Few-Shot Generative Conversational Query Rewriting

SIGIR 2020

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

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**?

Background

Conversational query rewriting

IR-style query expansion/term reweighting (信息检索风格的查询扩展或术语重加权)

User: “What are its side effects?” “side effects” → “adverse reactions”, “risks.”

NLP-style coreference resolution（自然语言处理风格的共指消解)

- Q1：Tell me about **aspirin**.
- Q2：What are **its** side effects?

→ What are **aspirin’s** side effects?

Neural-based query rewriting (基于神经网络的重写)

- **Advantages**: Can handle coreference and ellipsis, as well as adjust word order and semantic nuances.
- **Disadvantages**: Typically requires a large amount of training data. Output may be unstable and can potentially “hallucinate” information.

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

Background

NDCG：

DCG：Discounted Cumulative Gain

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$$

其中：

- rel_i：第 i 个检索结果的“相关性打分”（比如 0/1/2/3）；
- $\log_2(i+1)$：位置折扣，排名越靠后，惩罚越大；
- k：只看前 k 个结果（比如 NDCG@3 就只看前3条）。

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

IDCG：Ideal DCG

Method

The conversational query rewriting task is to rewrite a context dependent query Qk to a fully de-contextualized query Q′k , with the help of previous queries Q<k :

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

We use GPT-2 to directly generate the query words in Q′k one by one as:

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

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

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

Method

Rule-Based

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: $\tilde{S} = \{\tilde{Q}_1, ...\tilde{Q}_i ..., \tilde{Q}_N\}$ pairs can serve as weak supervision to approximate real conversational queries S and manual query rewrites S∗ .

- 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%)

Method

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:

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

EXPERIMENT

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.

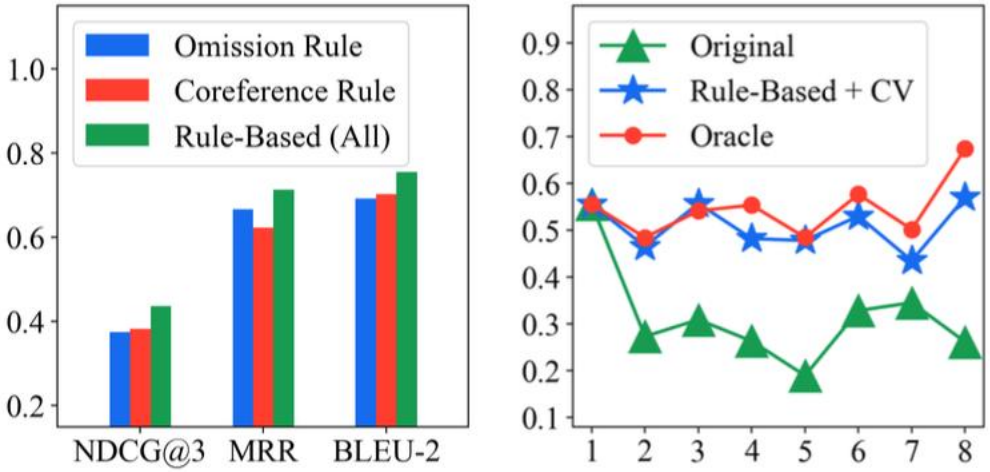
Method	BLEU-2	NDCG@3	QA-ROUGE
TREC CAsT Auto Runs			
clacBase*	–	0.360	–
pgbert*	–	0.413	–
CFDA_CLIP_RUN7*	–	**0.436**	–
CAsT Queries			
Original	0.659	0.304	0.231
AllenNLP Coref w/o sw	–	0.314	–
AllenNLP Coref w/ sw	0.750	0.437	0.278
Oracle	1.000	0.544	0.314
Zero-Shot Rewriter			
GPT-2 Raw	0.112	0.124	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**

BLEU-2	衡量生成的查询与人工重写之间的语言相似度
NDCG@3	排名指标，反映搜索系统前 3 条结果的相关性质量
QA-ROUGE	基于问答的 ROUGE 分数，评估生成查询在问答任务中的语言覆盖性

TREC CAsT Auto Runs	TREC 2019 官方比赛中的自动系统结果
Original	用户原始的会话查询（未重写）
AllenNLP Coref w/o sw	用 AllenNLP 做共指消解，不加 stopword
AllenNLP Coref w/ sw	加 stopword 的共指消解版本
Oracle	人工标注的理想重写版本（上限）
GPT-2 Raw	直接使用预训练 GPT-2，未做任何任务适配
MARCO Raw	将 GPT-2 MS MARCO 进行微调
Rule-Based	使用规则（省略+代指）自动构造改写
Rule-Based + CV w/o PLM	使用规则构造 + CV 微调，不使用语言模型
Self-Learn	用自监督方式训练出来的重写器（少量人工示例）
CV	使用人工数据做交叉验证微调的 GPT-2 模型
Rule-Based + CV	使用规则构造训练集 + 五折交叉验证微调
Self-Learn + CV	使用自监督合成数据 +五折交叉验证微调

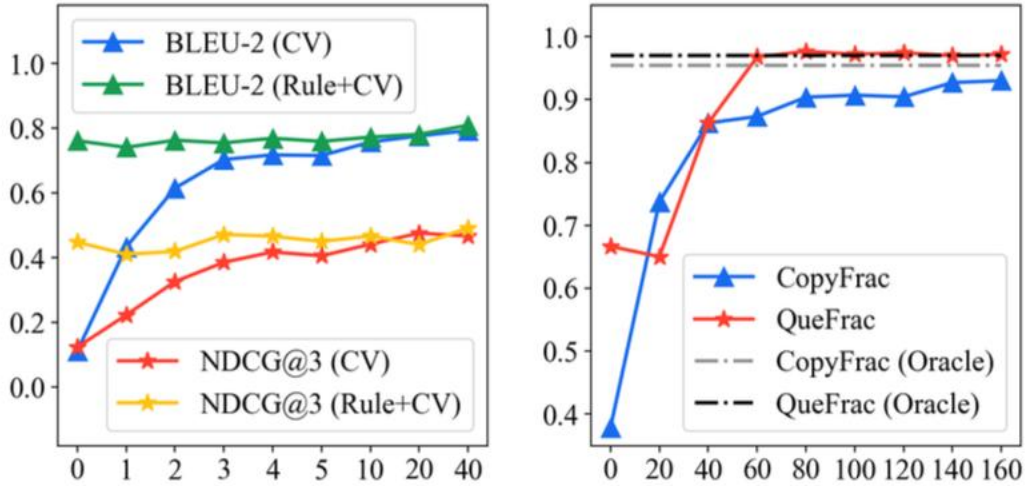
EXPERIMENT

- **QueFrac（Question Fraction）**：输出句中以疑问词（如 what、how）开头的比例；
- **CopyFrac（Copy Fraction）**：有多少新生成的词是从前一轮查询中“复制”来的。



(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.



(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).

GPT-2 学到了什么？（What is Learned?）

我们认为 GPT-2 不太可能只通过三轮会话就学会了复杂的对话结构现象（如代指、省略等）。

这些知识很可能**已经在预训练过程中被学到了**，因为未经过预训练的 GPT-2 表现在表 2 中显示接近随机猜测。

所以我们推测，GPT-2 在微调阶段只需要学习“**任务语法**”：

- 要生成的是 **问句（而不是一般文本）**；
- 要**用前文中提到的实体**替代代词或补充缺失信息。

EXPERIMENT

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?

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