Towards Facet-Driven Generation of Clarifying Questions for Conversational Search

Ivan Sekulić
Faculty of Informatics
Università della Svizzera italiana (USI)
Lugano, Switzerland
ivan.sekulic@usi.ch

Mohammad Aliannejadi
IRLab
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Fabio Crestani
Faculty of Informatics
Università della Svizzera italiana (USI)
Lugano, Switzerland
fabio.crestani@usi.ch

ABSTRACT

Clarifying an underlying user information need is an important aspect of a modern-day IR system. The importance of clarification is even higher in limited-bandwidth scenarios, such as conversational or mobile search, where a user is unable to easily browse through a long list of retrieved results. Thus, asking clarifying questions about user's potentially ambiguous queries arises as one of the main tasks of conversational search. Recent approaches have, while making significant progress in the field, remained limited to selecting a clarifying question from a predefined set or prompting the user with vague or template-based questions. However, with the recent advances in text generation through large-scale language models, an ideal system should generate the next clarifying question. The challenge of generating an appropriate clarifying question is twofold: (1) to produce the question in coherent natural language; (2) to ask a question that is relevant to the user query. In this paper, we propose a model that generates clarifying questions with respect to the user query and query facets. We fine-tune the GPT-2 language model to generate questions related to the query and one of the extracted query facets. Compared to competitive baselines, results show that our proposed method is both natural and useful, as judged by human annotators. Moreover, we discuss the potential theoretical framework this approach would fit in. We release the code for future work and reproducibility purposes.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**;
• **Information systems** → **Search interfaces**.

KEYWORDS

conversational search, clarifying questions, question generation

ACM Reference Format:
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 the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3471158.3472257

ICTIR '21, July 11, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472257

1 INTRODUCTION

A Conversational Search (CS) aims to satisfy a user information need through a written or spoken conversation with the user in natural language. With the recent rise in popularity of conversational assistants, such as Alexa, Cortana, Google Home or Siri, CS attracted significant attention from the research community. Early work on the topic identified several components of CS and pointed to multiple potential research directions [4, 33]. One of the key aspects of conversational search is a mixed-initiative paradigm, where user and system may take initiative at different points in time. Users take the initiative through the description of their information need and can at any point during the conversation ask for additional information. However, in the advocated mixed-initiative paradigm, the system, while providing necessary information, can take the initiative to clarify the user's intent by asking appropriate clarifying questions.

Asking clarifying questions has been shown to be beneficial both to the user and to the IR system. The system benefits from clarifications as it acquires additional information from the user, resulting in significant improvements in retrieval performance [3]. Additionally, user satisfaction has been reported to increase when prompted with clarifications in a voice-based system [21] or search engine [49]. Although significant progress in the area of clarifying user intent has been made, current approaches remain limited to either question selection from a predefined pool of questions [2, 3], template-based question generation [49], or unsuitable for a fully conversational setting due to the format of the question [49]. Each of the mentioned works make certain assumptions about the data and problem, such as being able to collect all possible clarifying questions [3], or being able to fit all the questions in a limited number of templates [49]. We argue that, while making such strong assumptions is necessary as a starting point towards studying the effect of clarifying questions, it does not represent a real-world scenario. Therefore, an ideal IR system should be able to generate any types of clarifying questions.

A good clarifying question needs to steer the conversation towards a clear formulation of the user's information need, aiming to facilitate retrieval of relevant documents. Two main challenges arise in generating a good clarifying question: (1) to decide if the content of the question is about a specific facet of the user's query or a somewhat more general aspect; (2) to generate the question in coherent and fluent natural language. The two challenges have been explored separately, but the unified solution is still lacking in the research community. In this paper, we propose an approach that addresses both of problems.

To tackle the first challenge, we propose to generate a clarifying question based on one or more query facets, thus tackling the problem of the lack of semantic guidance of recent question generation approaches [37]. Given a user's query, the conversational IR system should first perform a facet extraction step. Query facets can be extracted from the retrieved set of documents [22], knowledge graphs [15], query logs [49], or search engine query autocompletion features [39]. We propose a question generation model that produces a question based on the initial query and several keywords representing one of the extracted facets. In order to generate a coherent question, thus tackling the second challenge, we propose to employ the large-scale language model GPT-2, which is capable of generating text of near human quality [32].

More specifically, in this paper we formally define and propose a query- and facet-conditioned clarifying question generation model. To this end, we fine-tune a large-scale language model, specifically GPT-2, conditioned on the user query and one of its facets. Additionally, we construct a dataset of query-facet-question triplets, to use as ground-truth for training our model. The triplets are extracted from the ClariQ dataset [2] and are intended to simulate a real-world scenario, where facet terms would be automatically extracted by any of the aforementioned techniques. Moreover, by synthetically creating facet keywords we eliminate the possibility of propagated error from the facet extraction algorithm, ensuring the consistency in training of our models. Experimenting with different facet extraction methods goes beyond the scope of this paper and is left for future work. Instead, we evaluate the proposed method of generating questions with a number of automatic NLG metrics, as well as human annotators. In a crowdsourcing study, we compare the questions generated by our method with those generated by baseline models in terms of *naturalness* and *usefulness*. The results show that query- and facet-conditioned question generation outperforms the template-based and the query-conditioned GPT-2 in both dimensions.

