Open-Domain Question Answering Goes Conversational via Question Rewriting

Raviteja Anantha[★♦], **Svitlana Vakulenko**[★♠♥], **Zhucheng Tu**[♦], **Shayne Longpre**[♦], **Stephen Pulman**[♦], **Srinivas Chappidi**[♦]

[♦]Apple Inc.

[♠]University of Amsterdam, Amsterdam, the Netherlands

Abstract

We introduce a new dataset for **Q**uestion **Re**writing in **C**onversational **C**ontext (QReCC), which contains 14K conversations with 80K question-answer pairs. The task in QReCC is to find answers to conversational questions within a collection of 10M web pages (split into 54M passages). Answers to questions in the same conversation may be distributed across several web pages. QReCC provides annotations that allow us to train and evaluate individual subtasks of question rewriting, passage retrieval and reading comprehension required for the end-to-end conversational question answering (QA) task. We report the effectiveness of a strong baseline approach that combines the state-of-the-art model for question rewriting, and competitive models for open-domain QA. Our results set the first baseline for the QReCC dataset with F1 of 19.10, compared to the human upper bound of 75.45, indicating the difficulty of the setup and a large room for improvement.

1 Introduction

It is often not possible to address a complex information need with a single question. Consequently, there is a clear need to extend open-domain question answering (QA) to a conversational setting. This task is commonly referred to as conversational (interactive or sequential) QA (Webb, 2006; Saeidi et al., 2018; Reddy et al., 2019). Conversational QA requests an answer conditioned on both the question and the previous conversation turns as context. Previously proposed large-scale benchmarks for conversational QA, such as QuAC and CoQA, limit the topic of conversation to the content of a single document. In practice, however, the answers can be distributed across several documents that are relevant to the conversation, or the topic of the conversation may also drift. To investigate this phenomena and develop approaches suitable for the complexities of this task, we introduce a new dataset for open-domain conversational QA, called QReCC.[1] The dataset consists of 13.6K conversations with an average of 6 turns per conversation.

Question: Tell me about the benefits of **Yoga**?
Answer: Increased flexibility, muscle strength…
URL: https://osteopathic.org/what-is-osteopathic-medicine/benefits-of-yoga

Question: Does **it** help in reducing stress?
Rewrite: Does **Yoga** help in reducing stress?
Answer: Yoga may help reduce stress, lower blood pressure, and lower your heart rate.
URL: https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/yoga/art-20044733

Question: What are some of the main types?
Rewrite: What are some of the main types **of Yoga**?
Answer: Hatha, Kundalini, Ashtanga, …
URL: https://www.mindbodygreen.com/articles/the-11-major-types-of-yoga-explained-simply

Question: What are common poses in Kundalini Yoga?
Rewrite: What are common poses in Kundalini Yoga?
Answer: Lotus Pose, Celibate Pose, Perfect Pose, …
URL: https://www.kundaliniyoga.org/Asanas

Figure 1: A snippet of a sample conversation from QReCC with question rewrites and answer provenance links. **Orange** indicates coreference cases where the highlighted token should be replaced with its antecedent (in **bold**). **Blue** indicates the tokens that should be generated to make the question unambiguous outside of the conversational context.

A conversation in QReCC consists of a sequence of question-answer pairs. The answers to questions were produced by human annotators, who looked up relevant information on the web using a search engine. QReCC is therefore the first large-scale dataset for conversational QA that incorporates an information retrieval subtask. QReCC is accompanied with scripts for building a collection of passages from the Common Crawl and the Wayback Machine for passage retrieval.

QReCC is inspired by the task of question rewriting (QR) that allows us to reduce the task of conversational QA to non-conversational QA by

★ Equal contribution.
♥ Work done as an intern at Apple Inc.

[1] https://github.com/apple/ml-qrecc

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 520–534
June 6–11, 2021.

generating self-contained versions of contextually-dependent questions. QR was recently shown crucial for porting retrieval QA architectures to a conversational setting (Dalton et al., 2019). Follow-up questions in conversational QA often depend on the previous conversation turns due to ellipsis (missing content) and coreference (anaphora). Every question-answer pair in QReCC is also annotated with a question rewrite. We evaluate the quality of these rewrites as self-contained questions in terms of the ability of the rewritten question, when used as input to the web search engine, to retrieve the correct answer. A snippet of a sample QReCC conversation is given in Figure 1.

The dataset collection included two phases: (1) dialogue collection, and (2) document collection. First, we set up an annotation task to collect dialogues with question-answer pairs along with question rewrites and answer provenance links. Second, after all dialogues were collected we downloaded the web pages using the provenance links, and then extended this set with a random sample of other web pages from Common Crawl, preprocessed and split the pages into passages.

To produce the first baseline, we augment an open-domain QA model with a QR component that allows us to extend it to a conversational scenario. We evaluate this approach on the QReCC dataset, reporting the end-to-end effectiveness as well as the effectiveness on the individual subtasks separately.

Our contributions. We collected the first large-scale dataset for end-to-end, open-domain conversational QA that contains question rewrites that incorporate conversational context. We present a systematic comparison of existing automatic evaluation metrics on assessing the quality of question rewrites and show the metrics that best correlate with human judgement. We show empirically that QR provides a unified and effective solution for resolving references — both co-reference and ellipsis — in multi-turn dialogue setting and positively impacts the conversational QA task. We evaluate the dataset using a baseline that incorporates the state-of-the-art model in QR and competitive models for passage retrieval and answer extraction. This dataset provides a resource for the community to develop, evaluate, and advance methods for end-to-end, open-domain conversational QA.

