Generating Intent-aware Clarifying Questions in Conversational Information Retrieval Systems

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

Zhicheng Dou*
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
dou@ruc.edu.cn

Yujia Zhou
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
zhouyujia@ruc.edu.cn

Abstract

Generating clarifying questions can effectively clarify users' complicated search intent in conversational search systems. However, existing methods based on pre-defined templates are inadequate in understanding explicit user intents, making generated questions monotonous or inaccurate in some cases. In this paper, we define the "intent" of a query as a *verb* representing the potential behavior, action, or task the user may take. We study generating clarifying questions from a new perspective by incorporating the intents explicitly to form "intent-aware" questions with high informativeness and accuracy. Since obtaining gold intent-aware questions is expensive, we propose a rule-based method and a continual learning model to generate intent-aware questions as weak supervision signals. The former leverages search results to mine contextual intent-aware words or phrases, and the latter relies on parallel corpora to paraphrase template-based questions by incorporating the intents. The generated weak supervision data are then applied to fine-tune a BART-based model for end-to-end intent-aware question generation. We also explore to prompt a large language model to generate intent-aware questions. Experimental results on a public clarification dataset demonstrate that our proposed methods improve users' search experience compared to existing methods.

CCS Concepts

• **Information systems** → **Search interfaces**.

Keywords

Conversational Search, Search Clarification, Clarifying Question

ACM Reference Format:
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 (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627673.3679851

*Zhicheng Dou is the corresponding author.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679851

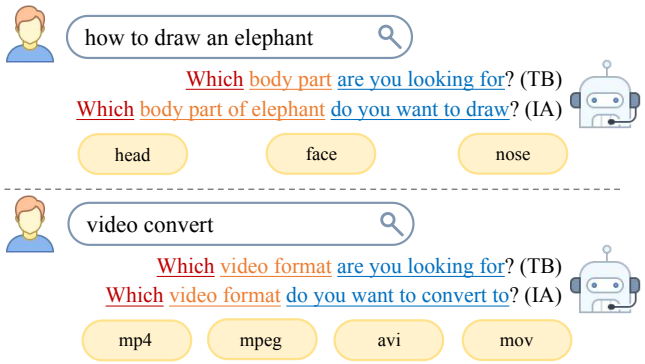


Figure 1: Examples of search clarification in conversational search systems. In this paper, we focus on improving the existing template-based (TB) clarifying question generation by defining and generating intent-aware (IA) questions.

1 Introduction

Search clarification has become an important module of conversational search systems and commercial search engines [29, 42]. It asks the user clarifying questions for their ambiguous or faceted intents [11, 21]. Figure 1 shows the basic clarification process [49]. For example, a user first submits a query "how to draw elephant" to the system. Since people may want to draw different body parts of the elephant, which means the query is faceted, the system provides several common body parts as clickable aspects (facets) and asks a question to help the user re-articulate her ambiguous or broad interest in detail. Among them, the facets usually represent a group of highly related sub-topics of the query, and the question is generated by the system for mixed-initiative interaction to improve users' search experience. When a user selects one facet, the query will be updated to retrieve new search results, and the clarification can be continued until the query clearly articulates the intent.

In this process, the quality of the *clarifying question* is important to give users a sense of intelligence to understand users' intent [1, 2, 12]. Studies also showed that a high-quality question can significantly improve the user engagement [49, 51]. At present, due to the complexity of Web search queries, clarifying questions are usually generated using *pre-defined templates* shown in Table 1. Usually, the key to generating a question is to first generate a *description* for the query (QUERY_DESC in Table 1) or the facets (FACETS_DESC in Table 1). For example, "medical condition" is a description for the query "headaches", and "type of headaches"

Table 1: Frequently used clarifying question templates.

N.	Template
T1	What do you want to know about **QUERY**?
T2	What do you want to know about this **QUERY_DESC**?
T3	Which **FACETS_DESC** are you looking for?
T4	Who are you shopping for?
T5	What do you want to do with **QUERY**?
T6	Which **QUERY** do you mean?

well describes a group of facets "[migraine, tension, cluster, hormone]". The description is then combined with a template as a question shown in Table 1 by rules [49, 57] or models [43, 56]. Although template-based methods are effective, they still have two considerable defects. **First**, template-based questions are all stylistically similar and diversity-lacking due to the limited template type, which lacks informativeness and reduces the user experience [51], especially for the template T1 in Table 1. **Second**, the process of finding descriptions loses information about *user intent*, which makes the questions *inaccurate* in some cases. For example in Figure 1, the facets of the query "video convert" are video formats to be converted. The template-based question "Which video format are you looking for?" is ambiguous because it is unclear whether the user is looking for "the format to be converted from" or "the format to be converted to". The above two limitations reduce the clarification quality and user's search experience.

In this paper, we focus on improving the quality of clarifying questions from a new perspective by making them *intent-aware*. First, to investigate the difference between template-based questions and ground-truth human-written questions, and explore what kind of question can provide a better user experience, we sample hundreds of questions from MIMICS dataset [39, 50] with different question templates and analyze their composition and features. We then hire five annotators who are familiar with the clarification scenario enough to manually write corresponding clarifying questions and compare them with the template-generated questions by performing the part-of-speech and frequency analysis. The results show that, for a large number of queries (about 42%), the five annotators believe that they can write a better clarifying question by modifying the question template with their subjective *intent information* in the form of *verb*. For example, as shown in Figure 1, when the query is "video convert", the *template-based* question is "Which video format are you looking for?", while the human-written question is "Which video format do you want to convert to?". Compared with the former question, the latter solves the ambiguity by explicitly emphasizing the user intent "convert to", which is action-aware and improves accuracy and informativeness.

Based on the analysis, it is demonstrated that **integrating user intents into template-based questions helps users accurately understand questions and improves users' experience**. Therefore, we propose *IQG*, an intent-aware question generation framework, to generate **intent-aware** clarifying questions as a new-perspective supplement and improvement of existing template-based questions. Since the analysis results significantly emphasize the importance of **verbs**, we define the "intent" of a query as a verb (or verb phrase) representing the potential behavior, action, or task the user may take. For example in Figure 1, the intent of the query "how to draw an elephant" might be "draw", and the intent of the query "video convert" should be "convert from" or "convert into".