Thus, our contribution is threefold:

- We present a novel approach to generating clarifying questions. We employ large-scale language models driven by query facets for the task.
- We show the plausibility of our approach, by proposing a semantically-controlled generative model. We fine-tune GPT-2 conditioned on an initial query and one of its facets. We release the code for future work and reproductibility purposes.[1]
- Alongside the automated evaluation of the proposed model, we perform human evaluation. Furthermore, to complete the study, we discuss the limitations, future work, and potential theoretical framework where our approach could be utilised.

The rest of the paper is organised as follows. Section 2 reviews related work concerning conversational search, natural language generation, and facet extraction. Next, in Section 3 we explain our approach for generating clarifying questions and the data used to fine-tune our models. Evaluation of the models is presented in Section 4, followed by a discussion on the limitations and future work in Section 5. Finally, we conclude the study in Section 6.

[1]Github repository: https://github.com/isekulic/CQ-generation.

2 RELATED WORK

Our work primarily belongs to the broad area of conversational information retrieval (IR), with a focus on clarifying questions in mixed-initiative conversational search. In this section, we review the recent literature on the topic.

2.1 Conversational Search

Due to the rise of conversational assistants, e.g., Alexa and Siri, conversational AI recently attracted a lot of attention from the research community. The same is true for the IR community, as the report from the Dagstuhl Seminar N. 19461 [4] identifies conversational search as one of the key areas of IR in the upcoming years. Moreover, the event highlighted several potential research directions and stressed the importance of user-system interaction in conversational search. Radlinski and Craswel [33] have also emphasised mixed-initiative paradigm as one of the desirable properties of a conversational search system. Under this paradigm, the system is not passive and can take initiative by prompting the user with clarifying questions or conversation-engaging information [19]. Similarly, a "System Ask-User Respond" framework has been proposed by Zhang et al. [52] for product recommendation in e-commerce, where questions about certain aspects of a query are prompted to the user to clarify their needs.

Other recent research efforts cover various aspects of conversational search systems. The TREC Conversational Assistant Track (CAsT) [9] fosters research in multi-turn passage retrieval task, where system needs to understand the conversational context and retrieve appropriate passages from the collection. However, the system in that setting is not proactive and is not encouraged to ask clarifying questions. Generating appropriate clarifying questions is the main focus of our paper, thus we review the related work on asking for clarifications in greater detail in the next Section. Other areas of conversational search include studies concerned with user intent classification [31], biases in conversational search [16], response ranking [9, 42, 44], user engagement prediction [43], and query rewriting [30, 47].

2.2 Asking Clarifying Questions

Clarifying the user information need is an important aspect of any IR system, and is especially important in a conversational setting. *Ad hoc* IR systems deal with ambiguous and faceted queries by result diversification, where users would be presented with relevant documents for several different aspects of the query, enabling users to scroll through them to find what they need [40]. The emphasis on clarification comes form the fact that conversational search is often carried out in limited-bandwidth scenarios, such as speech-only or mobile interfaces [1], thus making it impossible to present a large range of the results the user.

Asking clarifying questions with the goal of finding the underlying user information need has recently been studied. Aliannejadi et al. [3] proposed an offline evaluation methodology for asking clarifying questions with Qulac, a dataset consisting of question-answer pairs for faceted and ambiguous queries. A similar methodology has been applied in the ClariQ challenge [2], where the task is to select the most appropriate clarifying question from a pre-defined set of questions. They find that asking the right question can lead

to significant improvements in retrieval performance. Additionally, user satisfaction has been reported to increase when prompted with clarifications in a voice-based system [21]. Moreover, Zamani et al. [49] report highly positive feedback from users engaged with the clarification panes in search engine. Thus, it is evident that asking clarifying questions is beneficial for both, the user and the system.

Hashemi et al. [17] propose GuidedTransformer that leverages information from conversation history, retrieved documents, and potential clarifying questions. Their approach yields significant improvements in the question selection task on Qulac. Zou et al. [53] conduct an empirical study on users willingness to to respond to clarifying questions and their usefulness perceived by users, concluding that most users are willing to answer 6-10 yes-or-no questions.

Recently, Zamani et al. [49] proposed a template-based and a sequence-to-sequence generative model to produce clarifying questions regarding specific aspects of a user's query. Our work differs from their approach in three main aspects. First, they extract query aspects from 1.6 billion query reformulations from Bing logs, which are not publicly available. We assume that query facets can be extracted from the collection itself, by employing a suitable facet extraction method. This makes our approach more generalisable and supported by our collection, which prevents prompting the user with clarifying questions that potentially are not answerable with the documents from our collection. Second, Zamani et al. propose clarifying question generation for search engines in a form of clarification panes. Clarification pane consists of a general question and several clickable answers, which makes it unusable in purely conversational setting. Lastly, our generated questions are more natural. Interaction naturalness has been pointed out to be an important property of a conversational search system [4], as interactions in natural language are distinguished from the ones driven by keywords in classical IR.

