search clarification. In Web search clarification, the clarifying question [38] together with the candidate facets [8, 13, 14] are generated instead of selected (in conversational search systems) [1] or constructed by some rules (in CRS) [18], ensuring its essence of open-domain. Besides, some close-domain clarification generation methods [24, 25, 34, 41] also show their strength in some question-answering communities, yet they also cannot cover a wide range of Web search queries.

Conversational Recommender System (CRS). CRS aims at mining users' preferences through multiple turns of natural language conversation, so as to recommend to users the facets they may be interested in. Sun and Zhang [30] first proposed the concept of CRS and considered its several important issues. Later, researchers had tried various algorithms to perform conversational recommendation [3, 6, 16, 18, 19, 42, 46, 49, 50]. It is worth noticing that, in CRS, each commodity to be recommended has its own attribute set, such as the brand and CPU model of the computer. Therefore, it is an important step to select an attribute from them to ask the user. In this process, some strategies like max-entropy have been widely



Figure 2: Comparison of different potential clarification paths. In this paper, we emphasize the importance of satisfying the users' information needs with an as-short-as-possible path (clarification turns).

applied, or they can be achieved by applying more deep-inside natural language understanding models. However, different from CRS, in Web search, unlike static commodities, retrieved documents are real-time-updated with a huge quantity, and do not have fixed attribute sets, so it is necessary to dynamically *generate* the attribute set for each document.

Other Clarification Scenarios. In addition to the main-stream open-domain and close-domain search clarification scenarios mentioned above, search clarification can also be applied in some other novel scenarios. For example, Yu et al. [42] studied how to classify objects interactively. They gradually determined what users were thinking by asking them multiple turns of questions to clarify their intents in a multiple-choice manner. White et al. [40] proposed a novel scenario: guessing which image the user has in mind within 20 Yes/No questions. It also borrowed some ideas from CRS. Zhang and Zhu [47] studied about what information was omitted when publishing products on e-commerce websites, and informed publishers in the form of questions. Recently, Shi et al. [29] studied whether a certain step in gaming intelligence should be taken or a question asked of the user to resolve ambiguity. These studies provide potential application scenarios for search clarification.

3 MULTI-TURN WEB CLARIFICATION

3.1 The Necessity of Multi-turn Clarification

Existing studies of Web search clarification focus on generating high-quality clarifying questions and candidate facets given a user query in a single-turn setting. However, when the user intent is complex, single-turn clarification is not the best choice to satisfy the complicated user intent because single-turn methods are not aware of global potential clarification candidates, thereby making it easy to go through wrong paths in potential clarification distributions as shown in Figure 2. By applying multi-turn clarification, the system can gradually clarify the user's complicated search intent turn-byturn to improve their search efficiency and experience. For example, in CRS, the system asks the user about the attributes of commodities turn-by-turn for recommendation [6, 18, 19, 50]. In conversational search systems, the system asks the user to let the user deliver more information about her needs [1]. Besides, multi-turn clarification has also been applied in other close-domain applications like interactive classification [42] and twenty-questions task for image

guessing [40]. These application scenarios inspire us to expand multi-turn search clarification to the Web search.

In Web search, the situation is more complex. First, Web search queries are *open-domain*, covering all kinds of real-world intents. This makes it sometimes difficult to understand the user intent and emphasizes the very importance of multi-turn clarification compared with other close-domain scenarios such as CRS. Therefore, the clarification pane should be *generated* instead of *selected or constructed by some rules*. Second, the item to be recommended in Web search is *large-scale documents or natural language passages* instead of a set of people, images, or commodities with clear attributes. It is difficult to represent a document or passage using existing attribute-based approaches. Therefore, it also emphasizes the necessity for mining attribute sets for a specific document.

3.2 **Problem Reformulation**

To solve the problem of multi-turn Web search clarification, we need to first define and formulate this task. (1) **First**, the user submits an original query q, her target is relevant document set D^q . (2) **Next**, the system interacts with the user with multi-turn clarification: At each turn t, the system provides a clarification pane $C_t = (Q_t, S_t)$ composed of a clarifying question Q_t and a set of candidate facets S_t , and then the user selects a candidate facet and forms a new query q_t . After that, the system retrieves a new set of documents D_t with q_t and gets $A_t = (q_t, D_t)$ and generates a new clarification pane C_{t+1} based on A_t . (3) **Finally**, after the above k turns, we calculate the performance of retrieving documents from a large document set. The whole process can be represented formally:

$$\begin{array}{l} q, C_1, A_1, C_2, A_2, \cdots, C_k, A_k, \\ = q, (Q_1, S_1), (q_1, D_1), \cdots, (Q_k, S_k), (q_k, D_k). \end{array}$$
(1)

