A Wrong Answer or a Wrong Question? An Intricate Relationship between Question Reformulation and Answer Selection in Conversational Question Answering

Svitlana Vakulenko[†]
University of Amsterdam
s.vakulenko@uva.nl

Shayne Longpre, Zhucheng Tu, Raviteja Anantha
Apple Inc.
{slongpre, zhucheng_tu, raviteja_anantha}@apple.com

Abstract

The dependency between an adequate question formulation and correct answer selection is a very intriguing but still underexplored area. In this paper, we show that question rewriting (QR) of the conversational context allows to shed more light on this phenomenon and also use it to evaluate robustness of different answer selection approaches. We introduce a simple framework that enables an automated analysis of the conversational question answering (QA) performance using question rewrites, and present the results of this analysis on the TREC CAsT and QuAC (CANARD) datasets. Our experiments uncover sensitivity to question formulation of the popular state-of-the-art models for reading comprehension and passage ranking. Our results demonstrate that the reading comprehension model is insensitive to question formulation, while the passage ranking changes dramatically with a little variation in the input question. The benefit of QR is that it allows us to pinpoint and group such cases automatically. We show how to use this methodology to verify whether QA models are really learning the task or just finding shortcuts in the dataset, and better understand the frequent types of error they make.

1 Introduction

Conversational question answering (QA) is a new and important task which allows systems to advance from answering stand-alone questions to answering a sequence of related questions (Choi et al., 2018; Reddy et al., 2019; Dalton et al., 2019). Such sequences contain questions that usually revolve around the same topic and its subtopics, which is also common for a human conversation. The most pronounced characteristics of such question sequences are anaphoric expressions and ellipsis, which make the follow-up questions ambiguous outside of the conversation context. For example, consider a question "*When was **it** discovered?*". It is not possible to answer the question without resolving the pronoun **it** (example of an anaphoric expression). Ellipsis are even harder to resolve since they omit information without leaving any references. For example, a question "*When?*" can naturally follow an answer to the previous question, such as "*Friedrich Miescher discovered DNA.*"

Question rewriting (QR) was recently introduced as an independent component for conversational QA (Elgohary et al., 2019; Vakulenko et al., 2020; Yu et al., 2020). Query rewriting received considerable attention in the information retrieval community before but not in a conversational context (He et al., 2016). Ren et al. (2018) performed similar experiments using query reformulations mined from search sessions. However, half of their samples were keyword queries rather than natural language questions and their dataset was not released to the community. In this paper, we show that QR is not only operational in extending standard QA models to the conversational scenario but can be also used for their evaluation.

An input to the QR component is a question and previous conversation turns. The QR component is designed to transform all ambiguous questions, e.g., "*When?*", into their unambiguous equivalents, e.g., "*When was DNA discovered?*". Such unambiguous questions can be then processed by any standard QA model outside of the conversation history.

Clearly, the quality of question formulation interacts with the ability to answer this question. In this paper, we show that introducing QR component in the conversational QA architecture, by decoupling question interpretation in context from the question answering task, provides us with a unique opportunity to gain an insight on the interaction between the two tasks. Ultimately, we are interested in whether this interaction can potentially help us

[†] Work done as an intern at Apple Inc.

to improve the performance on the end-to-end conversational QA task. To this end we formulate our main research question:

How do differences in question formulation in a conversational setting affect question answering performance?

The standard approach to QA evaluation is to measure model performance on a benchmark dataset with respect to the ground-truth answers, such as text span overlap in the reading comprehension task, or NDCG@3 in the passage retrieval task. However, such evaluation setups may also have their limitations with respect to the biases in the task formulation and insufficient data diversity that allow models to learn shortcuts and overfit the benchmark dataset (Geirhos et al., 2020). There is already an ample evidence of the pitfalls in the evaluation setup of the reading comprehension task highlighting that the state-of-the-art models tend to learn answering questions using superficial clues (Jia and Liang, 2017; Ribeiro et al., 2018; Lewis and Fan, 2019).

We aim to further contribute to the research area studying robustness of QA models by extending the evaluation setup to the conversational QA task. To this end, we introduce an error analysis framework based on conversational question rewrites. The goal of the framework is to evaluate robustness of QA models using the inherent properties of the conversational setup itself. More specifically, we contrast the results obtained for ambiguous and rewritten questions outside of the conversation context. We show that this data is very well suited for analysing performance and debugging QA models. In our experiments, we evaluate two popular QA architectures proposed in the context of the conversational reading comprehension (QuAC) (Choi et al., 2018) and conversational passage ranking (TREC CAsT) (Dalton et al., 2019) tasks.

Our results show that the state-of-the-art models for passage retrieval are rather sensitive to differences in question formulation. On the other hand, the models trained on the reading comprehension task tend to find correct answers even to incomplete ambiguous questions. We believe that these findings can stimulate more research in this area and help to inform future evaluation setups for the conversational QA tasks.

2 Experimental Setup

To answer our research question and illustrate the application of the proposed evaluation framework in practice, we use the same experimental setup introduced in our earlier work (Vakulenko et al., 2020). This architecture consists of two independent components: QR and QA. It was previously evaluated against competitive approaches and baselines setting the new state-of-the-art results on the TREC CAsT 2019 dataset. The QR model was also shown to improve QA performance on both passage retrieval and reading comprehension tasks.

