Few-Shot Generative Conversational Query Rewriting

Shi Yu*[1], Jiahua Liu*[1], Jingqin Yang[1], Chenyan Xiong[2],
Paul Bennett[2], Jianfeng Gao[2], and Zhiyuan Liu[1*]
Tsinghua University[1], Microsoft Research AI[2]
{yus17, yang-jq17}@mails.tsinghua.edu.cn; alphaf52@gmail.com;
{chenyan.xiong, Paul.N.Bennett, jfgao}@microsoft.com; liuzy@tsinghua.edu.cn

ABSTRACT

Conversational query rewriting aims to reformulate a concise conversational query to a fully specified, context-independent query that can be effectively handled by existing information retrieval systems. This paper presents a few-shot generative approach to conversational query rewriting. We develop two methods, based on rules and self-supervised learning, to generate weak supervision data using large amounts of ad hoc search sessions, and to fine-tune GPT-2 to rewrite conversational queries. On the TREC Conversational Assistance Track, our weakly supervised GPT-2 rewriter improves the state-of-the-art ranking accuracy by 12%, only using very limited amounts of manual query rewrites. In the zero-shot learning setting, the rewriter still gives a comparable result to previous state-of-the-art systems. Our analyses reveal that GPT-2 effectively picks up the task syntax and learns to capture context dependencies, even for hard cases that involve group references and long-turn dependencies.

KEYWORDS

Conversational Search; Query Rewriting; Few-Shot Learning

ACM Reference Format:
Shi Yu*[1], Jiahua Liu*[1], Jingqin Yang[1], Chenyan Xiong[2], and Paul Bennett[2], Jianfeng Gao[2], and Zhiyuan Liu[1]. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401323

1 INTRODUCTION

Recent advances in deep learning and text understanding facilitate the transition of information retrieval systems from keyword-based queries and "ten-blue" links to more conversational experiences. Widely viewed as a next generation IR direction, Conversational IR is favored with its ability to satisfy users' complex information needs with multi-round interactions, while also providing convenient and precise information access through conversational interfaces and portable devices.

Table 1: A Conversational Search Example in TREC CAsT

Description: The Bronze Age collapse and the transition into a dark age.	
Turn	**Conversational Queries**
Q_1	Tell me about the Bronze Age collapse.
Q_2	What is the evidence for it?
Q_3	What are some of the possible causes?
Manual Query Rewrites	
Q_2^*	What is the evidence for **the Bronze Age collapse**?
Q_3^*	... the possible causes **of the Bronze Age collapse**?

A signature of Conversational IR is its multi-round interactions with the user, an opportunity to understand and assist with more complex tasks and a challenge to query understanding. Natural conversations are concise and context dependent. Statements refer to previous discussions, omit already mentioned concepts, and assume implicit context during the conversation. Table 1 shows one such example from TREC Conversational Assistance Track (CAsT) 2019. The user begins with a fully specified query (Q_1), but quickly starts to use references (Q_2) and omissions (Q_3), which is very different from typical keyword-based search sessions.

A natural direction to tackle this challenge is to rewrite the conversational queries to de-contextualized queries that include all necessary information. The manually rewritten queries (Q_2^* and Q_3^* in Table 1) can be much better handled by existing ad-hoc ranking systems. In TREC CAsT 2019, various approaches were developed for this *conversational query rewriting* task, including IR-style query expansion/term reweighting, NLP-style coreference resolution, and neural-based query rewriting. Still, conversational query rewriting is a challenging task: there is 30%+ NDCG drop from systems that use automatic query rewriting/reformulation, compared with their counterparts using manual rewrites [1].

One top performing conversational query rewriting system in TREC CAsT is ATeam's GPT-2 generative query rewriter (a later version can be found in [5]). They feed into a pre-trained transformer language model [4] the previous and current queries in the session (e.g. Q_1, Q_2 and Q_3), and fine-tune the model to generate the fully de-contextualized query rewrite (Q_3^*). The effectiveness and simplicity of this generative model make it a promising solution for conversational search. However, their GPT-2 was trained using their large quantity of manual query rewrites on their own conversational search queries. It is not clear whether the transformer language model can still be effectively learned without large

*The first two authors contributed equally.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401323

amounts of manual query rewrite labels, which are expensive to collect and are not always available for many domains [1].