Our task is to retrieve the documents (passages) satisfying the user's information need within a few clarification turns. It is worth noticing that the formulation is similar to the multi-turn clarification in conversational search systems [1, 10]. However, in conversational search systems, first, the user can only respond to the system by inputting a new sentence of natural language, which is time-wasting and experience-effecting. In our scenario, the user can respond just by clicking a candidate attribute staying consistent with existing single-turn Web search clarification [43–45], which is convenient for the user. Second, since Web search is complicated,

WWW '24, May 13-17, 2024, Singapore, Singapore

Ziliang Zhao and Zhicheng Dou



Figure 3: Our proposed framework taking the query "lost" as an example. It is composed of three main components: (1) Clarification candidate generation, (2) Optimal clarification selection, and (3) Document Retrieval.

we should **generate** instead of **selecting** clarification panes. In fact, existing studies in conversational search systems focus on selecting clarification panes from a question bank [1, 10], lacking universality for different queries. In contrast, for each specific query, we borrow the idea from CRS by generating various clarification panes according to the query and selecting the optimal one.

3.3 Framework Overview

To achieve our main goal, we design a framework containing three components, including (1) Clarification candidate generation: Unlike static attribute sets in CRS, the target to be recommended in Web search is dynamically updated with each turn of the facet selected by the user. Therefore, it is necessary to generate multi-dimensional candidates to have a global perspective of all potential panes. (2) Optimal clarification selection: Similar to CRS, we need to select the optimal clarification pane to deliver to the user. (3) Document *retrieval*: Retrieving relevant documents based on the user query and user-selected facets. To implement the above three components, we first design a rule-based method MulClari-Rule. This method relies on retrieved documents to generate multi-dimensional candidate clarification panes, and then select the optimal pane using the maximum information gain (or max-entropy, the same as below) strategy. Since the rule-based method makes it difficult to capture multi-turn semantic information, we further design another method MulClari-LLM leveraging the strong natural language understanding and generation ability of LLMs [2, 9, 22, 32, 33] to implement the clarification generation and selection process.

3.4 MulClari-Rule

3.4.1 Rule-based Clarification Candidate Generation. For a query q, we first obtain its corresponding potential multiple sets of candidate facets as shown in the middle part of Figure 3. This is done to allow the system to be aware of all potential facets of the query, staying consistent with existing systems such as CRS [18]. To obtain the multi-dimensional clarification candidates, we designed a method MulClari-Rule that combines a generative model and well-designed manual rules. This method consists of three steps: (1) First, generate an **independent facet candidate set** I^c containing many individual facets that do not have relations. (2) Then, since it is deemed that high-quality facets can be found in search result documents [43, 48], we only select the facets that have appeared in the

corresponding documents of the query as *I*^s, thereby **filtering out** some low-quality or wrongly-generated facets. (3) Finally, since we need facet dimensions divided by groups as shown in Figure 3, we **cluster the facets** using co-occurrence information from MIMICS and select high-quality facet dimensions and high-quality facets in each dimension as the final result.

For the first step, we use a Seq2Seq model (like BART [17]) to generate independent candidate facets relying on its strong generation ability. We first collect data pairs from MIMICS dataset [44] denoted as $(q, D) \rightarrow S_i$, where q is the user query, D is the top-10 search snippets, and S_i is one facet for a query. A query in MIMICS corresponds to up to five facets. In order to provide sufficient candidates, we use beam search to take the first 100 beams of facets generated by Seq2Seq as the preliminary facet candidate set I^c :

$$I^{c} = \text{beam_search}_{100} (\text{BART} (q, D)), \qquad (2)$$

where q and D are the same as above. These two are usually concatenated for facet generation [21, 27]. After that, we delete facets in I^c that do not appear in D to obtain a selected facets candidate set I^s , to ensure the quality of the facets. We apply BART as the instance for generation yet it can be replaced by some other models like T5. Besides, we add the disambiguation entities of the query in WikiData into I^c to deal with ambiguous queries. We also add provided intent keywords for queries in Qulac dataset [5] to fill the gap between Qulac dataset and MIMICS dataset.