After that, we study how to generate intent-aware clarifying questions. Since human-written questions are expensive to obtain, we first design a rule-based method and a continual learning model to generate a large amount of synthetic questions as **weak supervision signals**. Since rule-based methods are deemed suitable for generating clarifying questions [49, 57], we first design a rule-based method as a simple yet effective first-step attempt. The method extracts potential intent verbs from search results which contain abundant contextual information about the query, and then enhances existing question templates by incorporating the extracted verbs. However, the rule-based method may be sparse [49, 57]. As a supplement, we use large-scale paraphrasing corpora to build a continual learning framework to solve the sparsity problem. The continual learning model aims to borrow the characteristics of large-scale parallel corpora to enhance human-labeled data and to tackle the sparsity of the rule-based method. Finally, we fine-tune a pre-trained BART model using the generated data for an end-to-end intent-aware clarifying question generation. It is worth noting that *we do not completely dismiss template-based questions*. Since just a part of the queries are suitable for intent-aware questions, we also train the model with negative samples of template-based questions so that the model can determine when to generate a template-based or intent-aware question for a specific query automatically.

In existing studies of natural language processing (NLP), large language models (LLMs) are deemed good at capturing user intents and generating fluent natural questions [5, 10, 16, 17] compared with pre-trained language models (PLMs). In this paper, we also attempt to apply in-context learning to prompt LLMs to generate clarifying questions. We design the corresponding prompt and compare the effects of different numbers of demonstrations and whether to let the LLM directly generate or paraphrase questions.

In our experiments, we sample a subset of MIMICS [50], a large-scale Web search clarification dataset, as the evaluation data to evaluate the generated clarifying questions. Specifically, we first evaluate the questions by several sentence-level metrics including BLEU, ROUGE, and DISTINCT, then further evaluate the questions manually. Comprehensively, the experimental results show that our proposed intent-aware question generation models can generate more high-quality and user-satisfying questions versus strong baselines. With our proposed perspectives and methods, the existing search clarification scenario can be enhanced to provide a better user experience and user satisfaction in real-world conversational search systems. On the other hand, the LLM-based generation method shows low efficiency, with *9.25 times* inference time on average compared with the weak supervision method.

To sum up, the contributions of this paper include:

(1) We reveal the limitations of template-based questions and propose the IQG framework for generating intent-aware questions, which provides a new perspective for search clarification.

(2) We point out the importance of incorporating user intents in the form of verbs into clarifying questions by a carefully designed user study and data analysis.

(3) We implement a weak supervision method and an LLM-based method for intent-aware question generation. Experimental results demonstrate their effectiveness.

2 Related Work

2.1 Asking Open-domain Clarifying Questions

For conversational search systems, open-domain search clarification was first proposed by Aliannejadi et al. [1, 2]. In this scenario, a user retrieves web pages by conversation with the system. They built the Qulac dataset by crowdsourcing and proposed three models to retrieve questions from the dataset. After that, they tried to improve the performance of question selection [11, 35] and then analyzed the usefulness of user's responses [12]. Compared with question selection, question generation is considered more suitable for search clarification in information retrieval (IR) systems, because queries submitted by users are diverse and complex [57]. For asking clarifying questions in IR systems, Zamani et al. [49] first defined the clarification in IR systems. When a user submits a query to an IR system, the system generates a question and several facets for clicking shown in Figure 1. They also proposed three template-based and machine learning algorithms to generate questions with knowledge bases. They then analyzed interactions between users and the IR system to improve clarification quality [51]. Zamani et al. also built a dataset MIMICS [39, 50] based on the query log of Bing search engine. Each piece of data in MIMICS is composed of a query, several facets, and a clarifying question. In this paper, we focus on improving the quality of the "*questions*" displayed to users by incorporating user intent information, instead of exploring a better method for generating candidate aspect facets of a query. Since the existing MIMICS data does not contain explicit user intent information, and there is no intent-specific question, we additionally hire five annotators to expand the original MIMICS dataset.

2.2 Asking Close-domain Clarifying Questions

Close-domain search clarification usually focuses on some special cases like community question answering websites (StackExchange) [4, 31, 32, 40], conversational recommender system (recommending commodities to users) [7, 13, 38, 52, 53], and some other scenarios like interactive classification and behaviour decision [36, 48, 55]. It is more convenient to generate clarifying questions in a close-domain setting compared with an open-domain setting because the facets to be clarified are usually limited in a close-domain setting. For example, in a conversational recommender system, the attributes of each product are enumerable, such as brand, size, and price. Although close-domain methods and models can be applied to generate questions for Web search queries in some cases, in open-domain search clarification, user queries and document contents are mostly unstructured complicated information. This prevents us from utilizing structured data in some close-domain settings to generate clarifying questions.

2.3 Language Modeling

Traditional NLP tasks often require large-scale training datasets to implement fully supervised training tasks. After that, pre-training and fine-tuning has become one of the important research fields of NLP. The earliest pre-training models focused on obtaining semantic word embeddings [3, 19, 20, 25]. Later, the advent of Transformer network [41] has promoted the adventure of pre-trained language models (PLMs), including BERT [8], GPT [27, 28] etc. These PLMs significantly improved the performance of major NLP tasks [14, 37]. Recently, LLMs have become the research hotspot of NLP. By relying on the zero-shot instruction and in-context learning ability of LLMs, the effectiveness of many NLP tasks can be improved significantly [5, 10, 16, 17]. Our task is challenging because of the lack of intent-aware clarifying question data. In existing frameworks, one common way is to label a small amount of data (or obtain some weakly supervised data) and fine-tune a PLM on these data for generating intent-aware clarifying questions. Another feasible approach is to borrow the ability of LLMs by human-designed prompts and demonstrations. We implement both of the two frameworks and compare their effectiveness with experiments.

3 Clarifying Question Analysis

In this section, we utilize human-labeled clarifying questions and some automatic tools to conduct a comprehensive analysis. We first analyze the proportion of templates used in questions in an existing clarification dataset (MIMICS), and then analyze the manually-written ground-truth questions and template-based questions in terms of rewriting rate, part-of-speech, word frequency, etc. to show their differences. We finally conclude that some queries can use verbs to represent the user's search intent more specifically.

Existing clarifying question templates are shown in Table 1. Among them, template T6, T5, and T4 are three **special cases**, while template T3, T2, and T1 are suitable for **more common cases**: according to existing methods [49], when a description for facets can be found, the system will use **T3** first. Otherwise, if a description for the query can be found, **T2** will be preferred. If neither description for query nor facets can be found, template **T1** will be used to generate *generic question* and this will lead to lower user satisfaction [51]. The proportion of each question template calculated by the existing method [49] is shown in Figure 2(a). It is worth noting that using which template depends on both the query and its corresponding facets simultaneously.