In the following subsection, we describe the datasets, models, and metrics that were used in our evaluation. Note, however, that our error analysis framework can be applied to other models and metrics as well. The only requirement for applying the framework is the QR-QA architecture that provides two separate outputs in terms of question rewrites and answers. We show that the same framework can be applied for both reading comprehension and passage retrieval tasks, although they produce different types of answers and require different evaluation metrics.

2.1 Datasets

We chose two conversational QA datasets for the evaluation of our approach: (1) TREC CAsT for conversational passage ranking (Dalton et al., 2019), and (2) CANARD, derived from Question Answering in Context (QuAC) for conversational reading comprehension (Choi et al., 2018). Since TREC CAsT is a relatively small dataset, we used only CANARD for training our QR model. The same QR model trained on CANARD was then evaluated on both CANARD and TREC CAsT independently.

Following the setup of the TREC CAsT 2019, we use the MS MARCO passage retrieval (Nguyen et al., 2016) and the TREC CAR (Dietz et al., 2018) paragraph collections. After de-duplication, the MS MARCO collection contains 8.6M documents and the TREC CAR – 29.8M documents. The model for passage retrieval is tuned on a sample from the MS MARCO passage retrieval dataset, which includes relevance judgements for 12.8M query-passage pairs with 399k unique queries (Nogueira and Cho, 2019). We evaluated on the test set with relevance judgements for 173 questions across 20 dialogues (topics).

We use CANARD (Elgohary et al., 2019) and

QuAC (Choi et al., 2018) datasets jointly to analyse performance on the reading comprehension task. CANARD is built upon the QuAC dataset by employing human annotators to rewrite original questions from QuAC dialogues into explicit questions. CANARD contains 40.5k pairs of question rewrites that can be matched to the original answers in QuAC. We use CANARD splits for training and evaluation. Each answer in QuAC is annotated with a Wikipedia passage from which it was extracted alongside the correct answer spans within this passage. We use the question rewrites provided in CANARD and passages with answer spans from QuAC. In our experiments, we refer to this joint dataset as CANARD for brevity. Our model for reading comprehension was also pre-trained using MultiQA dataset (Fisch et al., 2019) to further boost its performance. MultiQA contains 75k QA pairs from six standard QA benchmarks.

2.2 Conversational QA Architecture

Our architecture for conversational QA is designed to be modular by separating the original task into two subtasks. The subtasks are (1) QR, responsible for the conversational context understanding and question formulation, and (2) QA that exactly corresponds to the standard non-conversational task of answer selection (reading comprehension or passage retrieval). Therefore, the output of the QR component is an unambiguous question that is subsequently used as input to the QA component to produce the answer.

Question Rewriting The task of question rewriting is to reformulate every follow-up question in a conversation, such that the question can be unambiguously interpreted without accessing the conversation history. The input to the QR model is the question with previous conversation turns separated with a $[SEP]$ token (in our experiments, we use up to maximum of 5 previous conversation turns). Using our running example, the input would be: *Friedrich Miescher discovered DNA* $[SEP]$ *When?* and the expected output from the QR model is: *When was DNA discovered?*.

Our QR model is based on a unidirectional Transformer (decoder) designed for the sequence generation task. It was initialised with the weights of the pre-trained GPT2 (Radford et al., 2019) and further fine-tuned on the QR task. The training objective in QR is to predict the output sequence as in the ground truth question rewrites produced by human annotators. The model is trained via the teacher forcing approach. The loss is calculated with the negative log-likelihood (cross-entropy) function. At inference time, the question rewrites are generated recursively turn by turn for each of the dialogues using the previously generated rewrites as input corresponding to the dialogue history, i.e., previous turns.

For our experiments, we adopt the same QR model architecture proposed in the previous work (Transformer++) (Anantha et al., 2020; Vakulenko et al., 2020). It was shown to outperform a co-reference baseline and other Transformer-based models on CANARD, TREC CAsT and QReCC.

Question Answering We experiment with two different QA models that reflect the state-of-the-art in reading comprehension and passage retrieval (Nogueira and Cho, 2019). All our QA models are initialised with a pre-trained $BERT_{LARGE}$ (Devlin et al., 2019) and then fine-tuned on each of the target tasks.

Our model for reading comprehension follows the standard architecture design for this task. We restrict our implementation to the simplest but a very competitive model architecture that more complex approaches usually build upon (Liu et al., 2019; Lan et al., 2019). This model consists of a Transformer-based bidirectional encoder and an output layer that predicts the answer span. The input to the model is the sequence of tokens formed by concatenating a question and a passage, the two are separated with a $[SEP]$ token.

Our passage retrieval approach is implemented following Nogueira and Cho (2019). It uses Anserini for the candidate selection phase with BM25 (top-1000 passages) and $BERT_{LARGE}$ for the passage re-ranking phase.[1] The re-ranking model was tuned on a sample from MS MARCO with 12.8M query-passage pairs and 399k unique queries.

2.3 Metrics

We use the standard performance metrics for each of the QA subtasks. We use normalized discounted cumulative gain (*NDCG@3*) and precision on the top-passage (*P@1*) to evaluate quality of passage retrieval (with a relevance threshold of 2 in accordance with the official evaluation guidelines for CAsT). We use *F1* metric for reading comprehen-

[1] https://github.com/nyu-dl/dl4marco-bert