Since facets in I^s are independent, we need to cluster related facets together to construct multi-dimensional facets. For example in Figure 3, for the query "lost", it can generate five facet dimensions, including (1) the information of this series, (2) roles, (3) seasons, (4) actors, and (5) other USA series. The facets in one dimension show a high correlation. To achieve this, we can rely on the cooccurrence information in the MIMICS dataset [44] and build a graph $G^M = \langle V^M, E^M \rangle$ containing the co-occurrence frequency. In the graph, one node V_i^M means one facet, and one edge E_{ij}^M represents the co-occurrence frequency between the facet V_i^M and V_j^M . Then, we cluster the generated facets candidate set I^s based on G^M . Specifically, we initialize a new graph G^I . For two facets I_i^s and I_j^s , if they exist in V^M and E_{ij}^M exists, we add the two nodes into G^I as V_i^I and V_j^I , and then build an edge E_{ij}^I between these two nodes. Finally, we take out all k fully connected components

WWW '24, May 13-17, 2024, Singapore, Singapore



Figure 4: Our proposed LLM-based method MulClari-LLM.

in G^I as k dimensions of generated facets, sort the facets in each dimension in descending order of their frequency in G^I , and select the top-5 facets with the highest frequency in each dimension as the results, staying consistent with that in MIMICS dataset.

We conducted additional processes to ensure the quality of the generated facets. First, in order to avoid repeating clarifications in multiple turns, we record the facets presented to users in history and delete the clarification candidates containing these facets in subsequent generations. In addition, after using BART for singlefacet generation, we use the part-of-speech analysis tool Stanza [20] to convert all plural facets into singular and perform deduplication.

Besides, given the user query q and corresponding facets S, we can apply some robust clarifying question generation algorithms [28, 39, 43, 48] to generate a question $Q: (q, S) \rightarrow Q$, to form a whole clarification pane as shown in Figure 1. For a specific query, several panes could be generated as clarification candidates.

3.4.2 Rule-based Optimal Clarification Selection. After generating multi-dimensional clarification candidates, we need to deliver the optimal one to the user. Due to the lack of relevant data, we aim to first propose a simple yet effective rule-based approach to achieve this goal. To our best knowledge, in some systems like CRS [6, 19, 40, 46], max-entropy (ME, or max information gain) strategy has been widely applied for selecting the optimal attribute [40]. Inspired by the ME strategy, we also design an optimal clarification selection strategy relying on the appearance distribution of the facets in the top search results of the query. We aim to select the clarification pane most widely distributed in retrieved documents. Specifically, for one clarification pane, we first list documents where each facet appears, as shown in the middle part of Figure 3. We then calculate the information gain $gain(\cdot)$ for each facet *s* as gain(s) =H(D|q) - H(D|q, s) where q is the user query, s is one facet, and D is the document set. Due to the large number of candidate documents, we only selected the top 50 documents retrieved by BM25 as the candidates. H(D|q) is set to be 1, and H(D|q, s) is the rate of the documents that do not contain s in top-50 documents of q.

We further define the information gain of a dimension of facets S_i as the average information gain for each facet in this dimension:

$$\operatorname{gain}(S_i) = \frac{1}{|S_i|} \sum_{s \in S_i} \operatorname{gain}(s).$$
(3)

Finally, we select the dimension with the highest information gain as the optimal clarification. The selected facet dimension together with the generated question are then delivered to the user.

3.4.3 Document Retrieval. We implement and apply the BM25 algorithm to retrieve relevant documents of the query ("lost" for example) and return the newly generated clarification pane and the retrieved documents to the user. When the user clicks one of the provided facets ("episodes for example in Figure 3), the query will be updated by concatenating the original query and the clicked facet ("lost episodes" for example) to retrieve a new document list.

3.5 MulClari-LLMs

MulClari-Rule has two limitations: First, it is still not good at modeling multi-turn context. When it selects the optimal clarification at each turn, it just focuses on maximizing the information gain without considering the pre-context. Second, we assume that highquality facets should occur in top-retrieved documents, but it cannot cover all potential high-quality facets [27]. Recently, LLMs have performed well in various NLP tasks due to their strong in-context learning and zero-shot instruction ability. The ability of LLM can essentially help model our multi-turn clarification process. Therefore, besides the MulClari-Rule, we further propose an LLM-based method MulClari-LLM to try to improve the multi-turn clarification effectiveness. Specifically, as for the clarification candidate generation and optimal clarification selection, we design prompts with additional information and let the LLM generate clarification candidates and select the optimal one. For the document retrieval module, we still apply BM25, staying consistent with MulClari-Rule.

Table 1: Evaluation results of document retrieval of the original query, the baseline methods, and our proposed two methods. " \dagger " denotes that the proposed method significantly outperforms all baseline models with p < 0.05.