This work studies learning with GPT-2 in conversational query rewriting using few or even zero manual rewriting labels. We propose two approaches that generate weak supervision signals for this task using the ad hoc search sessions abundant in search logs. The first is a rule-based approach which uses two simple rules to omit or co-refer repeated noun phrases in search sessions. The second is a self-supervised learning approach that uses a handful of manually created query rewrites and conversational queries to train a GPT-2 model, as a simplifier, to convert the ad hoc search sessions to more context-dependent, conversational-like queries. These approaches provide large amounts of weak supervision data for the GPT-2 rewriter to learn the context dependencies in concise conversational search queries.

In the few-shot setting where only TREC CAsT's manual query rewrites of 50 conversational sessions are used, ranking with our query rewrites outperforms the best automatic runs in CAsT 2019 by 12% NDCG@3. In the zero-shot setting, where no manual query rewrites are used, our weakly supervised GPT-2 still gives comparable result to the previous best automatic run in CAsT.

We further explore the capability of GPT-2 in few-shot learning by fine-tuning only on a handful of manual query rewrites. We observe that, surprisingly, the pre-trained transformer is able to pick up this task with *as few as three conversational sessions*. We find that GPT-2 quickly and effectively learns the task syntax: to generate questions instead of stories and to resolve the context dependencies using previous turns. We also observe that the model accurately deals with hard cases such as ones containing long-term and multiple coreferences.[1]

2 PRELIMINARIES

This section describes the application of GPT-2 on the conversational query rewriting task.

Conversational Query Rewriting. Conversational search systems aim to find relevant documents for queries in a conversational search session $S = \{Q_1, ...Q_k..., Q_N\}$ [1]. The conversational queries are often concise and their information needs are often presented in the previous queries.

The conversational query rewriting task is to rewrite a context dependent query Q_k to a fully de-contextualized query Q'_k, with the help of previous queries $Q_{<k}$:

$$Q'_k = \texttt{QueryRewriter}(Q_k; Q_{<k}), \tag{1}$$

which better reflects user intent and is easier for ad hoc search.

We use GPT-2 [1, 4] to directly generate the query words $\{w'_1, ... w'_i..., w'_M\}$ in Q'_k one by one as:

$$w'_i = f(w'_{<i}; Q_k, Q_{<k}). \tag{2}$$

where f is transformer decoder and the input is in the format of:

$$Q_1 \circ [SEP] \circ ... \circ [SEP] \circ Q_k \circ [BOS] \circ [w'_1, ..., w'_{i-1}], \tag{3}$$

Both training and inference use standard GPT-2 [4], which is adapted to our task to generate queries instead of plain text [1].

[1]Our code, data, and analyses results are publicly available at https://github.com/thunlp/ConversationQueryRewriter.

In training, the target query $Q^*_k = \{w^*_1 ... w^*_m\}$, either ground truth labels or weak supervision labels, are used to train the model.

Ranking with Query Rewrites. With the de-contextualized query rewrite Q'_k, standard ad hoc ranking can be used to complete the conversational search task. We use the standard BM25 to retrieve 100 documents and a BERT ranker to rerank them [2, 3].

3 WEAK SUPERVISION

One concern of generative query rewriting is that gold query rewrites Q^* are expensive to obtain. This section describes how we leverage the ad hoc search sessions, available in search logs, to construct weak supervision data to mimic conversational search sessions with target query rewrites.

As current search engines are still moving towards conversational experiences, a typical ad hoc session is less likely to include many coreferences or omissions. Users may not expect search engines to resolve context dependency and tend to write fully specified queries. These fully specified queries, on the other hand, can be used as Q^* in the conversational query rewriting task.

We consider ad hoc search sessions as pseudo target query rewrites, $\tilde{S^*} = \{\tilde{Q}^*_1, ...\tilde{Q}^*_i..., \tilde{Q}^*_N\}$, and convert them to conversation-like sessions: $\tilde{S} = \{\tilde{Q}_1, ...\tilde{Q}_i..., \tilde{Q}_N\}$. Then $(\tilde{S}, \tilde{S^*})$ pairs can serve as weak supervision to approximate real conversational queries S and manual query rewrites S^*.

