Can You Unpack That?
Learning to Rewrite Questions-in-Context

Ahmed Elgohary, Denis Peskov, Jordan Boyd-Graber*
Department of Computer Science, UMIACS, iSchool, Language Science Center
University of Maryland, College Park
{elgohary, dpeskov, jbg}@cs.umd.edu

Abstract

Question answering is an AI-complete problem, but existing datasets lack key elements of language understanding such as coreference and ellipsis resolution. We consider sequential question answering: multiple questions are asked one-by-one in a conversation between a questioner and an answerer. Answering these questions is only possible through understanding the conversation history. We introduce the task of question-in-context rewriting: given the context of a conversation's history, rewrite a context-dependent into a self-contained question with the same answer. We construct, CANARD, a dataset of 40,527 questions based on QuAC (Choi et al., 2018) and train Seq2Seq models for incorporating context into standalone questions.

1 Introduction

Question Answering (QA) is an AI complete problem (Webber, 1992), but existing QA datasets do not rise to the challenge: they lack key NLP problems like anaphora resolution, coreference disambiguation, and ellipsis resolution. The logic needed to answer these types of questions requires deeper NLP understanding that simulates the context in which humans naturally answer questions.

Neural techniques question answering have improved (Devlin et al., 2018) machine reading comprehension (Rajpurkar et al., 2016, MRC): computers can take a single question and extract answers from datasets like Wikipedia. However, QA models struggle to generalize when questions do not look like the standalone questions systems in training data: e.g., new genres, languages, or closely-related tasks (Yogatama et al., 2019).

Conversational question answering (Reddy et al., 2019, CQA) is a generalization that ask *multiple* questions in an information-seeking dialogs. Unlike MRC, CQA requires models to link questions together to resolve the conversational dependencies between them: each question needs to be understood in the conversation context. For example, the question *"What was he like in that episode?"* cannot be understood without knowing what *"he"* and *"that episode"* refer to, which can be resolved using the conversation context.

We reduce challenging, interconnected CQA examples to independent, stand-alone MRC to create CANARD—**C**ontext **A**bstraction: **N**ecessary **A**dditional **R**ewritten **D**iscourse—a new dataset[1] that rewrites QuAC (Choi et al., 2018) questions. We crowdsource context-independent paraphrases of QuAC questions and use the paraphrases to train and evaluate question-in-context rewriting.

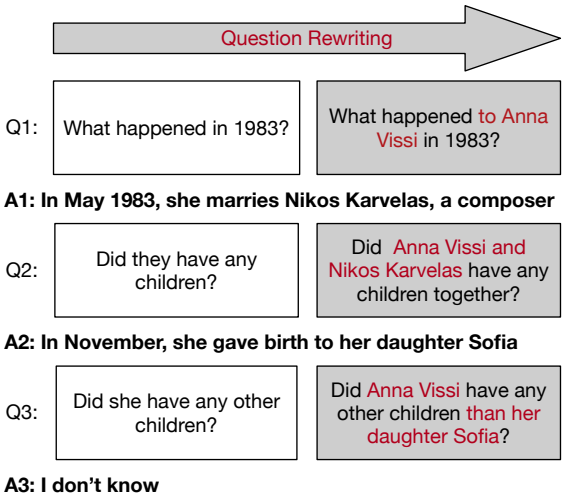

Figure 1: Question-in-context rewriting task. The input to each step is a question to rewrite given the dialog history which consists of the dialog utterances (questions and answers) produced before the given question is asked. The output is an equivalent, context-independent paraphrase of the input question.

* Now at Google AI Zürich

[1] http://canard.qanta.org

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5918–5924,
Hong Kong, China, November 3–7, 2019.

Characteristic	Ratio
Answer Not Referenced	0.98
Question Meaning Unchanged	0.95
Correct Coreferences	1.0
Grammatical English	1.0
Understandable w/o Context	0.90

Table 1: Manual inspection of 50 rewritten context-independent questions from CANARD suggests that the new questions have enough context to be independently understandable.

Section 2 formally defines the task of question de-contextualization. Section 3 constructs CANARD, a new dataset of question-in-context with corresponding context-independent paraphrases. Section 5 analyzes our rewrites (and the underlying methodology) to understand the linguistic phenomena that make CQA difficult. We build several baseline rewriting models and compare their BLEU scores to our human rewrites in Section 4.

2 Defining Question-In-Context Rewrites

We formally define the task of question-in-context rewriting (de-contextualization). Given a conversation topic t and a history H of $m-1$ turns, each turn k is a question q_i and an answer a_i; the task is to generate a rewrite q'_m for the next question q_m based on H. Since q_m is part of the conversation, its meaning often involves references to parts of its preceding history. A valid rewrite q'_m should be self-contained: a correct answer to q'_m by itself is a correct answer to q_m combined with the question's preceding history H.

Figure 1 shows the assumptions of CQA and how they are made explicit in rewrites. The first question omits the title of the page (Anna Vissi), the second question omits the answer to the first question (replacing both Anna Vissi and her husband with the pronoun "they"), and the last question adds a scalar implicature that must be resolved.