Model	MRR	P@1	nDCG@1 (@5 @20
original q	0.1836	0.1027	0.0863 0.0	0.0914
Generation- qD	0.2114	0.1218	0.1081 0.1	0.1012
Generation-q	0.2031	0.1169	0.0963 0.0	0.0875
Labeling	0.1848	0.1032	0.0871 0.0	0.0735
Classification	0.1729	0.0903	0.0778 0.0	0.0710
Extraction	0.1681	0.0843	0.0721 0.0	0.0644
MulClari-Rule MulClari-LLM	0.2286 [†] 0.2374 [†]	0.1332 [†] 0.1389 [†]	0.1241 [†] 0.1 0.1263 [†] 0.1	245 [†] 0.1143 [†] 239 [†] 0.1167 [†]

3.5.1 LLM-based Clarification Candidate Generation. The LLMbased method MulClari-LLM is illustrated in Figure 4. The process is shown on the left side. We first use a well-designed natural language prompt to let the LLM cluster the independent facet set in Section 3.4.1 to build multiple dimensions of facets. The prompt first describes the form of single-turn clarification and the need for multi-turn clarification. After that, we give the model several demonstrations to help guide the LLM generation. As shown in the middle part of Figure 4, the model outputs multiple dimensions of facets and retrieves documents with BM25 simultaneously.

3.5.2 LLM-based Optimal Clarification Selection. After the LLMbased clarification candidate generation, the LLM can generate several dimensions of related facets. Since our purpose is to let MulClari-LLM select one dimension that is deemed the optimal one for retrieving better documents, we further provide the model with the top 50 retrieved documents with BM25 as pseudo relevance feedback. We let the LLM select one clarification from the generated candidates to deliver to the user as shown in Figure 4.

4 EXPERIMENTS

4.1 Evaluation Data

For multi-turn clarification, we use the Qulac dataset [1] to evaluate the document ranking results. This dataset contains 198 ambiguous or faceted queries, each of which has a group of corresponding labels for related and unrelated documents. For the evaluation of the quality of the first-turn clarification pane, we use a subset of MIM-ICS [44] to evaluate the quality of facets. In fact, our experiments essentially combine the advantages of two mainstream clarification datasets Qulac and MIMICS. The advantage of Qulac is that its form is closer to human dialogue, and there are corresponding annotations for relevant and irrelevant documents for each query, which is convenient for evaluation. The advantage of MIMICS is that it consists of a large number of real-world queries sampled from a search engine, making it more suitable for Web Search.

4.2 Evaluation Metrics

For multi-turn clarification, the effectiveness is measured by considering the performance of retrieval after updating the user query. Following existing studies [1], we apply several groups of evaluation metrics to evaluate the document ranking results, including (1) mean reciprocal rank (MRR), (2) precision of the top 1 retrieved document (P@1), and (3) normalized discounted cumulative gain for the top 1, 5, and 20 retrieved documents (nDCG@1, nDCG@5, nDCG@20). The three groups of evaluation metrics are important in different search scenarios, including traditional search engines (MRR, nDCG@5, and nDCG@20) and conversational search systems with limited screens (P@1 and nDCG@1).

Furthermore, we also evaluate the quality of the first clarification pane generation. Therefore, we use four sets of single-turn evaluation metrics widely used in existing studies [11, 12, 27] to evaluate the generated facets. (1) Term overlap (Precision, Recall, and F1): the term overlap score evaluates the lexical similarity between the generated facets and ground-truth facets by comparing their same terms. (2) Exact match (Precision, Recall, and F1): the exact match score evaluates whether the generated and ground-truth facets are the same. (3) Set BLEU (1, 2, 3, and 4) scores: the BLEU score calculates the n-gram overlap between two sets of texts. It is widely applied in various NLP tasks. (4) Set BERT (Precision, Recall, and F1) score: the Set BERT score calculates the similarity between two sets of texts from a semantic perspective, which makes up for the shortcomings of the previous three metrics.

4.3 **Baseline Methods**

For multi-turn clarification, we implement four types of PLMbased [7, 17, 35] single-turn clarification generation approaches [27] to obtain clarification panes, including generation, labeling, classification, and extraction. The four approaches are trained with the MIMICS dataset with different paradigms, and they perform well in single-turn Web search clarification. To extend them to fit the multi-turn clarification setting, after the user clicks one facet, the updated query will be used independently to retrieve a new document list and generate a new clarification pane.

For evaluating the first clarification pane, we only evaluate the generated facets. This is because, in our proposed methods, the quality of the clarifying question is determined by the facets [43]. Besides, existing clarifying question generation methods have been good enough for generating clarifying questions that are not necessary for evaluation. For candidate facets, we also apply the four well-performing approaches mentioned above as baselines.