To tackle the limitations of template-based questions mentioned in Section 1, it is important to know the difference between template-based and human-written questions. Therefore, we first randomly sample two hundred of "(query, aspect facets, question)" triples with T1, T2, and T3[1] as the question template respectively (600 in total) from MIMICS dataset [50]. We then hire five annotators (three males and two females) to manually rewrite a clarifying question given each query and corresponding facets. The annotators all have Ph.D. or Master's degrees with different majors and are well-trained in the whole process of Web search clarification by going through dozens of examples and experiencing an online clarification system. Each annotator is assigned 40 questions with the template T1, T2, and T3 respectively (120 in total). An annotator should first determine whether a template-based question should be rewritten. If she deems that the template-based question is not satisfying enough, she writes down a new question. After that, we do frequency and part-of-speech analysis for the results.

We first analyze the proportion of different question templates being rewritten by the five annotators. The rewriting rate of each

[1]Since special cases T4, T5, and T6 are few in the dataset, we only focus on studying T1, T2, and T3, and we believe that it does not influence the main conclusion.

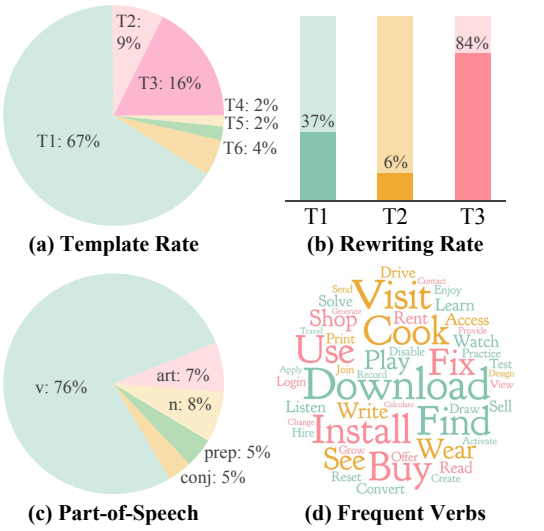


Figure 2: The analytical results of template-based and human-written clarifying questions.

template is summarized in Figure 2(b). It can be found that questions using template T1 and T3 are more prone to be rewritten, with the rate of 37% and 84% respectively, while T2 is few (6%). We believe that it is because templates T1 and T3 (especially T3) correspond to more *transactional queries* containing much explicit user intent, while most T2-based questions correspond to *informational queries*. For example, the transactional query "gta game" (and its corresponding facets) aims to *download* the game to an operating system. On the contrary, the informational query "headaches" aims to "look for" information about the condition. Based on the observations, we estimate that **for about 41% of queries, existing template-based questions can be rewritten**, which means that existing clarifying questions can still be improved significantly.

We then analyze the part-of-speech of the words in the human-written questions yet not in the original template-based questions. Results as shown in Figure 2(c). We find that **compared with template-based questions, most of the parts-of-speech added to human-written questions are verbs**. In other words, people are prone to use verbs to articulate their search tasks explicitly for some queries. The most popular verbs (including "download", "buy", "visit", etc.) are shown in Figure 2(d). These verbs explain the user's potential actions or tasks when issuing a query. For example, when a user searches for "google chrome exe", her task-aware intent may be "*install* google chrome exe". When a user searches for "gta game", her intent may be "*download* gta" or "*play* gta".

To sum up, the analytical results show that there is still a huge gap between template-based and human-written questions for a large number of queries, especially for transactional queries. To mitigate this gap, **it is essential to explicitly incorporate user intents into the template-based questions in the form of verbs**. To achieve this, there are two important problems to be solved: (1) without large-scale intent-aware data, how to obtain its potential search intent for a specific query, and (2) how to integrate the original query to generate intent-aware clarifying questions.

4 Intent-aware Question Generation

Our analysis emphasizes the importance of intent-aware clarifying questions. In this section, we introduce the intent-aware question generation (IQG) framework. First, as shown in Figure 3, since gold intent-aware data are expensive to obtain, we define a rule-based method together with a continual learning method to collect weak supervision data. The rule-based method extracts intents from search results of the user query and then integrates the intents into several *enhanced templates* to form intent-aware questions. The continual learning framework relies on large-scale parallel corpora and a small amount of human-labeled data to generate intent-aware questions. We implement these two approaches to generate weak supervision data and then use the generated data to fine-tune a BART model for end-to-end intent-aware question generation. We name the method IQG-WS (weak supervision). We further design another method IQG-LLM to generate intent-aware questions by prompting the large language model.