Rosset et al. [37] tackle the task of question suggestion in a "People Also Ask" search engine setting. They argue that a useful question is not simply related to the topic of a user's query, but should also be "conversation leading" and provide meaningful information for the user's next step. They propose a BERT-based and a generative GPT-2-based model for question suggestion. They find that questions generated by GPT-2 are syntactically correct, but less useful than the ones selected by BERT from a pre-defined pool of questions. The authors suggest that a reason for the inferior performance of GPT-2 might be due to the lack of explicit guidance in semantics. We overcome this shortcoming by grounding our question on query facets.

2.3 Natural Language Generation

The rise of large pretrained language models brought significant progress in various tasks of IR and NLP, including natural language generation (NLG). One of the most prominent models is GPT [32], with variations of GPT-2 and GPT-3. These are deep autoregressive generative models, trained on large amount of textual data, which makes them an extremely powerful natural language generators. Early work includes a hierarchical recurrent encoder-decoder for generative context-aware query suggestion [45]. There are also several attempts in generating a more controllable text, e.g., CTRL [20] and Grover [50]. Closer to our work of conditioning language generation on several keywords and a query is PPLM [10]. PPLM steers the language generated with a decoding scheme using keywords and classifiers. Moreover, Peng et al. [30] propose a semantically-conditioned GPT for conversational response generation from dialogue acts. Our clarifying question generation model differs in two main parts: (1) We generate questions, not any text; (2) We condition question generation on two parts: query and facets. Generating questions has also been studied by the natural language processing (NLP) community [34, 35], where the task is to produce questions about a given document, rather to clarify user information need.

2.4 Facet Extraction

Facet extraction from search results has been studied for a long time in *ad hoc* information retrieval settings. Deveaud et al. [11] propose a Latent Concept Modelling (LCM) method for modelling latent search concepts in order to better understand the conceptual view of an underlying user's need. Their method is based on Latent Dirichlet Allocation (LDA) to identify specific query-related topics from the top K documents retrieved. They define the extracted topics as latent concepts of user's query. Kong and Allan [22] develop a graphical model to recognise facets from the set of candidate terms, extracted from the retrieved documents. They formally define the difference between query subtopic/aspect, semantic class, and a facet. Query facets can also be extracted from search engine query logs [49], query autocompletion [39], or knowledge graphs [15]. We note that any of these methods can yield query facets suitable for the input of our proposed clarifying question generation model.

3 CLARIFYING QUESTION GENERATION

In this section, we describe our approach to generating clarifying questions that are conditioned on a specific aspect of a given query. We first formally define the problem and describe the data acquisition for training such a model. Then, we propose a training method for fine-tuning GPT-2. Finally, we present and analyse the results on the dataset used for fine-tuning in order to quantify its effectiveness with both automatic metrics and human judgements acquired through crowdsourcing.

3.1 Semantically-Guided Question Generation

We define our task of generating clarifying questions as a sequence generation task. Formally, given a facet f and a query q, the model needs to construct a valid clarifying question cq. Facet f is one of the facets taken from the set of extracted facets, as described in section 5.1. Query q is issued by a user. A clarifying question is then defined as a function of the query and its facet:

$$cq' = QuestionGenerator(q, f) \tag{1}$$

To the best of our knowledge, we are the first to condition the clarifying question generation on both the query and the facet.

3.1.1 Language Modelling. Current state-of-the-art methods for text generation are based on large-scale auto-regressive language models [8, 32]. The goal of language modelling is to learn probability distribution $p_\theta(\boldsymbol{x})$, given example sequences $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$, where n is a sequence length and θ are parameters of our model. In auto-regressive language generation, we decompose language

modelling into next-word prediction, factorising the distribution $p_\theta(\boldsymbol{x})$ using the chain rule of probability:

$$p_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} p_\theta(x_i|x_{<i}) \tag{2}$$

We utilise pre-trained GPT-2, an auto-regressive model trained to learn $p_\theta(\boldsymbol{x})$ as in Equation 2. By doing so, we take advantage of the fact that GPT-2 is trained on large amount of text data, which already makes it powerful for language generation, as shown by it's performance on various downstream tasks [32]. However, as our goal is not just to generate any text sequences, but to generate questions conditioned with the initial query and one of its facets, we essentially model $p_\theta(\boldsymbol{x}|q, f)$ where q and f are a query and its facet, respectively. Building on top of recent approaches to semantically-controlled language generation [20, 30], we model:

$$p_\theta(\boldsymbol{x}|q, f) = \prod_{i=1}^{n} p_\theta(x_i|x_{<i}, q, f) \tag{3}$$

Learning of parameters θ is done by minimising the negative log-likelihood of the conditional probabilities in Equation 3, that is:

$$\mathcal{L}_\theta(D) = -\sum_{j=1}^{|D|}\sum_{i=1}^{n} log p_\theta(x_i^j|x_{<i}^j, q^j, f^j) \tag{4}$$

where $D = \{(x^j, q^j, f^j)\}_{j=1}^{N}$ is the dataset of triplets consisting of clarifying questions x, queries q, and facet terms f. We fine-tune our model on D, in order to be able to: (1) generate questions, not just any textual sequence; (2) generate questions about a given facet of a given query.