4.4 Implementation Details

For our evaluation data, we obtain Qulac¹ and MIMICS² from their websites respectively. We also obtain the annotation of the document relevance of Qulac for evaluating the document ranking results as well as the top-10 search snippets of each query from Bing search engine for enhancing the query. For the BART model in Section 3.4.1 and the baseline models to be compared, we optimize the BART-base model³ with AdamW optimizer with the learning rate of 1.0×10^{-4} and the batch size of 32. We hold out 10% of the MIMICS data as a validation dataset. Deep learning libraries including PyTorch and Transformers are used for training, beam searching, and validation. In the training, validation, and evaluation for facet generation, we remove the facet terms in the MIMICS dataset that

¹Qulac dataset: https://github.com/aliannejadi/qulac

²MIMICS dataset: https://github.com/microsoft/MIMICS

³BART-base: https://huggingface.co/facebook/bart-base

Table 2: Evaluation results for facets generation. The best result for each metric is marked in bold. "†	" denotes that the proposed
method achieves significant improvement compared with all baseline methods with $p < 0.05$.	

	Term Overlap		Exact Match		Set BLEU			Set BERT					
Model	Prec	Recall	F1	Prec	Recall	F1	1-gram	2-gram	3-gram	4-gram	Prec	Recall	F1
Generation-qD	0.1423	0.1457	0.1440	0.0936	0.0912	0.0924	0.2147	0.1885	0.1724	0.1623	0.5333	0.5395	0.5364
Generation-q	0.1351	0.1375	0.1363	0.0875	0.0912	0.0893	0.2084	0.1816	0.1686	0.1510	0.5351	0.5328	0.5339
Labeling	0.1615	0.1833	0.1717	0.1024	0.1275	0.1136	0.2192	0.1897	0.1767	0.1622	0.5371	0.5338	0.5354
Classification	0.0938	0.0956	0.0947	0.0512	0.0584	0.0546	0.0849	0.0766	0.0662	0.0608	0.5415	0.5382	0.5398
Extraction	0.1034	0.1522	0.1231	0.0463	0.0531	0.0495	0.2065	0.1771	0.1633	0.1529	0.5369	0.5413	0.5391
MulClari-Rule	0.1528	0.2527^{\dagger}	0.1904^{\dagger}	0.0398	0.0575	0.0470	0.2113	0.1825	0.1544	0.1323	0.5318	0.5359	0.5332
MulClari-Rule-Best	0.3268 [†]	0.4129^\dagger	0.3648^\dagger	0.1081	0.1925^\dagger	$\boldsymbol{0.1412}^\dagger$	0.3408 [†]	0.2803^\dagger	0.2592^\dagger	0.2410^\dagger	0.5413	0.5399	0.5405
MulClari-LLM	0.0803	0.0885	0.0842	0.0086	0.0079	0.0082	0.1053	0.0764	0.0524	0.0389	0.5277	0.5302	0.5287
MulClari-LLM-Best	0.1414	0.1726	0.1554	0.0622	0.0814	0.0705	0.2173	0.1862	0.1689	0.1557	0.5359	0.5332	0.5346

overlap with the query terms. For example, for the query "watches" and one of its corresponding facets "rolex watches", we modify the facet as "rolex". We conduct this to ensure the consistency of the output. For the LLM, we use the GPT-3.5-Turbo⁴ in this paper, which can be replaced by other LLMs.

4.5 Experimental Results

4.5.1 Multi-turn Clarification Evaluation. We first evaluate whether the documents returned after k clarification turns are more satisfactory to the users. In this section, we first set k = 2, which is about the "inflection point" value of the clarification turns. In other words, after more than two turns of clarification, the improvement rate of document retrieval performance slows down. We also conduct experiments with the increase of the turn k, which will be discussed in Section 4.7. In addition, since a clarification pane contains multiple candidate facets for users to click, and clicking each candidate facet will retrieve a different list of documents, we concatenate each candidate facet and query provided by each clarification pane to generate a new query retrieved document list, and average the evaluation metrics generated by these document lists to obtain the final score of the current clarification pane. For the multi-turn situation (k > 1), we consider all possible combinations of the clarification paths and select the optimal one as the final evaluation result.