2 Related Work

QReCC builds upon three publicly available datasets and further extends them to the open-domain conversational QA setting: Question Answering in Context (QuAC) (Choi et al., 2018), TREC Conversational Assistant Track (CAsT) (Dalton et al., 2019) and Natural Questions (NQ) (Kwiatkowski et al., 2019). QReCC is the first large-scale dataset that supports the tasks of QR, passage retrieval, and reading comprehension (see Table 1 for the dataset comparison).

Open-domain QA. Reading comprehension (RC) approaches were recently extended to incorporate a retrieval subtask (Chen et al., 2017; Yang et al., 2019; Lee et al., 2019). This task is also referred to as machine reading at scale (Chen et al., 2017) or end-to-end QA (Yang et al., 2019). In this setup a reading comprehension component is preceded by a document retrieval component. The answer spans are extracted from documents retrieved from a document collection, given as input. The standard approach to end-to-end open-domain QA is (1) use an efficient filtering approach to reduce the number of candidate passages to the top-k of the most relevant ones (usually BM25 based on the bag-of-words representation); and then (2) re-rank the subset of the top-k relevant passages using a more fine-grained approach, such as BERT based on vector representations (Yang et al., 2019).

Conversational QA. Independently from end-to-end QA, the RC task was extended to a conversational setting, in which answer extraction is conditioned not only on the question but also on the previous conversation turns (Choi et al., 2018; Reddy et al., 2019). The first attempt at extending the task of information retrieval (IR) to a conversational setting was the recent TREC CAsT 2019 task (Dalton et al., 2019). The challenge was to rank passages from a passage collection by their relevance to an input question in the context of a conversation history. The size of the collection in CAsT 2019 was 38.4M passages, requiring efficient IR approaches. As efficient retrieval approaches operate on bag-of-words representations they need a different way to handle conversational context since they can not be trained end-to-end using a latent representation of the conversational context. A solution to this computational bottleneck was a QR model that learns to sample tokens from the conversational context as a pre-processing step before QA.

Table 1: The datasets that QReCC extends to open-domain conversational QA (QuAC, CAsT and NQ) and the datasets that are complementary to QReCC (CANARD and SaaC). RC - Reading Comprehension, PR - Passage Retrieval, QR - Question Rewriting.

Dataset	#Dialogues	#Questions	Task	Provenance
QuAC (Choi et al., 2018)	13.6K	98K	RC	-
NQ (Kwiatkowski et al., 2019)	0	307K	RC	-
CAsT (Dalton et al., 2019)	80	748	PR	-
CANARD (Elgohary et al., 2019)	5.6K	41K	QR	QuAC
OR-QuAC (Qu et al., 2020)	5.6K	41K	PR+RC	QuAC+CANARD
SaaC (Ren et al., 2020)	80	748	QR+PR+RC	CAsT
QReCC (our work)	13.7K	81K	QR+PR+RC	QuAC+NQ+CAsT

Question Rewriting. CANARD (Elgohary et al., 2019) provides rewrites for the conversational questions from the QuAC dataset. QR effectively modifies all follow-up questions such that they can be correctly interpreted outside of the conversational context as well. This extension to the conversational QA task proved especially useful while allowing retrieval models to incorporate conversational context (Voskarides et al., 2020; Vakulenko et al., 2020; Lin et al., 2020).

More recently, Qu et al. introduced OR-QuAC dataset that was automatically constructed from QuAC and CANARD datasets. OR-QuAC uses the same rewrites and answers as the ones provided in QuAC and CANARD. In contrast to OR-QuAC, the answers in QReCC are not tied to a single Wikipedia page. The answers can be distributed across several web pages. QReCC's passage collection is also larger and more diverse: 11M passages from Wikipedia in OR-QuAC vs. 54M passages from CommonCrawl in QReCC. The answers in OR-QuAC are single spans, whereas QReCC answers were produced by human annotators instructed to imitate natural conversational answers and may include several spans from different parts of the same web page.

TREC CAsT 2019 paved the way to conversational QA for retrieval but had several important limitations: (1) no training data and (2) no answer spans. First, the size of the CAsT dataset is limited to 80 dialogues, which is nowhere enough for training a machine-learning model. This was also the reason why CANARD played such an important role for the development of retrieval-based approaches even though it was collected as a RC dataset. Second, the task in TREC CAsT 2019 was conversational passage retrieval not extractive QA since the expected output was ranked passages and not a text span. We designed QReCC to overcome both of these limitations.

The size of the QReCC dataset is comparable with other large-scale conversational QA datasets (see Table 1). The most relevant to our work is the concurrent work by Ren et al., who extended the TREC CAsT dataset with crowd-sourced answer spans. Since the size of this dataset is inadequate for training a machine-learning model and can be used only for evaluation, the authors train their models on the MS MARCO dataset instead, which is a non-conversational QA dataset (Bajaj et al., 2016). Their evaluation results show how the performance degrades due to the lack of conversational training data. TREC CAsT will continue in the future and the QReCC dataset provides a valuable benchmark helping to train and evaluate novel conversational QA approaches.

3 Dialogue Collection

To simplify the data collection task we decided to use questions from pre-existing QA datasets as seeds for dialogues in QReCC. We used questions from QuAC, CAsT and NQ. While QuAC and CAsT datasets contain question sequences, NQ is not a conversational dataset but contains standalone questions from web search. We use the NQ dataset to increase and diversify the number of samples beyond QuAC and CAsT by generating more rewrites for cases beyond coreference resolution. The majority of the follow-up questions in QuAC require coreference resolution for QR. Therefore, we explicitly instructed the annotators to use NQ as a start of a conversation and then come up with relevant follow-up questions, which would require generation of missing content, i.e., ellipsis, instead of coreference resolution for QR.

The task for the annotators was also to answer questions using a web search engine. Question rewrites were used as input to a search engine. This