3 Dataset Construction

We elicit paraphrases from human crowdworkers to make previously context-dependent questions *unambiguously* answerable. Through this process, we resolve difficult coreference linkages and create a pair-wise mapping between ambiguous and context-enriched questions. We derive CANARD from QuAC (Choi et al., 2018), a sequential question answering dataset about specific Wikipedia sections. QuAC uses a pair of workers—a "student" and a "teacher"—to ask and respond to questions. The "student" asks questions about a topic based on only the title of the Wikpedia article and the title of the target section. The "teacher" has access to the full Wikipedia section and provides answers by selecting text that answers the question. With this methodology, QuAC gathers 98k questions across 13,594 conversations. We take their entire dev set and a sample of their train set and create a custom JavaScript task in Mechanical Turk that allows workers to rewrite these questions. JavaScript hints help train the users and provides automated, real-time feedback.

We provide workers with a comprehensive set of instructions and task examples. We ask them to rewrite the questions in natural sounding English while preserving the sentence structure of the original question. We discourage workers from introducing new words that are unmentioned in the previous utterances and ask them to copy phrases when appropriate from the original question. These instructions ensure that the rewrites only resolve conversation-dependent ambiguities. Thus, we encourage workers to create minimal edits; in Section 5.2, we take advantage of this to use BLEU for evaluating model-generated rewrites.

We display the questions in the conversation one at a time, since the rewrites should include only the previous utterance. After a rewrite to the question is submitted, the answer to the question is displayed. The next question is then displayed. This repeats until the end of the conversation. The full set of instructions and the data collection interface are provided in the appendix.

We apply quality control throughout our collection process. During the task, JavaScript checks automatically monitor and warn about common errors: submissions that are abnormally short (e.g., 'why'), rewrites that still have pronouns (e.g., 'he wrote this album'), or ambiguous words (e.g., 'this article', 'that'). Many QuAC questions ask about 'what/who else' or ask for 'other' or 'another' entity. For that class of questions, we ask workers to use a phrase such as 'other than', 'in addition to', 'aside from', 'besides', 'together with' or 'along with' with the appropriate context in their rewrite.

We gather and review our data in batches to screen potentially compromised data or low quality workers. A post-processing script flags suspicious

ORIGINAL: Was this an honest mistake by the media?
REWRITE: Was the claim of media regarding Leblanc's room come to true?
ORIGINAL: What was a single from their album?
REWRITE: What was a single from horslips' *album*?
ORIGINAL: Did they marry?
REWRITE: Did Hannah Arendt and Heidegger marry?

Table 2: Not all rewrites correctly encode the context required to answer a question. We take two failures to provide examples of the two common issues: Changed Meaning (top) and Needs Context (middle). We provide an example with no issues (bottom) for comparison.

rewrites and workers who take and abnormally long or short time. We flag about 15% of our data. *Every* flagged question is manually reviewed by one of the authors and an entire HIT is discarded if one is deemed inadequate. We reject 19.9% of submissions and the rest comprise CANARD. Additionally, we filter out under-performing workers based on these rejections from subsequent batches. To minimize risk, we limit the initial pool of workers to those that have completed 500 HITs with over 90% accuracy and offer competitive payment of $0.50 per HIT.

We verify the efficacy of our quality control through manual review. A random sample of fifty questions sampled from the final dataset is reviewed for desirable characteristics by a native English speaker in Table 1. Each of the positive traits occurs in 90% or more of the questions. Based on our sample, our edits retain grammaticality, leave the question meaning unchanged, and use pronouns unambiguously. There are rare occasions where workers use a part of the answer to the question being rewritten or where some of the context is left ambiguous. These infrequent mistakes should not affect our models. We provide examples of failures in Table 2.

We use the rewrites of QuAC's development set as our test set (5,571 question-in-context and corresponding rewrite pairs) and use a 10% sample of QuAC's training set rewrites as our development set (3,418); the rest are training data (31,538).

	Dev	Test
Copy	33.84	36.25
Pronoun Sub	47.72	47.44
Seq2Seq	51.37	49.67
Human Rewrites*	59.92	

Table 3: BLEU scores of the baseline models on development and test data. Seq2Seq improves up to four points over naive baselines but still well below human accuracy. Human accuracy (*) is computed from a small subset of the validation set.

4 Baselines

We compare three baseline models for the question-in-context rewriting task. In the **Copy** baseline, the rewrite q'_m is set to be the same as the input question q_m without making any changes.

We also try a **Pronoun Substitution** baseline in which the first pronoun in q_m is replaced with the topic entity of the conversation. We use the title of the corresponding Wikipedia article to the original QuAC conversation as the topic entity. Similar to the Copy baseline, the training data is not used in that baseline.

Unlike the previous baselines which do not use our rewrites as training data, the third baseline is a neural sequence-to-sequence (**Seq2Seq**) model with attention and a copy mechanism (Bahdanau et al., 2015; See et al., 2017). We construct the input sequence by concatenating all utterances in the history H, prepending them to q_m, and adding a special separator token between utterances. We use a bidirectional LSTM encoder-decoder model with shared the word embeddings between the encoder and the decoder.[2]

Since questions are written by humans, a human rewrites are the upper-bound for this task. However, annotators (especially crowdworkers) can be inconsistent or disagree. To estimate the human accuracy, we collect 100 pairs of rewritten questions; each pair has two rewrites of the same question (in its given context) by two different workers. We manually verify that all rewrites are valid and then use the pair of rewrites as a hypothesis and a reference.

Table 3 shows the BLEU scores produced by the baselines and humans over both the validation and the test sets.[3] Although a well-trained standard

[2]We initialize the embeddings with GloVE (Pennington et al., 2014) and train with a batch-size of 16 for 200000 steps. We use OpenNMT (Klein et al., 2018) implementation.

[3]We use multi-bleu-detok.perl (Sennrich et al., 2017)