3.1.2 Inference. In order to generate clarifying questions, we use a combination of state-of-the-art sampling techniques to generate a textual sequence from the trained model. Namely, we utilise temperature-controlled stochastic sampling with top-k [14] and top-p (nucleus) filtering [18]. By tuning the temperature, we increase the likelihood of high probability words and decrease the likelihood of low probability words, or vice versa. We do so by directly adjusting the softmax over $p_\theta(\boldsymbol{x})$, making the probability to predict the i-th token from the vocabulary:

$$p_i = \frac{exp(x_i/T)}{\sum_j exp(x_j/T)} \tag{5}$$

where x_i are the logits for i-th token in the vocabulary, and T is the temperature. Moreover, we restrict the sampling to only the top-k most likely next tokens, redistributing the probability mass over the remaining k tokens. As some tokens can be samples from a sharp distribution, while others from a flat distribution, top-k sampling shows some shortcomings. Parameter k is fixed and set before sampling, so the possibility of sampling an irrelevant token from a sharp distribution increases. At the same time, setting a low k might restrain model's token variety. To overcome these potential issues, we experiment with top-p (nucleus) sampling [18], where we consider the minimum number of next possible tokens whose summed probabilities amount to $p \in [0, 1]$. Again, the probability mass is then redistributed among the remaining tokens and standard sampling is performed from the reduced set of tokens. Our final experiments are performed with temperature T set to 0.7, k to 0, and p to 0.9, as this combination of parameters showed promising results in the early analysis, and is supported by previous research [18].

Table 1: Statistics of ClariQ-Fkw dataset used for fine-tuning our generative model. Symbol #N represents "number of".

	train	dev
Number of samples	1756	425
Average #N facet terms	1.9	1.8
Std of #N facet terms	1.0	0.9
Number of unique queries	187	50
Avg #N questions per query	9.39	8.5
Std #N questions per query	2.7	2.6

3.2 Dataset Construction

To the best of our knowledge, a dataset suitable for our purpose of training a sequence generation model conditioned on two different segments, i.e., a query and its facet, is not available in the IR/NLP community. Moreover, as we need to generate *questions*, the specificity of our needs increases. Thus, we adapt a simple data filtering to transform ClariQ data samples to the appropriate (q, f, cq) triplets. ClariQ [2] consists of queries and corresponding clarifying questions, which are acquired from crowdsourcing. However, since questions are not about a specific facet, we extract the facet keywords using the following procedure:

(1) We discard the beginning of the question. Since most of the questions fall under a few templates, with 10 different prefixes we cover around 80% of the dataset.
(2) We keep only content-bearing words, specifically verbs, nouns, and noun phrases. The part-of-speech tagging is done with NLTK [6]. This ensures that our synthetically created facet keywords resemble the output of any of the most typical facet extraction methods.
(3) Finally, from the remaining set of words we remove the ones that also appear in the query. This is done in order to keep our dataset as general as possible, as the facets or subtopics in real-world scenario are unlikely to contain words from the initial query.

This procedure gives us a dataset of more than 2000 triplets, whose examples can be seen in Table 2. Table 1 shows statistics of the created dataset – ClariQ-FKw, where FKw stands for Facet Keywords. We see that the average number of facet terms is 1.9, with standard deviation of 1.0. Thus, for future work, we suggest that facet extraction method follows similar characteristics. Additionally, we notice a high number of questions for the same query. This further enforces our model to consider both the queries and the facet terms when forming a clarifying question.

3.3 Fine-tuning GPT-2

We fine-tune the GPT-2 [32] model as our clarifying question generation function *QuestionGenerator* from Equation 1. Specifically, we form the input to the GPT-2 model as follows:

$$input\_seq = f[SEP]q[bos]cq[eos] \tag{6}$$

Table 2: Training triplets from ClariQ-FKw as an input to the GPT-2 model, with clarifying question as a language modelling target.