To perform this conversion $(\tilde{S^*} \rightarrow \tilde{S})$, we propose two approaches, based on rules and self learning, respectively.

Rule-Based. The first approach uses two simple rules to mimic two discourse phenomena in conversations: *omission* and *coreference*. We perform the following operations on search sessions:

- `Omission.` A noun phrase is omitted if it occurs after a preposition and appears in previous queries;
- `Coreference.` Otherwise, previously appeared singular and plural noun phrases are respectively replaced with "it" (96%), "he" (2%), or "she" (2%), and "they" (75%) or "them" (25%).

Both operations can be done efficiently on a vast amount of sessions.

Self-Learn. The second approach uses self-supervised learning and trains a GPT-2 model, known as query simplifier, to generate the conversation-like sessions $\tilde{S}$ using $\tilde{S^*}$. Differing from query rewriting that aims to "put contexts back" to the query, the query simplifier learns to generate contextual queries containing few information presented in previous queries of the same session.

The query simplifier uses a handful manual query rewrites, and learns to simplify the fully specified query to a contextual query as:

$$Q_k = \texttt{QuerySimplifier}(Q^*_k; Q_{<k}). \tag{4}$$

Except reversing the source and target $(S^* \rightarrow S)$, the same GPT-2 setup described in the previous section is used. The query simplifier, trained with a few manual query rewrites, is then applied to the ad hoc search sessions (MS MARCO) to generate more conversation-like sessions $(\tilde{S^*} \rightarrow \tilde{S})$.

4 EXPERIMENTAL METHODOLOGIES

Our experiments use the TREC CAsT 2019 benchmark for evaluation and the ad hoc sessions from MS MARCO for weak supervision.

TREC CAsT Conversation Search Benchmark. The dataset consists of 50 conversational search sessions S, each containing

Table 2: Overall Results on TREC CAsT 2019 Conversational Search Task. * marks scores from [1]. All our runs use the same ranking model. BLEU-2 are compared with Oracle Queries. QA-ROUGE evaluates the answer quality.

Method	BLEU-2	NDCG@3	QA-ROUGE
TREC CAsT Auto Runs			
clacBase*	–	0.360	–
pgbert*	–	0.413	–
CFDA_CLIP_RUN7*	–	**0.436**	–
CAsT Queries			
Original	0.659	0.304	0.231
AllenNLP Coref w/o sw	–	0.314	–
AllenNLP Coref w/ sw	0.750	0.437	0.278
Oracle	1.000	0.544	0.314
Zero-Shot Rewriter			
GPT-2 Raw	0.112	0.124	0.196
MARCO Raw	0.380	0.172	0.183
Rule-Based	**0.755**	**0.437**	0.266
Few-Shot Rewriter			
Rule-Based + CV w/o PLM	0.178	0.065	0.151
Self-Learn	0.750	0.435	0.263
CV	0.793	0.467	0.280
Rule-Based + CV	**0.809**	**0.492**	**0.291**
Self-Learn + CV	0.804	0.491	**0.291**

around ten conversational queries. The task is to retrieve and rank relevant passages for each query in S from the MS MARCO passage collection and TREC Complex Answer corpora. Standard TREC relevance judgments are provided. CAsT provides official manually rewritten queries for 50 conversational topics [1]. We also manually label answer text for TREC CAsT questions and evaluate question answering result.[2]

Evaluation Metrics. The main metric in CAsT is NDCG@3 averaged on all turns. We also evaluate the similarity between automatic rewrites and ground truth using BLEU-2 and the question answering result using ROUGE-L.

Weak Supervision Dataset and Preprocessing. The ad hoc search sessions are collected from MS MARCO[3]. It includes 152K artificial sessions, with MS MARCO queries automatically aligned to Bing search sessions. We process the DEV sessions to contain more question-like queries, by only retaining those with question words, and converting them to the weak supervision data (Sec. 3).

Baselines. We compare with the following query reformation baselines. They all use the same ad hoc ranking as ours.