Table 1 presents the results of the document retrieval after two turns of clarification. The Generation-qD and -q means the input is composed of the query q and the snippets D, and only the query q respectively. We can conclude from the results that, (1) **First**, most of the baseline models and our proposed methods perform better than the original query in retrieving documents after two turns of clarification. This confirms that search clarification plays an important role in Web search. (2) **Second**, our proposed two methods outperform all baselines significantly with p < 0.05. This result demonstrates our main conclusion: compared to existing single-turn clarification methods, our proposed multi-turn strategy is more suitable for clarifying users' ambiguous or faceted intent in Web conversational search. (3) **Third**, compared with MulClari-Rule, the LLM-based model MulClari-LLM achieves better results in most of the evaluation metrics. As discussed in Section 3.5, MulClari-Rule is

not good at modeling multi-turn interactions, while MulClari-LLM is suitable to model the multi-turn process essentially. Therefore, it shows better performance than MulClari-Rule in a multi-turn document retrieval setting.

4.5.2 First-turn Clarification Pane Evaluation. Our above experiments have shown that compared to the existing single-turn Web Search Clarification methods, our proposed multi-turn methods can retrieve better documents for users. However, in addition to evaluating the quality of retrieving documents, we are also interested in ensuring the quality of the generated clarification candidates, especially the facets. To achieve this, we evaluate the first-turn clarification quality because the first-turn clarification quality significantly determines the quality of subsequent turns and plays a very important role at the beginning. We want to answer two questions: (1) Is the dimension of multi-turn selection consistent with the ground truth in existing real-world single-turn datasets (such as MIMICS)? In other words, we want to understand whether the goal of our multi-turn method is consistent with that of single-turn methods and data. (2) Since we first generate multi-dimensional clarification candidates, we would like to observe whether these candidates include single-turn ground truths, even if it is not selected as optimal clarification to deliver to the user.

The experimental results are shown in Table 2. We can summarize from the result table that, (1) Compared to MulClari-LLM, the clarifications selected by MulClari-Rule are usually more close to the MIMICS dataset, showing higher performance in most of the evaluation metrics. However, MulClari-Rule's performance in multi-turn document retrieval (see Table 1) is not as good as the improvement brought by MulClari-LLM. This indicates that the single-turn clarification generation in existing studies is less effective than our proposed model in improving document retrieval ability in multi-turn scenarios. In other words, the target in the multi-turn scenario is different from it in the single-turn scenario. (2) We record the best facet dimension results of the multi-dimensional facets generated by MulClari-Rule and MulClari-LLM as MulClari-Rule-Best and MulClari-LLM-Best respectively. It can be seen that compared to the optimal clarification pane selected by the model for delivering to the users (MulClari-Rule and MulClari-LLM in Table 2), the best clarification panes (facets) show great improvement in various metrics. This indicates that there are also many facets

⁴GPT-3.5-turbo: https://platform.openai.com/playground?model=text-davinci-003

WWW '24, May 13-17, 2024, Singapore, Singapore



Figure 5: Performance comparison with the baselines for different numbers of clarification turns.

Table 3: Our additional experimental results for MulClari-LLM. "w/o." in the table means "without".

Model	MRR	P@1	nDCG@1	nDCG@5	nDCG@20
LLM	0.2374	0.1389	0.1263	0.1239	0.1167
LLM-com w/o. E w/o. D	0.2313 0.2062 0.2136	0.1347 0.1263 0.1127	0.1246 0.0992 0.1015	0.1277 0.1016 0.1003	0.1118 0.0932 0.0951

corresponding to the ground truth in the MIMICS dataset that are included in the clarification candidates generated by our method. However, these facets were not selected to deliver to the user in the optimal clarification selection step.

4.6 Additional Experiments for MulClari-LLM

The above experiments have illustrated the effectiveness of our proposed LLM-based methods and MulClari-LLM. However, some details still have not been discussed. In this section, we conduct some additional experiments to explore some details contained in MulClari-LLM. First, the clarification candidate generation and optimal clarification selection are two separate processes. In fact, in MulClari-LLM, we can combine the two processes as one whole process by modifying the prompt and letting the LLM output the best clarification pane without generating multi-dimensional candidates. The result is noted as "LLM-com" in Table 3. It is found that all metrics show a slight decrease. It proves that generating clarification candidates first is important and effective. Besides, we are also interested in the effectiveness of the demonstrations and the retrieved documents. Therefore, we remove these two modules respectively and report their results in Table 3 as "w/o. E" and "w/o. D" respectively. We see that, after removing the demonstrations, the performance shows a significant decrease, confirming that the demonstrations are important for LLMs to complete multi-turn clarification tasks. However, the top retrieved documents D are not that important for LLMs, which just show a slight decrease.