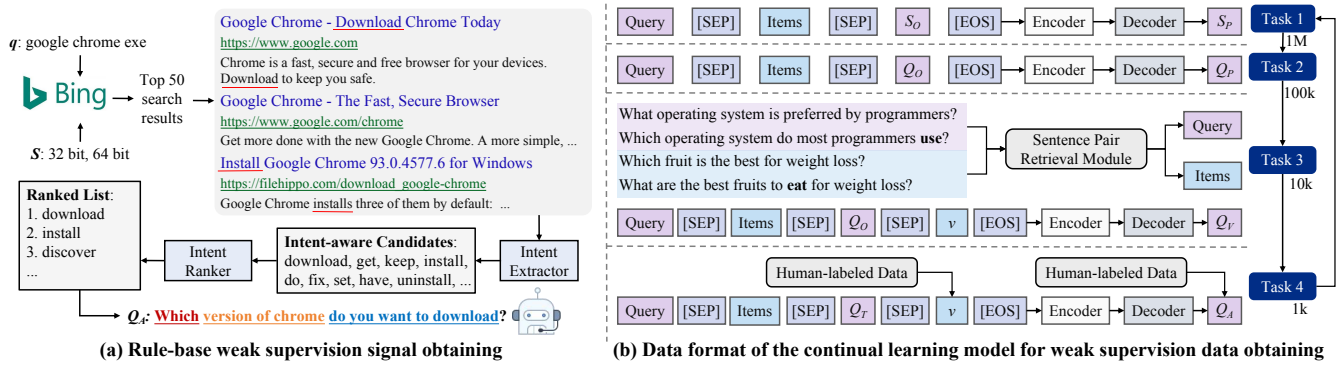
4.1 Rule-based Method

Our rule-based method is designed to generate intent-aware questions by human-designed rules. Studies have shown that search results contain abundant contextual information of a specific query [9, 57]. Based on this, We deem that user intents in the form of verbs can also be extracted from the search results of the query. For example, for the query "google chrome exe", the search results usually contain many intent-specific verbs like "download", "install", and "fix", as shown in Figure 3. To achieve this, we concatenate the query q with each facet in the facets set S into a combined query submitted to the search engine. For example, for the query "google chrome exe" and facets [32 bit, 64 bit], the submitted query is "google chrome exe 32 bit 64 bit". We do this to obtain co-occurrence information of q and S in the search results. We then get top-k search results and extract all plain texts in the HTML to get a text set T. By implementing Stanford Stanza [18, 26, 54], we can easily get the part-of-speech of each word in the text set and obtain a verb set $\mathcal{V}$. Each verb in the set is a candidate. For each candidate v. we define the feature score $f(v)$ as the **summation** of five sub-features: frequency feature $f_{\text{f}}(v)$, query distance feature $f_{\text{dq}}(v)$, facets distance feature $f_{\text{dS}}(v)$, query pattern feature $f_{pq}(v)$, and facets pattern feature $f_{\text{pS}}(v)$. Studies have shown that these features are effective in mining information from search result pages [57].

We rank all verb candidates based on their $f(v)$ and select the candidate with the maximum $f(v)$ as v_0. We also set a threshold τ, if $f(v_0) \geq \tau$, the algorithm returns v_0. Otherwise, it does not return a verb, which means that the current query q is unsuitable for generating an intent-aware question. Instead, it generates a template-based question using existing methods. The above five sub-features are defined as follows.

4.1.1 Frequency Feature. Verbs with higher frequency in texts are likely to have a strong correlation with the query and facets. Therefore, we count verb v in top result texts as the frequency feature:

$$f_{\text{f}}(v) = \tanh\left(k_{\text{f}} \sum_i N(T_i, v)\right), \tag{1}$$



(a) Rule-base weak supervision signal obtaining

(b) Data format of the continual learning model for weak supervision data obtaining

Figure 3: Two approaches to obtain weak-supervision signals: (a) A rule-based method extracting search intents in the form of verbs from search results by human-designed features and generating intent-aware clarifying questions by intent-enhanced question templates. (b) A continual learning framework borrowing parallel corpora to generate intent-aware clarifying questions with few human-labeled data.

where T_i is the i-th result and $N(T_i, v)$ is the frequency that verb v occurs in T_i. The hyper-parameter k_f is used to adjust the importance of each feature, and the tanh() function is used to scale the value of the feature, which is the same as below.

4.1.2 Query Distance Feature. Intuitively, a verb nearing the query q in a piece of text is more likely to have a stronger relationship with the query. Therefore, we set a distance function to calculate the distance between the verb v and the query q, then calculate the query distance feature f_{dq} based on:

$$f_{\text{dq}}(v) = \tanh\left(k_{\text{dq}} \sum_i d_q(v, q, T_i)\right). \tag{2}$$

Here $d_q()$ is defined as the reciprocal of the sum of distances:

$$d_q(v, q, T_i) = \sum_j \frac{1}{\text{dist}(v, q_{ij})}, \tag{3}$$

where q_{ij} is a query occurring in T_i. The distance function $\text{dist}(a, b)$ should consider the distance between two words a and b:

$$\text{dist}(a, b) = \begin{cases} |\text{p}(a) - \text{p}(b)| & \text{if } |\text{p}(a) - \text{p}(b)| \le 10; \\ +\infty & \text{for else.} \end{cases} \tag{4}$$

Here $\text{p}(x)$ returns the position of x in a paragraph. If the distance (number of words) between a and b is longer than ten words, we deem they no longer have a connection anymore.

4.1.3 Facets Distance Feature. Similar to the query distance feature, if a verb v is close to an aspect facet, we deem that v is also likely to express user intent. For example, in the top results of query "google chrome exe", except for "download google chrome", there are still some "download 32 bit" and "download 64 bit" in web pages. These verbs should also be assigned a higher score. We also integrate the entropy (ent) of facet scores to avoid attending partial facets and ensure that all facets should be considered. The facets distance feature is defined as follows:

$$f_{\text{dS}}(v) = \tanh\left(k_{\text{dS}} \sum_i d(v, S, T_i) \cdot \text{ent}(S)\right), \tag{5}$$

where $d()$ shares the same definition of $d_q()$ in Equation (3).

4.1.4 Query Pattern Feature. Similar to existing studies [9, 49], patterns are important to extract verbs representing intents from texts, like "**download** google chrome exe" and "**eat** the food". For a query q, if a verb v appears followed by q, then v is more likely to be a user intent for q. Therefore, we define a template "[V] ART QUERY" where "ART" is a definite & indefinite article like "the", "a", and "this", or a null character "". We then calculate f_{pq} as follows:

$$f_{\text{pq}}(v) = \tanh\left(k_{\text{pq}} \sum_i I(v, T_i) \cdot I(q, T_i)\right). \tag{6}$$

For the function $I(x, T)$, when x occurs in T, $I(x, T) = 1$, otherwise when x does not occurs in T, $I(x, T) = 0$.

4.1.5 Facets Pattern Feature. Similar to query pattern feature f_{pq}, we also define a template "[V] ART FACET" to obtain verbs occurring before facets. Staying consistent with $f_{\text{dS}}(v)$, we further add the entropy of facets as a factor to ensure that all facets are considered.

$$f_{\text{pS}}(v) = \tanh\left(k_{\text{pS}} \sum_i I(v, T_i) \cdot I(S, T_i) \cdot \text{ent}(S)\right). \tag{7}$$

Recent studies show that most questions can be formulated by a few templates [51]. Therefore, we try to extend existing templates to make them intent-aware. By observing the human-written questions used in Section 3, we summarize two types of intent: intent for query q, or intent for its corresponding facets S. The two types of intent can be formulated as two templates respectively:

- T1: Which FACETS_DESC are you looking for to [V] q?
- T2: Which FACETS_DESC do you want to [V]?