Initial Query	Facet terms	Clarifying Question
Tell me about cass county missouri	list homes sale	are you interested in a list of homes for sale in cass county
What is von Willebrand Disease?	treatments	are you interested in learning about treatments for von willebrand disease
What is von Willebrand Disease?	types	are you interested in the types of von willebrand disease
Tell me about atypical squamous cells	aytypical desciption	are you interested in a desciption of aytypical squamous cells
Tell me about atypical squamous cells	result test	are you interested in atypical squamous cells in a test result
Tell me about atypical squamous cells	urine	are you interested in atypical squamous cells in urine
Tell me more about Rocky Mountain News	archives	are you interested in news archives
Tell me more about Rocky Mountain News	information park national	are you interested in information about the national park
Find me information about the sales tax in Illinois.	state	are you interested in how the illinois state tax is determined

where $[bos]$, $[eos]$, and $[SEP]$ are special tokens indicating the beginning of sequence, the end of sequence, and a separation token, respectively. Query q, facet terms f, and clarifying question cq are tokenized prior to constructing the full input sequence to the model. Additionally, we further feed the model with segment embeddings, which indicate different segments of the input sequence, namely q, f, and cq. All of the text pre-processing and exact formation of the input sequence is available in our Github repository.[2]

To calculate the language modelling loss $\mathcal{L}$, we project the hidden-state on the word embedding matrix to get logits and apply a cross-entropy loss:

$$\mathcal{L} = \sum_{i=1}^{N} p_i log(\widetilde{p}_i) \tag{7}$$

where p is the true distribution and $\widetilde{p}$ is the predicted distribution of our model. The loss is applied only on the clarifying question (cq) part of the sequence, while predecending tokens of the facet and the query are masked out.

We fine-tune the models with a batch size of 32, learning rate of 5×10^{-5} for 8 epochs. The hyperparameters were chose based onn previous research on text generation and a grid search of the optimal learning rate and number of epoch on the development set. The training takes about one hour on GeForce GTX 1080 Ti GPU. We use the HuggingFace [48] implementation of the GPT-2 model. During inference, we omit the clarifying question cq part from the Equation 6 of the input sequence to the model. We generate the question by sampling token by token, as described in Section 3.1.2.

4 EVALUATION

We evaluate our question generation model using a number of standard NLG metrics. Additionally, as some NLG metrics received heavy criticism from the research community, we make use of crowdsourcing for gathering human judgements. With the evaluation, we aim to assess the plausibility of query- and facet-conditioned GPT-2 for the task of generating clarifying questions.

4.1 Baselines

We compare our query and facet-conditioned model (QF-GPT) to two competitive baselines. These are:

[2]Github URL after anonymity period ends.

4.1.1 Template-based Question Generation (TB). A template-based approach to generating clarifying questions produces a question by simply filling a slot in a pre-defined question. More specifically, we construct the questions by filling the slot *[facet term]* in a question "Are you interested in *[facet term]*" with one of the query facet terms. This specific question was chosen as it is the most common way of constructing clarifying questions in the Qulac dataset [3]. The template-based approach has been widely used in various IR tasks [49, 51].

4.1.2 Query-conditioned GPT-2 (Q-GPT). As our second baseline model, we fine-tune GPT-2 to generate clarifying questions as described in Section 3, but without feeding facet terms as input to the model. This simulates the behaviour of most chatbots, as they solely rely on conversational history, rather than the explicit conversational aspect that should be discussed. This model resembles the approach of Rosset et al. [37], who employ a GPT-2-based model for conversational question suggestion. They train their model on query-question pairs in a pointwise setting. One limitation of that approach, as pointed out by the authors, is the lack of explicit semantic guidance. Our hypothesis is that this baseline model will generate fluent responses. However, clarifying questions are generated solely based on the query and the memorised generic utterances in the weights of the model, rather than being about specific aspect of the query. Moreover, the generated questions have a risk of not being answerable by our collection.

4.2 Automated Metrics

We evaluate our generated questions against reference questions from ClariQ. To this aim, we compute a number of standard metrics for evaluating generated language. The first two are widely adopted metrics: BLEU [29] and ROUGE [26], which are based on n-gram overlap between the generated text and the reference text. Additionally, we compute METEOR [5], which was reported to have higher correlation with human judgements than BLEU and ROUGE [5]. METEOR mitigates the shortcomings of BLEU and ROUGE by not just counting the overlap of n-grams, but also considering their stems, WordNet synonyms, and paraphrases. Furthermore, we compute the EmbeddingAverage, defined as cosine similarity between the mean of the word embeddings of each token in the generated and the target questions [24].

Table 3: Evaluation of generated questions against gold standard in ClariQ. TB, Q-GPT, and QF-GPT stand for templated-based baseline, query-conditioned GPT-2 baseline, and the proposed query- and facet-conditioned GPT-2, respectively. EAC stands for EmbeddingAverageCosine. Bleu-N indicates BLEU metric calculated on N-grams.