4.7 Experiments for Clarification Turns

The performance of document retrieval is related to the specificity of the query, while the specificity of the query is related to the clarification turn, as shown in Figure 1. Therefore, the clarification turn affects the retrieval performance. Figure 5 shows the retrieval performance of our proposed methods as well as the baselines in different clarification turn $t \in \{1, 2, 3\}$. It is obvious that almost all metrics increase with the increase of t. However, when the turn t increases from 1 to 2, the increment is more significant than the increment when the turn t increases from 2 to 3. This indicates that the previous turns (like k = 1 or k = 2) are more meaningful for clarifying the user's intent. Similarly, some methods also show a decrease in some evaluation metrics when k = 3. This proves that the clarification effect does not necessarily increase with the number of turns. Some irrelevant documents can be wrongly retrieved when the length of the user query is long. This inspires us that, in the future, it is helpful to study how to automatically determine when to stop clarification and only return documents in multi-turn conversational Web search.

5 CONCLUSION

Multi-turn clarification has been applied in various kinds of conversational search systems. However, multi-turn Web search clarification is still not comprehensively studied. In this paper, we try to extend the framework, process, and concepts of existing multiturn clarification systems to the Web search for clarifying users' ambiguous or faceted search intents actively. We first define three important components of multi-turn Web search clarification including clarification candidate generation, optimal clarification selection, and document retrieval. Based on the framework, we design a rule-based method MulClari-Rule to generate clarification candidates and select the optimal clarification based on the frequency information of the facets, and then design an LLM-based method MulClari-LLM by utilizing the in-context learning and zero-shot instruction ability of LLMs, which further improves the effectiveness of multi-turn Web search clarification. The evaluation results on the Qulac and MIMICS datasets show that, first, our proposed methods achieve better performance in improving the document retrieval ability compared with existing single-turn clarification generation methods. Second, our proposed methods can also ensure the quality of generated clarification panes. We conduct some additional experiments to further illustrate our conclusions.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author. This work was supported by the National Natural Science Foundation of China No. 62272467, the fund for building world-class universities (disciplines) of Renmin University of China, Beijing Outstanding Young Scientist Program No. BJJWZYJH012019100020098, and Public Computing Cloud, Renmin University of China. The work was partially done at the Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE, and Beijing Key Laboratory of Big Data Management and Analysis Methods.

WWW '24, May 13-17, 2024, Singapore, Singapore

REFERENCES

- Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, et al. 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. 475–484.
- [2] Mark Chen, Jerry Tworek, Heewoo Jun, et al. 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021).
- [3] Qibin Chen, Junyang Lin, Yichang Zhang, et al. 2019. Towards Knowledge-Based Recommender Dialog System. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 1803–1813.
- [4] Jeffrey Dalton, Sophie Fischer, Paul Owoicho, et al. 2022. Conversational Information Seeking: Theory and Application. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 3455–3458.
- [5] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. 2021. Towards multimodal conversational information seeking. In Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval. 1577– 1587.
- [6] Yang Deng, Yaliang Li, Fei Sun, et al. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1431–1441.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 4171–4186.
- [8] Zhicheng Dou, Sha Hu, Yulong Luo, et al. 2011. Finding dimensions for queries. In Proceedings of the 20th ACM international conference on Information and knowledge management. 1311–1320.
- [9] Zhengxiao Du, Yujie Qian, Xiao Liu, et al. 2021. Glm: General language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360 (2021).
- [10] Helia Hashemi, Hamed Zamani, and W Bruce Croft. 2020. Guided Transformer: Leveraging Multiple External Sources for Representation Learning in Conversational Search. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1131–1140.
- [11] Helia Hashemi, Hamed Zamani, and W Bruce Croft. 2021. Learning multiple intent representations for search queries. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 669–679.
- [12] Helia Hashemi, Hamed Zamani, and W Bruce Croft. 2022. Stochastic Optimization of Text Set Generation for Learning Multiple Query Intent Representations. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 4003–4008.
- [13] Weize Kong and James Allan. 2013. Extracting query facets from search results. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 93–102.
- [14] Weize Kong and James Allan. 2014. Extending faceted search to the general web. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 839–848.
- [15] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, et al. 2020. Analysing the Effect of Clarifying Questions on Document Ranking in Conversational Search. In Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. 129–132.
- [16] Wenqiang Lei, Xiangnan He, Yisong Miao, et al. 2020. Estimation-actionreflection: Towards deep interaction between conversational and recommender systems. In Proceedings of the 13th International Conference on Web Search and Data Mining. 304–312.
- [17] Mike Lewis, Yinhan Liu, Naman Goyal, et al. 2020. BART: Denoising Sequenceto-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 7871–7880.
- [18] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, et al. 2018. Towards deep conversational recommendations. Advances in neural information processing systems 31 (2018).
- [19] Zujie Liang, Huang Hu, Can Xu, et al. 2021. Learning Neural Templates for Recommender Dialogue System. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 7821–7833.
- [20] Christopher D Manning, Mihai Surdeanu, John Bauer, et al. 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 55–60.
- [21] Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2023. A Comparative Study of Training Objectives for Clarification Facet Generation. In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. 1–10.