The first template T1 is used when the intent is focusing on the query q. For example the query "prom dresses" and its facets [long, short], the potential intent is that the user wants to *wear* a prom dress (query), thus the clarifying question could be "Which length are you looking for to wear prom dresses?". The second template T2 focuses on the intent for facets S. For example the query "google chrome exe" and its facets [32 bit, 64 bit], the intent is to download a specific version (facets) of chrome, thus the question should be "Which version do you want to download?". To determine whether

the intent is mainly focused on query q or facets S, for a verb v, we can simply compare the feature values of q and S mentioned in Section 4.1: f_{dq} and f_{pq} focus on the query, and f_{dS} and f_{pS} focus on the facets. Therefore, we set $f_{\text{q}}(v) = f_{\text{dq}}(v) + f_{\text{pq}}(v)$ and $f_{\text{S}}(v) = f_{\text{dS}}(v) + f_{\text{pS}}(v)$. If $f_{\text{q}}(v) \geq f_{\text{S}}(v)$, which means that feature values related to query q are larger than that related to facets S, we use T1 as the template to form a question. Otherwise, T2 will be used. In fact, template-based questions may be prone to grammatical or semantic errors. However, they are still considered a simple yet effective solution in early exploration, especially when there is no available data for training a model.

4.2 Continual Learning Model

The rule-based method strictly relies on the two templates proposed in Section 4.1, while ignoring other possible question formats. This makes some intents difficult to be expressed by (intent-aware) question template accurately, and the generated questions are sparse, which means that it can only generate intent-aware questions for queries that contain explicit intent information in their search result pages. To this end, we can directly use human-labeled data for training to generate more questions. However, few amount of human-labeled data can easily lead to over-fitting [28, 30], while large-amount human-written data are expensive to obtain. To alleviate this problem, we propose a novel framework as shown in Figure 3(b). **First**, we deem that writing an intent-aware question given the template-based question is equal to a paraphrasing process [22, 47]. **Second**, many existing parallel (paraphrasing) corpora have similar features to our task, which can be utilized as weak supervision signals to enhance our small amount of human-labeled data. Based on the two observations, **we reform the intent-aware question generation as a paraphrasing task and propose enhancing human-written questions with paraphrasing data to generate intent-aware clarifying questions**. Since different kinds of paraphrasing data have different features, we further divide the data into four levels and propose a continual learning framework [37] to learn to paraphrase level-by-level.

The four-level data shown in Figure 3(b) include:

(1) 1M general paraphrasing **sentences** (S_O, S_P) for learning the common paraphrasing process.

(2) 100K parallel **questions** (Q_O, Q_P) focus on learning to paraphrase questions.

(3) 10K **parallel questions containing verb modification** (Q_O, Q_V). For example in Figure 3(b), the question "Which fruit is the best for weight loss" can be rewritten as "Which are the best fruits to *eat* for weight loss?" by adding a new verb "eat". These data help us learn to paraphrase by adding a verb.

(4) 1K **human-labeled question pairs** (Q_T, Q_A) which is our task goal. As the data volume decreases in turn, the task is getting closer to our goal.

To effectively combine these data and their corresponding training tasks, following previous studies, we design a continuous learning [23, 37] framework. As shown in Figure 3, we define an epoch as a learning process starting from task 1 and ending with task 4. We split the training data of each task into batches of size 1K (task 1 for 1K batches, task 2 for 100 batches, task 3 for 10 batches, and task 4 for only 1 batch). In each training epoch, we fetch a batch from each task respectively and form an identical batch permutation, resulting in 1M epochs in total. This process ensures that the training data for each epoch is different, and the model will not forget old tasks when trained on a new task.

In task 4, we concatenate the query q, the facets S, the template-based question Q_T, and the human-labeled intent verb v with a special token "[SEP]" as the input, and let the model output the intent-aware question Q_A. However, in common paraphrasing datasets like ParaNMT and Quora, the concept of the query and the facets is missing. As a result, the training data format between tasks 1-3 and task 4 is different, which affects the model training. To imitate the data format of task 4, we retrieve the most similar query and corresponding facets from the MIMICS dataset in tasks 1, 2, and 3. To this end, we first apply BERT [8] to encode the embedding of sentence pair $E_S = \text{BERT}(Q_O \text{ [SEP] } Q_V)$, then encode the embedding of the concatenated query and facets $E_D = \text{BERT}(q \text{ [SEP] } S)$, and finally calculate the cosine similarity between the embeddings of E_S and E_D and select the query and corresponding facets with the highest score. As for task 3, we further imitate task 4 by adding the verb explicitly in the input to stay consistent with task 4.

4.3 IQG-WS: Generation with Weak Supervision

To take advantage of the above two algorithms simultaneously, we sample 40K queries in the MIMICS dataset, and then run the rule-based method and continuous learning model to collect 20K pieces of data as weak supervision signals respectively, and finally randomly mix them for further fine-tuning. The input of each piece of data includes a query, its corresponding facets, and a template-based question Q_T. The output is the intent-aware question Q_A generated by our proposed rule-based method or continual learning model. In the 40K pieces of weak supervision data, about 17.8k pieces of data (occupying 45%) are rewritten. We implement IQG-WS based on the sequence-to-sequence framework BART [14] and train it with the weak supervision data.

4.4 IQG-LLM: LLM-based Generation

The emergence of LLM (such as ChatGPT) improves the performance of many NLP tasks [5, 17]. As one feasible solution, one can prompt an LLM to generate clarifying questions [21, 33]. In this paper, we design a natural language prompt for intent-aware question generation composed of the following components:

(1) **Task description**: introducing the search clarification process and the need for incorporating user intent information.

(2) **Demonstrations**: examples of (q, S, D, Q_T, Q_A) in front of the generation instruction, where q is the query, S indicates the aspect facets, D is the top search result snippets, Q_T is the template-based question, and Q_A is the target intent-aware question. To enable the model to determine whether to generate an intent-aware question or a template-based question automatically, we further display several negative examples of template-based questions that are not necessary to be paraphrased.

(3) **Generation instruction**: given a new piece of test data (q, S, D, Q_T), let the LLM output the predicted intent-aware question (Q_A). The question can be consistent with the template-based question Q_T or it can be paraphrased by incorporating an explicit user search intent. In this paper, we use the OpenAI GPT 3.5-Turbo

for generation[2], which can be directly replaced by other LLMs. We further study the impact of different numbers of demonstrations k and whether to generate or paraphrase on the generation ability of the LLM-based method. For related experiments, see Section 5.3.