Model	Bleu-1	Bleu-2	Bleu-3	EAC	METEOR	ROUGE-L
TB	0.316	0.169	0.101	0.890	0.212	**0.394**
Q-GPT	0.316	**0.210**	**0.150**	0.862	0.165	0.315
QF-GPT	**0.320**	0.186	0.119	**0.906**	**0.289**	0.285

4.3 Human judgements

Recent studies have revealed several flaws of the standard heuristic-based NLG metrics [7, 27, 28, 36, 38]. The criticism comes from the low correlation of the automated metrics with human judgements, thus making the metrics untrustworthy or even misleading. Moreover, Stent et al. [46] found that several automatic metrics, including BLEU, correlate negatively with human judgements on fluency of generated text. Thus, in order to properly evaluate our generated clarifying questions, we opt for human annotations. As stated before, a good clarifying question should be in a coherent, fluent natural language and relevant to the topic of the conversation. For that reason, we evaluate two different aspects of our generated questions: *naturalness* and *usefulness*, described in the next Section.

Evaluation is done in a pairwise setting, i.e., an annotator is presented with two questions generated by two different models and has to choose which one is more natural, or more useful, depending on the task. We evaluate the performance of the models in a pairwise setting, as it has been shown to be more reliable and more consistent across annotators than for example the Likert scale [25]. We compare our main facet-guided clarifying question generation model with the two baseline models described in Section 4.1. Additionally, we compare the two baselines among themselves.

4.3.1 Naturalness. An important feature of a conversational search system is its natural responses in fluent and coherent natural language [4]. Inspired by several studies in various tasks of NLG [30, 38], we define *naturalness* as a question being natural, fluent, and likely generated by a human. Similar definitions exist in a wide range of work, including fluency [7, 46] and humanness [41]. For example, a clarifying question "Would you like to know more about magnesium-rich foods?" is more natural and fluent than "Are you interested in magnesium foods?".

4.3.2 Usefulness. Rosset et al. [37] define a *usefulness* metric to describe conversation-leading clarifying questions. They argue that questions can be relevant to the user's query, but that does not make them necessarily useful. For example, given a query "Tell me about kiwi fruit.", a question such as "Would you like to know about kiwi?" is arguably relevant to the query, but it is useless, as it is too broad. Moreover, a clarifying question such as "Are you interested in the business model of NZ kiwi fruit company?" is also related, but also useless due to it being far too specific. This definition of usefulness can be related to adequacy [7, 46] and informativeness [30].

4.4 Crowdsourcing

We use the crowdsourcing platform Amazon Turk[3] to acquire annotators, i.e., workers, for our evaluation study. We use in total more than 100 different workers, all based in United States, with minimum approval rate of 94% and minimum number of accepted HITs so far of 1000. We limit the number of annotations per worker to 25 question pairs, in order to eliminate tiredness. Moreover, each question pair is judged by three different workers yielding more than 2000 labels in total, across all of the experiments. We use majority voting to decide on the final label. We compute Fleiss' kappa κ to assess the degree of agreement per annotated pair. The outcome reaches low, fair, and moderate agreement, depending on the compared models. Workers with suspiciously low performance on the manually curated test pair questions were eliminated from the study.

4.5 Results

4.5.1 Automated Metrics. Results of our automatic evaluation are presented in Table 3. We notice that it is not clear which model is the best based solely on automated metrics. Specifically, QF-GPT yields the best results among the models in terms of Bleu-1, EAC, and METEOR. However, Q-GPT shows the best performance in terms of Bleu-2 and Bleu-3, while TB outperforms all other models in terms of ROUGE-L. This is a well-known issue in evaluating generated text, as the automated metrics rarely highly correlate with the real scenario [7, 27, 28, 36, 38]. For that reason, we rely on human judgements to more accurately estimate the performance of the models. With an automated evaluation, we can only add to the large body of work on criticism of automated metrics for NLG.

4.5.2 Human Evaluation. Comparisons of the baseline models, i.e. Template-based and GPT-2-query, with our proposed facet-guided GPT-2 on *naturalness* and *usefulness* are presented in Tables 4 and 5, respectively. We conducted a binomial test for each of the model comparison that we've made. The p-values are presented in Table 6.

When assessing *naturalness* of questions generated by our model and the baseline models, we notice several key observations:

- GPT-2-based models (Q-GPT and QF-GPT) produce more natural clarifying questions than the template-based model (TB). This was our initial hypothesis and main motivation behind utilising GPT-2 for the task.
- GPT-2-based models produce questions of similar *naturalness*, as the difference between query-only model (Q-GPT) and query- and facet-guided GPT-2 model is small (51 to

[3] https://www.mturk.com

Table 4: Results on *Naturalness* between query- and facet-guided GPT-2 (QF-GPT), query-only GPT-2 (Q-GPT), and the template-based (TB) models.

		Wins		
		TB	Q-GPT	QF-GPT
Loses	TB	-	69	59
	Q-GPT	31	-	49
	QF-GPT	41	51	-

Table 5: Results on *Usefulness* between query- and facet-guided GPT-2 (QF-GPT), query-only GPT-2 (Q-GPT), and the template-based (TB) models.