- [22] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35 (2022), 27730–27744.
- [23] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In Proceedings of the 2017 conference on conference human information interaction and retrieval. 117–126.
- [24] Sudha Rao and Hal Daumé III. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2737–2746.
- [25] Sudha Rao and Hal Daumé III. 2019. Answer-based Adversarial Training for Generating Clarification Questions. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 143–155.
- [26] Pengjie Ren, Zhongkun Liu, Xiaomeng Song, et al. 2021. Wizard of search engine: Access to information through conversations with search engines. In Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval. 533–543.
- [27] Chris Samarinas, Arkin Dharawat, and Hamed Zamani. 2022. Revisiting Open Domain Query Facet Extraction and Generation. In Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval. 43–50.
- [28] Julian Seitner, Christian Bizer, Kai Eckert, et al. 2016. A Large DataBase of Hypernymy Relations Extracted from the Web.. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 360–367.
- [29] Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to execute actions or ask clarification questions. In *Findings of the Association for Computational Linguistics: NAACL 2022.* 2060–2070.
- [30] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In The 41st international acm sigir conference on research & development in information retrieval. 235–244.
- [31] Leila Tavakoli, Johanne R Trippas, Hamed Zamani, et al. 2022. MIMICS-Duo: Offline & Online Evaluation of Search Clarification. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 3198–3208.
- [32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
- [33] Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).
 [34] Jan Trienes and Krisztian Balog. 2019. Identifying unclear questions in community
- [34] Jan Trienes and Krisztian Balog. 2019. Identifying unclear questions in community question answering websites. In *European Conference on Information Retrieval*. Springer, 276–289.
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017), 5998–6008.
- [36] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, et al. 2017. Exploring conversational search with humans, assistants, and wizards. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 2187–2193.
- [37] Jian Wang and Wenjie Li. 2021. Template-guided Clarifying Question Generation for Web Search Clarification. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 3468–3472.
- [38] Zhenduo Wang, Yuancheng Tu, Corby Rosset, et al. 2023. Zero-shot Clarifying Question Generation for Conversational Search. In Proceedings of the ACM Web Conference 2023. 3288–3298.
- [39] Zhongyuan Wang, Haixun Wang, Ji-Rong Wen, et al. 2015. An inference approach to basic level of categorization. In Proceedings of the 24th acm international on conference on information and knowledge management. 653–662.
- [40] Julia White, Gabriel Poesia, Robert Hawkins, et al. 2021. Open-domain clarification question generation without question examples. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 563–570.
- [41] Jingjing Xu, Yuechen Wang, Duyu Tang, et al. 2019. Asking clarification questions in knowledge-based question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 1618–1629.
- [42] Lili Yu, Howard Chen, Sida I Wang, et al. 2020. Interactive Classification by Asking Informative Questions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2664–2680.
- [43] Hamed Zamani, Susan Dumais, Nick Craswell, et al. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, 418–428.
- [44] Hamed Zamani, Gord Lueck, Everest Chen, et al. 2020. Mimics: A large-scale data collection for search clarification. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 3189–3196.
- [45] Hamed Zamani, Bhaskar Mitra, Everest Chen, et al. 2020. Analyzing and Learning from User Interactions for Search Clarification. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1181–1190.
- [46] Yiming Zhang, Lingfei Wu, Qi Shen, et al. 2022. Multiple Choice Questions based Multi-Interest Policy Learning for Conversational Recommendation. In

WWW '24, May 13-17, 2024, Singapore, Singapore

Ziliang Zhao and Zhicheng Dou

- Proceedings of the ACM Web Conference 2022. 2153–2162.
 [47] Zhiling Zhang and Kenny Zhu. 2021. Diverse and specific clarification question generation with keywords. In Proceedings of the Web Conference 2021. 3501–3511.
 [48] Ziliang Zhao, Zhicheng Dou, Jiaxin Mao, et al. 2022. Generating clarifying questions with web search results. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 234–244.
- [49] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, et al. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 1006-1014.
- [50] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, et al. 2020. Towards Topic-Guided Conversational Recommender System. In Proceedings of the 28th International Conference on Computational Linguistics. 4128-4139.