5 Experiments

5.1 Settings

5.1.1 Data. We select 1,000 samples from the MIMICS dataset [39, 50] as the training data for task 4 in Section 4.2. Among them, half are positive samples with human-labeled intent-aware questions. The other half are negative samples that are deemed not necessary to be rewritten, ensuring that the model does not generate an intent-aware question when it is not necessary. We further sample another 200 pieces of evaluation samples from MIMICS. The training and evaluation data are manually labeled with intent-aware or template-based questions according to the process in Section 3. For generating weak supervision data, we sample 20,000 queries from MIMICS for the rule-based method and another 20,000 queries for the continual learning model. We ensure that the selected MIMICS data contains 50% ambiguous queries and 50% faceted queries. We select ParaNMT-50M [45] and Quora[3] as two parallel corpora used in the continual learning model.

5.1.2 Baselines. We implement **RTC** and **QLM** [49] as two baselines to generate clarifying questions. RTC first finds descriptions for queries or facets, then combines these descriptions with predefined templates shown in Table 1 to generate a question. QLM is a seq2seq model that is trained with questions generated by RTC as weak-supervision data for generalization. To implement RTC, we apply WebIsA [34] and Concept Graph [44, 46] dataset to find descriptions for queries and facets. Since the two algorithms are not intent-aware, We further design **RTC-I** and **QLM-I**. RTC-I uses our proposed rule-based method to improve the existing template with intent information. QLM-I uses the data generated by RTC-I for weak supervision to improve generalization. We also fine-tune a BART model as a PLM-based baseline which is directly fine-tuned with our human-labeled data in an end-to-end way since it is commonly used in recent clarifying question generation studies [43, 56].

5.1.3 Evaluation. The best evaluation approach for clarification is to measure the downstream retrieval capability [2]. However, it is very difficult because the evaluation of downstream retrieval capability relies on a question pool, yet our questions are generated and could not be in the pool [35]. According to existing research, when users are more satisfied with the generated clarifying questions, the questions are often more likely to lead to better retrieval results [51, 57]. Therefore, following these works, we focus on evaluating the quality of the questions.

To evaluate the generated results compared with the human-written ground truths, we apply BLEU and ROUGE as two automatic evaluation metrics. We also apply DISTINCT [15] as an additional metric to measure the diversity of generated questions since we deem that it is also important to solve the monotony problem of template-based questions [51]. Besides, since automatic metrics cannot completely measure the question quality [50], in our main experiments, we further apply **human annotation** to evaluate the questions. Following previous work [49, 57], we assign the labeling task to three annotators and obtain the labels based on **majority voting**. The hired annotators are all knowledgeable of the task and its applications. For each generated clarifying question, we ask the annotators to choose a score among 0, 1, 2, and 3, representing the accuracy and informativeness of a question. The scoring criteria and examples are shown in Table 2. We guarantee that, in fewer than 5% cases, there is no agreement among the annotators (i.e., no label with more than 1 voter). In such cases, the three annotators will hold a quick meeting to discuss the final labeling result. We further calculate a linear and an exponential score of human evaluation [49], denoted as "linear" and "exp" in Table 3 respectively.

5.1.4 Algorithm Implementation Details. In the rule-based method, we adjust the parameters k and τ by grid search with the step of 0.1, from the range of $(0, 1]$ and $(0, 5]$ respectively. All pre-trained language models are fine-tuned based on BART-base [14]. The batch size is set to 32 and the max lengths of the input and output are set to 64. We use AdamW with the learning rate of 2×10^{-5} to optimize the cross-entropy loss function. We apply Scikit-learn [6, 24] to split the training set and validation set with the rate of 9:1 when training all mentioned models, and apply an early stop strategy to control the training process until the loss does not decrease on the validation set when training each task in each epoch. In the IQG-LLM method, the number of demonstrations defaults to 5.

5.2 Overall Results

Since we have got human-labeled ground-truth data, we first evaluate whether our generated intent-aware questions are similar to the human-written questions. Therefore, we first apply BLEU-1, 2, ROUGE-1, 2, L, five commonly used metrics to measure the similarity between sentences. We also apply the DISTINCT [15] (Dist-1, 2) to evaluate the n-gram diversity since it is also important for clarifying questions based on our analysis in Section 3.

The results are shown in Table 3. For all metrics, **our proposed weak supervised method (IQG-WS) outperforms the four baselines significantly**. Specifically, by incorporating the intent into two baselines RTC and QLM, the improvement in BLEU and ROUGE metrics nearly doubled. This proves that our proposed intent-aware questions are more similar to human-written questions compared with template-based questions, which supports our motivation. Our proposed model further outperforms QLM-I and achieves state-of-the-art results, which demonstrates the effectiveness of the weak supervision data generated by the rule-based method and continual learning model. As for Dist-1 and Dist-2, RTC-I and QLM-I both show a slight increase, while our method increases significantly to 0.335 and 0.514 respectively. This is because although RTC-I and QLM-I both incorporate intent information, the question format is still restricted by the templates, yet our proposed method can generate more flexible formats. As for the LLM-based method (IQG-LLM), it achieves slightly lower BLEU and ROUGE yet higher DISTINCT score compared with IQG-WS. This shows that, first, few-shot prompted LLM is competitive in generating intent-aware questions compared with our proposed weak supervised method IQG-WS. Second, the questions generated by the LLM are more flexible in format, thereby improving the DISTINCT score.

[2]https://platform.openai.com/playground?model=text-davinci-003
[3]https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

Table 2: Criteria of scoring and evaluating clarifying questions by human.

Rating	Criteria
0	Wrong or illogical question. The question has semantic error or obvious grammatical error, or is difficult to understand.
1	Generic clarifying question like "What do you want to know about QUERY?" and "Select one to refine your search."
2	Correct clarifying question which is not very specific and need further thinking to understand , but still acceptable to users.
3	Correct and easy-to-understand clarifying question which can effectively articulate user intents and provide good experience.

Examples			
Query: centimeters , **Facets**: [inch, feet, meter, kilometer]		**Query**: google chrome exe, **Facets**: [32 bit, 64 bit]	
0	What do you want to *measure* a distance?	0	Which **version** are you looking for to **release** chrome?
1	What do you want to know about **centimeters**?	1	Select one to refine your search.
2	Which **unit** are you looking for?	2	Which **version** are you looking for?
3	Which **length unit** do you want to **convert** into?	3	Which **version of chrome** do you want to **download**?