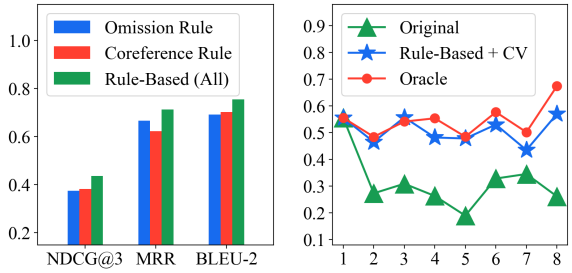
`Original` uses the original queries from TREC CAsT.

`AllenNLP Coref` uses the query reformulations (with or without stopwords) provided by CAsT where AllenNLP is used to resolve coreferences in search sessions.

`GPT-2 Raw` directly applies the pre-trained GPT-2 for query rewriting without fine-tuning.

`MARCO Raw` fine-tunes GPT-2 on MS MARCO sessions for a language modeling task instead of the rewriting task.

[2]The answers are available at https://github.com/thunlp/ConversationQueryRewriter.
[3]https://github.com/microsoft/MSMARCO-Conversational-Search



(a) Different Rules **(b) Conversational Depth**

Figure 1: Performances in Different Scenarios. X-axis in (b) shows turn depths and Y-axis is NDCG@3.

`Oracle` uses the ground truth query rewrites provided by CAsT. This is the oracle run and falls in the manual category of CAsT [1].

We also include three automatic runs from CAsT: `clacBase`, an expert query reformulation system, `pgbert`, a GPT-2 rewriter with external manual labels, and `CFDA_CLIP_RUN7`, a BERT based query expansion system. The last two systems achieve the highest ranking accuracy among all automatic runs in CAsT 2019 [1].

Implementation Details. The query rewriter is initialized using the pre-trained GPT-2 (medium) in Pytorch-Transformers.

In the *zero-shot* setting, only the weak supervision data of the converted MARCO sessions are used to fine-tune GPT-2. We include for comparison two `Raw` baselines and our `Rule-Based` method.

In the *few-shot* setting, we also fine-tune on manual rewrites via five-fold cross validation (CV). We split the folds by sessions and *no testing fold is revealed to model training*. Our methods in this setting include `Rule-Based + CV w/o PLM`), `Self-Learn`, `CV`, `Rule-Based + CV`, and `Self-Learn + CV`. We refer readers to our code repo for details.

Our GPT-2 uses batch size 2, learning rate 5e-5, and max sequence length 150. Fine-tuning on weak supervision data converges after one epoch. Cross validation runs until convergence.

The ad hoc ranking uses Anserini BM25 with INQUERY stopword removal. The BERT ranker fine tunes BERT (base) *only* using MS MARCO passage ranking labels; the CAsT relevance labels are only used in testing; our results are directly comparable with CAsT runs.

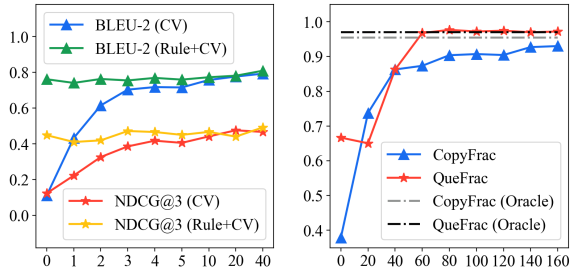
5 EVALUATION RESULTS

This section evaluates the effectiveness of our query rewriter in conversational search and analyzes the behavior of GPT-2.

5.1 Conversational Search Accuracy

The overall Results in TREC CAsT are presented in Table 2. As expected, the concise and contextual dependent nature of conversational search challenges existing ad hoc ranking and coreference resolution systems: There is a significant gap between `Original` or `AllenNLP Coref` and manual `Oracle` queries. However, the gap is substantially narrowed by our GPT-2 query rewriter.

In the few-shot setting, GPT-2 trained with `CV` already outperforms the best CAsT auto runs, `pgbert` and `CFDA`. Together with weak supervision data, `Rule-Based + CV` or `Self-Learn + CV`



(a) Training Sessions **(b) Training Steps**

Figure 2: Performances of GPT-2 with different fine-tuning amounts: conversational sessions with manual rewrites (a) and fine-tuning steps (b). The Y-axes show the corresponding metric in (a) and (b).

improves the state-of-the-art by 10+%. The improvement is mainly attributed to better query rewriting: our simple BERT (base) ranker, when using `Oracle` queries, is less effective than `pgbert` and `CFDA` teams' manual runs; they obtained 0.57+ NDCG@3, compared to ours 0.544 [1]. The BLEU scores correlate well with NDCG—better query rewriting leads to better search accuracy. Our query rewriter also maintains a stable accuracy in later turns, as shown in Fig. 1b, which indicates that our rewriter effectively captures the multi-turn context as the conversation proceeds.