		Wins		
		TB	Q-GPT	QF-GPT
Loses	TB	-	63	60
	Q-GPT	37	-	57
	QF-GPT	40	43	-

Table 6: The p-values of binomial statistical test for the significance of the comparisons between different models for *naturalness* and *usefulness*.

	Naturalness	Usefulness
TB & Q-GPT	0.01	0.01
TB & QF-GPT	0.08	0.05
Q-GPT & QF-GPT	0.92	0.19

49). This is also expected, as both methods produce fluent questions.

Moreover, from the study on *usefulness* of generated questions, we observe:

- GPT-2-based models outperform the template-based model. The initial hypothesis was that query-aware GPT-2 (Q-GPT) might not perform so well on *usefulness*, as the questions are not grounded in any specific facet of the query. However, careful examination of the annotated questions suggested that even though the model is not guided by an explicit facet, it is still generating questions that are very often facet-based. The crucial difference to the QF-GPT being that we can not control which facet Q-GPT will ask questions about, as it can only ask about the ones that are implicitly saved in GPT-2's weights.
- QF-GPT outperforms Q-GPT, which confirms our hypothesis that facet-guided question generation is more useful. According to the Table 6, the difference is not statistically significant. However, although Q-GPT is capable of producing clarifying questions related to user's query, we have no control over the content of the questions. Thus, facet-driven QF-GPT poses itself as a stronger choice for the task.

Human judgements confirm our hypothesis that GPT-2 can generate fluent and natural clarifying questions, while allowing explicit semantic guidance, when trained accordingly. Next, we perform qualitative study of actual questions generated by the proposed model.

4.5.3 Qualitative Study. Table 7 shows several examples of questions generated by our proposed facet-guided model and the baselines. We can observe that all questions generated by GPT-2-based models are indeed fluent and coherent. However, a key difference is that query-conditioned GPT-2 generates questions that are often not entirely relevant to the topic of the conversation. This is known as the hallucination of generative models, where generated responses do not correspond to the real-world [12, 13]. By grounding our question generation model in facets, we gain control of the conversation and eliminate the hallucination effect, thus making our model more useful for clarifying the user need.

5 LIMITATIONS AND FUTURE WORK

5.1 Facet Extraction from Retrieved Documents

In this study, we proposed to generate clarifying questions about certain query facets. However, we have dealt only with facets acquired through controlled keyword extraction, leaving automatic facet extraction as a separate part in a conversational system for future work. We now describe several potential approaches for acquiring query facets and formally define the required properties such methods should have in order to be easily included into our model.

Recent work on clarifying question generation extracted query aspects from the search log of a major search engine [49]. As such logs are not widely available, we suggest extracting the facets from the set of documents retrieved in response to the initial query. Several methods for facet extraction from a set of documents already exist in the literature [11, 22] and are largely based on clustering and language modelling approaches. Formally, given a query q, we retrieve a set of documents $\mathcal{D} = \{D\}$ from the collection C. We then extract a set of N facets $\mathcal{F} = \{F\}$, where N is the hyperparameter representing the total number of facets to extract. Each facet F consists of a set of terms (keywords) representing it: $F = \{t\}$. Facet terms F can then be fed into our generative model in order to produce useful clarifying questions.

5.2 Conversation History

Our approach is currently limited to generating clarifying questions from the initial query only. However, one important aspect of conversational search are multi-turn interactions. Extension of our model to multi-turn conversations includes understanding user's answer to our clarifying question and deciding whether to ask a follow up clarifying question, or otherwise to show the user retrieved relevant documents. In order to ask a follow up clarifying question we can simply generate a question about some other facet of a query. However, for better results, we should also consider user's answer, as they often provide additional information and not just a *yes-or-no* answer [23]. As GPT-2 has been show to be able to capture multiple turns of information-seeking conversations [30, 47], our first attempt of extending the model would be to feed

Table 7: Examples of generated clarifying questions given the initial request and one of the facet terms.

Initial request	Tell me about kiwi		
Facet terms	information fruit	biology bird	people background historical
Template-based	Are you interested in information fruit?	Are you interested in biology bird?	Are you interested in people background historical?
Q-GPT	Are you looking for kiwi clothing?	Are you looking for kiwi reviews?	Are you interested in kiwi fitness?
QF-GPT	Are you interested in kiwi fruit?	Are you interested in kiwi birds?	Are you interested in kiwi history?
Initial request	What is von Willebrand Disease?		
Facet terms	treatments	symptoms	types
Templated-based	Are you interested in treatments?	Are you interested in symptoms?	Are you interested in types?
Q-GPT	Are you looking for a specific web page?	Are you looking for a specific medication?	Do you want to know the causes of this disease?
QF-GPT	Do you want to know what treatments are used to treat the von Willebrand disease?	Are you looking for a list of symptoms of von Willebrand?	Are you looking for a list of the diseases?

the conversation history together with the initial query and the extracted query facets to the GPT-2-based model.