Table 3: Experimental results of clarifying question generation on MIMICS dataset. "†" denotes that the result is significantly outperformed by our proposed method with $p < 0.05$. The best results are in bold and the second best results are underlined.

Method	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	ROUGE-L	Dist-1	Dist-2	3	2	1	0	linear	exp
RTC	0.394†	0.223†	0.434†	0.164†	0.430†	0.110†	0.190†	0.310	0.115	**0.445**	0.125	1.605†	2.960†
QLM	0.402†	0.235†	0.446†	0.173†	0.434†	0.091†	0.177†	0.270	0.145	0.415	0.120	1.515†	2.740†
RTC-I	0.603†	0.438†	0.629†	0.338†	0.620†	0.125†	0.237†	0.515	0.140	0.110	**0.235**	1.935†	4.135†
QLM-I	0.675†	0.515†	0.716†	0.432†	0.688†	0.138†	0.274†	0.590	0.155	0.085	0.170	2.165†	4.680†
BART	0.668†	0.497†	0.703†	0.447†	0.692†	0.165†	0.298†	0.575	0.135	0.120	0.170	2.115	4.550
IQG-WS	**0.752**	**0.549**	**0.773**	**0.515**	**0.768**	0.335	0.514	**0.645**	0.110	0.120	0.125	**2.275**	**4.965**
IQG-LLM	0.742	0.546	0.752	0.485	0.755	**0.402**	**0.597**	0.575	**0.185**	0.165	0.075	2.260	4.745

Besides automatic metrics, we also carry out human evaluation according to the annotation criteria shown in Table 2. The evaluation results are shown on the right side of Table 3. We first list the proportion of questions scored by 3, 2, 1, and 0 respectively, then calculate a linear score and an exponential score [49, 57] to represent the overall question quality. The experimental results show that, first, RTC-I and QLM-I both improve the human annotation results compared with RTC and QLM, with the linear score increasing from 1.605 to 1.935 and 1.515 to 2.165 respectively. Besides, the improvement of QLM-I over RTC-I is significantly higher than that of QLM over RTC. This is because the data for training QLM are not intent-aware, while QLM-I is trained with intent-aware questions, which improves the generalization of RTC-I. Our weak supervised model (IQG) also outperforms all baselines significantly with the linear score of 2.275 and the exponential score of 4.965. It is also worth mentioning that, as for the LLM-based model (IQG-LLM), it gets more 1 and 2 points yet less 0 and 3 points, and achieves the linear score of 2.260 and the exponential score of 4.745.

5.3 Ablation Studies

To prove the effectiveness of the modules in our proposed methods, we further conduct three ablation studies (AS): **(AS1)** whether the weak supervision data generated by our proposed rule-based method and continual learning model both effective respectively and which part of data contributes more to the result? **(AS2)** In the continual learning model, whether the four tasks all contribute to the result and whether paraphrasing parallel corpora is effective in enhancing human-written questions? **(AS3)** Are the demonstrations and various components in the LLM-based method effective?

The results are shown in Table 3. In AS1, we use the rule-based method and continual learning (CL) model independently to generate 40K pieces of weak supervision data respectively, denoted as "w/o. CL" and "w/o. Rule", to ensure the fairness of the data volume. Table 3 shows that, no matter which of the two kinds of data is removed, almost all the automatic and manual evaluation metrics decrease, illustrating that the two methods for obtaining weak supervision data both have a positive effect on the results. The data generated by the two can complement each other, thus improving the generation quality. On the other hand, the results without the rule-based method are slightly better than the results without the CL model. Furthermore, by removing the CL model, the Dist-1 and Dist-2 metrics show a tremendous decrease, from 0.335 to 0.265 and 0.514 to 0.356 respectively. However, by removing the rule-based method, Dist-1 and Dist-2 just show a slight increase. This is because questions generated by the CL model are more flexible in format because of the multi-level training data borrowed from general parallel corpora.

In AS2, we focus on the continual learning model and try to prove the effectiveness of the four tasks to the results. The results show that, by removing tasks 1 to 3 individually, the quality of generated questions decreases gradually. By removing task 4, the results showed a significant decline, which is even lower than the lowest baseline. This emphasizes the importance of human-labeled data. We also try to only use task 4 to train the model without the continual learning framework. The significantly decreased results also show that tasks 1 to 3 can enhance task 4. This proves the validity of our proposed continual learning framework.

In AS3, we study how the demonstration number k, the search result snippets D, and how the template-based question Q_T affects

Table 4: Ablation study results on the MIMICS dataset.

Method	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	ROUGE-L
IQG-WS	0.752	0.549	0.773	0.515	0.768
w/o. Rule	0.726	0.531	0.755	0.480	0.735
w/o. CL	0.703	0.522	0.743	0.462	0.727
w/o. task1	0.713	0.529	0.756	0.501	0.739
w/o. task2	0.708	0.536	0.756	0.462	0.718
w/o. task3	0.670	0.498	0.725	0.429	0.692
w/o. task4	0.241	0.199	0.210	0.085	0.177
task4 only	0.655	0.484	0.713	0.412	0.686
zero-shot	0.374	0.198	0.385	0.210	0.356
$k = 1$	0.629	0.396	0.680	0.395	0.684
$k = 3$	0.684	0.501	0.713	0.432	0.708
IQG-LLM	0.742	0.546	0.752	0.485	0.755
No Para	0.717	0.508	0.722	0.451	0.696

the generation ability of the LLM. The ablation results are shown in Table 4. The results show that, first, with the increase of the demonstration number k from 0 to 5, the generation quality improves. Specifically, the LLM performs badly in a zero-shot setting. Besides, according to the result of "No Para" in Table 4, the paraphrasing paradigm performs better slightly than the generation paradigm in all five metrics. This is because the template-based questions provide additional information and question format.