Surprisingly, GPT-2 (`CV`) provides effective rewrites when only cross validated on 50 CAsT sessions; `Rule-Based`, in the zero-shot setting, is on par with best TREC CAsT automatic runs (Fig. 1a shows their individual effectiveness). In comparison, directly applying (`GPT-2 Raw`) or only fine-tuning using ad hoc sessions (`MARCO Raw`) yield sub-par results. It is impressive that the pre-trained transformer can learn conversational query rewriting, a challenging task for previous techniques, in such a data efficient manner.

5.2 Few-Shot Study

This study further investigates GPT-2's capability of generalization.

How Few Shot? Fig. 2a shows GPT-2 fine-tuned with fewer sessions, with or without weak supervision. Exceptionally, GPT-2 learns to generate reasonable query rewrites with *only three conversational sessions or 30 manual labels*; it matches best CAsT auto runs with as few as 10 sessions.

What is Learned? It is unlikely that GPT-2 learns the discourse phenomena from just three sessions. They are likely to be captured in pre-training since the non-pre-trained `GPT-2` does not outperform substantially random guess, as in Table 2.

We hypothesize that GPT-2 only needs to learn the "syntax" of the rewriting task during fine-tuning: to generate questions and to replace pronouns with or add concepts mentioned in previous turns. Fig. 2b plots the fraction of questions (QueFrac) in GPT-2 (`CV`) rewrites, indicated by question words, and the percentage of new words being copied from previous queries (CopyFrac), at different fine-tuning steps. GPT-2 adapts to query rewriting very quickly with very little fine-tuning. Our effectiveness perhaps is more from properly "unleashing" the language understanding power already in the pretrained language model.

Table 3: GPT-2 Query Rewrites on CAsT Topic 31 and 64.

Q_6	What causes **throat cancer**?
Q_7	What is the first sign of it?
Q_8	Is it the same as **esophageal cancer**?
Q_9	What's the difference in their symptoms?
Oracle	What's the difference in **throat cancer and esophageal cancer's** symptoms?
Output	What's the difference between **throat cancer and esophageal cancer**?
Q_1	What are the types of **pork ribs**?
Q_2	What are baby backs?
Q_3	What are the differences with spareribs?
Q_4	What are ways to **cook** them?
Q_5	How about on the bbq?
Oracle	How **do you cook pork ribs** on the bbq?
Output	How about on the bbq?

5.3 Case Study

Table 3 provides two examples from GPT-2 (`Rule-based + CV`). We found it surprising that in the first case, GPT-2 accurately resolves the group coreference from "their" to two cancer types, with one of the two from three turns ago. The second example presents a common error made by our rewriter: it fails to add proper context perhaps because in this case it is not clear what the context the term "about" refers to. In our manual analyses, we found that GPT-2's errors are more often due to missing complete contexts than due to adding false information.

6 CONCLUSION

This work demonstrates the effectiveness of GPT-2 for conversational query rewriting. Fine-tuned using weak supervision data generated by rules or a handful of manual rewriting labels, our GPT-2 query rewriter is able to create new state-of-the-art on the TREC CAsT conversational search benchmark—outperforming previous methods including query expansion, contextual ranking, and coreference resolution, many of which use large-scale pre-trained models and deep neural networks.

ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (No. 2018YFB1004503) and the National Natural Science Foundation of China (NSFC No. 61732008, 61532010).

REFERENCES

[1] Jeff Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The Conversational Assistance Track Overview. In *TREC 2019*. NIST.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*.
[3] R. Nogueira and K. Cho. 2019. Passage Re-ranking with BERT. *ArXiv* abs/1901.04085 (2019).
[4] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
[5] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. Question Rewriting for Conversational Question Answering. *ArXiv* abs/2004.14652 (2020).