5.3 Multiple Facets

One of the planned extensions of this work is to generate clarifying questions about multiple query facets, rather than just one at the time. Looking at the first example of Table 7 and the query *"Tell me about kiwi"*, our current model would generate a question about one of the facets, such as *"Do you want information about kiwi fruit?"*. However, in order to potentially minimise the number of conversational turns needed to satisfy the user information need, we could ask a question in line with *"Are you interested in kiwi fruit, kiwi bird, or New Zealand people?"*. The challenge here is to create a dataset of such questions suitable for training of generative language models, like GPT-2.

6 CONCLUSIONS

In this paper, we presented a facet-guided model for generating clarifying questions in mixed-initiative conversational search. We showed that large-scale language models, in particular GPT-2, are quite fit for the task, when fine-tuned properly. More specifically, we semantically guide GPT-2 question generation by conditioning the question on the user's original query and one of the query facets. Human judgements acquired through crowdsourcing show that clarifying questions generated by our proposed model are both natural and useful, compared to competitive baselines. Our results and discussions serve as a preliminary step towards generating clarifying questions from the query facets. Our goal was to demonstrate the capability of large-scale language models for generating clarifying questions, by showing that a model such as GPT-2 can be guided and driven towards a certain topic or goal in a conversation. Our results demonstrated the superiority of generated questions over template-based questions.

REFERENCES

[1] Mohammad Aliannejadi, Morgan Harvey, Luca Costa, Matthew Pointon, and Fabio Crestani. 2019. Understanding Mobile Search Task Relevance and User Behaviour in Context. In *CHIIR*. ACM, 143–151.

[2] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ). (2020).

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

[4] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). In *Dagstuhl Reports*, Vol. 9. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[5] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[6] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

[7] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

[8] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2978–2988.

[9] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624* (2020).

[10] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164* (2019).

[11] Romain Deveaud, Eric SanJuan, and Patrice Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17, 1 (2014), 61–84.

[12] Yuheng Du, Shereen Oraby, Vittorio Perera, Minmin Shen, Anjali Narayan-Chen, Tagyoung Chung, Anu Venkatesh, and Dilek Hakkani-Tur. 2020. Schema-Guided Natural Language Generation. *arXiv preprint arXiv:2005.05480* (2020).

[13] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. *arXiv preprint arXiv:2104.08455* (2021).

[14] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).

[15] Leila Feddoul, Sirko Schindler, and Frank Löffler. 2019. Automatic facet generation and selection over knowledge graphs. In *International Conference on Semantic Systems*. Springer, Cham, 310–325.

[16] Emma J Gerritse, Faegheh Hasibi, and Arjen P de Vries. 2020. Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 133–136.

[17] 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.

[18] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).

[19] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[20] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).

[21] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward voice query clarification. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1257–1260.

[22] 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.

[23] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 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.

[24] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. 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)*. 540–551.

[25] Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087* (2019).

[26] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[27] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2122–2132.

[28] François Mairesse, Milica Gasic, Filip Jurcicek, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 1552–1561.

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[30] Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328* (2020).

[31] Chen Qu, Liu Yang, W Bruce Croft, Yongfeng Zhang, Johanne R Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 25–33.

[32] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[33] 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.

[34] 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.

[35] 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.

[36] Alan Ritter, Colin Cherry, and William B Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 172–180.

[37] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020*. 1160–1170.

[38] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2020. A Survey of Evaluation Metrics Used for NLG Systems. *arXiv preprint arXiv:2008.12009* (2020).

[39] Alexandre Salle, Shervin Malmasi, Eugene Agichtein, and Oleg Rokhlenko. 2021. Studying the Effectiveness of Conversational Search Refinement through User Simulation. In *Proceedings of the European Conference on Information Retrieval (ECIR)*.

[40] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search Result Diversification. *Found. Trends Inf. Retr.* 9, 1 (March 2015), 1–90. https://doi.org/10.1561/1500000040

[41] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. 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)*. 1702–1723.

[42] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2020. Extending the Use of Previous Relevant Utterances for Response Ranking in Conversational Search. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC*.

[43] Ivan Sekulic, Mohammad Aliannejadi, and Fabio Crestani. 2021. User Engagement Prediction for Clarification in Search. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 12656)*. Springer, 619–633.

[44] Ivan Sekulić, Amir Soleimani, Mohammad Aliannejadi, and Fabio Crestani. 2020. Longformer for MS MARCO document re-ranking task. *arXiv preprint arXiv:2009.09392* (2020).

[45] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 553–562.

[46] Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *international conference on intelligent text processing and computational linguistics*. Springer, 341–351.

[47] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A Comparison of Question Rewriting Methods for Conversational Passage Retrieval. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings*.

[48] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* (2019), arXiv–1910.

[49] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*. 418–428.

[50] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *arXiv preprint arXiv:1905.12616* (2019).

[51] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1512–1520. https://doi.org/10.1145/3394486.3403202

[52] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 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] Jie Zou, Evangelos Kanoulas, and Yiqun Liu. 2020. An Empirical Study on Clarifying Question-Based Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2361–2364.