5.4 Case Study

To compare the generated results intuitively, we select four queries and corresponding facets and use our proposed approaches to generate corresponding clarifying questions shown in Table 5. In this table, Q_T is the template-based question, and "-R" and "-C" means that the rule-based method and continual learning model are being removed respectively. For the first three queries, our BART model fine-tuned with two sets of weak supervision data generates high-quality questions which clearly clarify the user's intent, while for the last query "facial tingling", there is a logical error. This shows that although we have achieved significant improvement in automatic and manual evaluation, our proposed model still has errors in understanding some complex intents. It also can be seen that, after removing the rule-based method, the generated questions become more flexible. For the query "vob player", the "-R" method generates a question starting with "on which" that is different from others. However, it is prone to misunderstand the intent because of the lack of contextual information: for example, the query "nsw tafe portal" adds the wrong verb "serve" as the intent. On the other hand, by removing the continual learning model ("-C"), the generated questions all stay the same format with our designed two templates, but it would also lead to logical errors. For the second query, it generates "Which people do you want to login?" as the question, which lacks a preposition "as". These cases further illustrate that the two weak supervision methods we proposed can complement each other, thus improving the quality of the results. Finally, we can conclude that template-based questions show too many informational characteristics, but the intent of many transactional queries cannot be grasped, resulting in inaccurate generated questions.

Table 5: Four example outputs of different algorithms.

q	vob player	S	[windows, mac, android]
Q_T	Which OS are you looking for?		
Our	Which OS do you want to download the vob player on?		
-R	On which OS do you want to download vob player?		
-C	Which OS are you looking for to download vob player?		
q	nsw tafe portal	S	[student, staff]
Q_T	Which people are you looking for?		
Our	Which people do you want to login as?		
-R	Which group of people do you want to serve as?		
-C	Which people do you want to login?		
q	colts reddit	S	[seahawk, steeler, texan]
Q_T	Which team are you looking for?		
Our	Which team in colts reddit do you want to see?		
-R	What colts reddit team do you want to watch?		
-C	Which team are you looking for to beat colts reddit?		
q	facial tingling	S	[left side, right side]
Q_T	Which body are you looking for?		
Our	Which body are you looking for to tingle?		
-R	Which part of body do you want to tingle?		
-C	Which body do you want to tingle?		

6 Limitation

In this section, we list limitations of this paper. First, the advantages of intent-aware question is proven by our small-scale user study, which is insufficient. One possible solution is to utilize online search engines for online experiment, while it is usually difficult to access main-stream commercial search engines. Second, we use verbs to represent the user intent. However, it is not optimal because we do not actually know the user's real intent. In this paper, the intent is more like action, task, or transaction. The intent information may can be mined from other resources, like search log.

7 Conclusion

We study generating clarifying questions from a new perspective by incorporating user intent into questions to improve user experience. We first conduct user studies proving that intent-aware questions can improve user satisfaction for a large number of queries compared with template-based questions. We then design a rule-based method to generate intent-aware questions with search results, and a continual learning framework to generate questions leveraging parallel corpora. The two methods generate weak supervision data which are then applied to fine-tune a generative model for end-to-end generation. We also try to generate intent-aware questions by prompting large language models. The experimental results on two datasets demonstrate the effectiveness of our proposed motivation and methods. Therefore, they can be implemented in real-world conversational search systems to provide a better user experience.

Acknowledgments

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, 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.

References

[1] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, et al. 2021. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 4473–4484.

[2] 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.

[3] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems* 13 (2000).

[4] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, et al. 2017. What do you mean exactly? Analyzing clarification questions in CQA. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 345–348.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Lars Buitinck, Gilles Louppe, Mathieu Blondel, et al. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[7] 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.

[8] 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.

[9] Zhicheng Dou, Sha Hu, Kun Chen, et al. 2011. Multi-dimensional search result diversification. In *Proceedings of the fourth ACM international conference on Web search and data mining*. 475–484.

[10] Aysa Fan, Haoran Zhang, Luc Paquette, and Rui Zhang. 2023. Exploring the Potential of Large Language Models in Generating Code-Tracing Questions for Introductory Programming Courses. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 7406–7421.

[11] 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.

[12] 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.

[13] Wenqiang Lei, Xiangnan He, Yisong Miao, et al. 2020. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 304–312.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, et al. 2020. BART: Denoising Sequence-to-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.

[15] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 110–119.

[16] Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Weining Qian, and Yunshi Lan. 2023. Prompting Large Language Models with Chain-of-Thought for Few-Shot Knowledge Base Question Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 4329–4343.

[17] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[18] 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*. 55–60.

[19] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems* 30 (2017).

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[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] Tong Niu, Semih Yavuz, Yingbo Zhou, et al. 2021. Unsupervised Paraphrasing with Pretrained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 5136–5150.

[23] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks* 113 (2019), 54–71.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[26] Peng Qi, Yuhao Zhang, Yuhui Zhang, et al. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 101–108.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, et al. 2018. Improving language understanding by generative pre-training. (2018).

[28] Alec Radford, Jeffrey Wu, Rewon Child, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[29] 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.

[30] Colin Raffel, Noam Shazeer, Adam Roberts, et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2021. Towards Facet-Driven Generation of Clarifying Questions for Conversational Search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 167–175.

[36] 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.

[37] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8968–8975.

[38] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 235–244.

[39] 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.

[40] Jan Trienes and Krisztian Balog. 2019. Identifying unclear questions in community question answering websites. In *European Conference on Information Retrieval*. Springer, 276–289.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[42] 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.

[43] 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.

[44] 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.

[45] John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 451–462.

[46] Wentao Wu, Hongsong Li, Haixun Wang, et al. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 481–492.

[47] Qian Yang, Zhouyuan Huo, Dinghan Shen, et al. 2019. An end-to-end generative architecture for paraphrase generation. 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)*. 3132–3142.

[48] 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.

[49] Hamed Zamani, Susan Dumais, Nick Craswell, et al. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The WebConf 2020*. 418–428.

[50] 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.

[51] 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.

[52] Yongfeng Zhang, Xu Chen, Qingyao Ai, et al. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 177–186.

[53] Yiming Zhang, Lingfei Wu, Qi Shen, et al. 2022. Multiple Choice Questions based Multi-Interest Policy Learning for Conversational Recommendation. In *Proceedings of the ACM Web Conference 2022*. 2153–2162.

[54] Yuhao Zhang, Yuhui Zhang, Peng Qi, et al. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association* 28, 9 (2021), 1892–1899.

[55] Zhiling Zhang and Kenny Zhu. 2021. Diverse and specific clarification question generation with keywords. In *Proceedings of the Web Conference 2021*. 3501–3511.

[56] Ziliang Zhao, Zhicheng Dou, Yu Guo, Zhao Cao, and Xiaohua Cheng. 2023. Improving Search Clarification with Structured Information Extracted from Search Results. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3549–3558.

[57] 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*. 385–394.