Reinforced Question Rewriting for Conversational Question Answering

Zhiyu Chen, Jie Zhao, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, Shervin Malmasi
Amazon.com, Inc., Seattle, WA, USA
{zhiyuche,zhaozjie,njfn,besnikf,olegro,malmasi}@amazon.com

Abstract

Conversational Question Answering (CQA) aims to answer questions contained within dialogues, which are not easily interpretable without context. Developing a model to rewrite conversational questions into self-contained ones is an emerging solution in industry settings as it allows using existing single-turn QA systems to avoid training a CQA model from scratch. Previous work trains rewriting models using human rewrites as supervision. However, such objectives are disconnected with QA models and therefore more human-like rewrites do not guarantee better QA performance.

In this paper we propose using QA feedback to supervise the rewriting model with reinforcement learning. Experiments show that our approach can effectively improve QA performance over baselines for both extractive and retrieval QA. Furthermore, human evaluation shows that our method can generate more accurate and detailed rewrites when compared to human annotations.

1 Introduction

Interacting through conversations is a natural information-seeking procedure for humans, therefore it is important for AI assistants like Apple Siri and Amazon Alexa to enable and improve such experiences. In recent years Conversational Question Answering (CQA) has gained more attention, where a user can ask a series of related questions and ideally obtain answers that leverage the conversational context. Different from widely-studied question answering (QA) tasks that happen in single-turn (Rajpurkar et al., 2016, 2018; Tay et al., 2018; Tang et al., 2019), the interpretation of conversational questions in CQA depends on questions and answers from previous turns.

Previous approaches to CQA usually train new models from scratch, which can achieve promising results but also are expensive in terms of obtaining domain-specific training data. In industry settings, there are many single-turn QA models deployed. Training new CQA models with additional annotations to replace each existing single-turn QA model is expensive, and generally not feasible. Moreover, discarding existing single-turn models and datasets is impractical, and studying how to reuse these existing resources to tackle CQA merits attention.

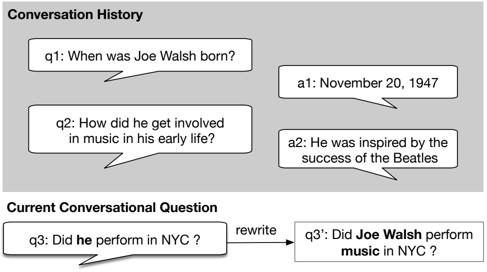

Conversation History
q1: When was Joe Walsh born?
a1: November 20, 1947
q2: How did he get involved in music in his early life?
a2: He was inspired by the success of the Beatles
Current Conversational Question
q3: Did **he** perform in NYC ?
rewrite
q3': Did **Joe Walsh** perform **music** in NYC ?

Figure 1: A conversational question rewriting example.

Existing approaches to this task, called Conversational Question Rewriting (CQR), often train sequence-to-sequence models supervised by human rewrites to generate self-contained questions (Ren et al., 2018; Vakulenko et al., 2021). Such methods have several limitations. First, the CQR training objective is disconnected from CQA performance. The annotation process of existing rewriting datasets has no knowledge of the QA systems, and more human-like rewrites do not guarantee better CQA performance. Second, the rewriting model does not take into account the feedback from downstream QA systems. In industry settings, multiple single-turn QA systems trained with different datasets serve in the backend. It is impractical to replace them with new CQA models, and we argue that their output can still be used as signals to help train rewriting models.

To overcome these limitations, we propose an effective CQR approach upon the recent success of Reinforcement Learning (RL) techniques for text generation (Rennie et al., 2017). RL enables flexible ways to incorporate training objectives in the

Proceedings of EMNLP 2022 Industry Track, pages 367–380
December 9–11, 2020. ©2022 Association for Computational Linguistics

form of reward functions. We systematically analyze different rewards and their effectiveness in terms of final QA performance, as well as the quality of the question rewrites (i.e. the question still has to be understandable and interpretable by humans). To optimize QA performance, we propose various QA rewards to measure the likelihood of a question yielding a better answer. In comparison with the QA rewards, we also propose to use the same RL approach with question rewriting (QR) rewards reflecting the similarity between a model-generated question and the human's ground-truth.

We summarize our contributions as follows:

- To the best of our knowledge, we are the first to study how to incorporate QA signals to improve CQR using RL.
- We systematically propose and compare using different training signals as rewards.
- We conduct experiments on two CQA tasks to show our approach is effective.
- A user study shows that our method can generate more accurate and detailed rewrites when compared to human annotations.

2 Related Work

Conversational Question Answering. Recently, conversational QA has been studied which presents new challenges for QA models such as being able to resolve conversational dependencies so that a conversational question can be interpreted by QA models with conversational context. QuAC (Elgohary et al., 2019) and CoQA (Reddy et al., 2019) are two datasets for extractive CQA where answers are extracted from passages. CAsT-19 (Dalton et al., 2020) is a benchmark for retrieval CQA and the target is to return relevant passages given a question. QReCC (Anantha et al., 2021) combines retrieval and extractive CQA where the answers are extracted from passages returned by a retrieval component. Kim et al. (2021) propose to train the CQA model and rewriter simultaneously, which is impractical for industry setting. A directly related work to ours is Vakulenko et al. (2021) which proposes to rewrite questions for CQA. However, they do not consider taking the QA feedback into the CQR training which is studied in our work.

RL for Nature Language Generation. Reinforcement learning methods have been widely applied for various language generation tasks. Li et al. (2016) propose to apply deep reinforcement learning in dialogue generation to model future rewards related to conversational properties, such as informativeness, coherence and ease of answering. Ranzato et al. (2016) propose Mixed Incremental Cross-Entropy Reinforce (MIXER) for sequence prediction to directly optimize the metrics used at test time, such as BLEU or ROUGE. They show MIXER outperforms several strong baselines for greedy generation on text summarization, image caption and machine translation. Nogueira and Cho (2017) train a query rewriter based on the rewards relying on the ground-truth ranking list for information retrieval. Buck et al. (2018) use RL for single-turn question rewriting by maximizing the answers' quality which requires ground-truth. Similar to our F1 reward, Wu et al. (2021) design rewards from ground-truth answers to train a conversational query rewriter. Instead, we propose alternative rewards indicating the confidence of answers from a model itself which do not require human annotations.

3 Problem Definition

In CQA, each conversation contains a sequence of (question, answer) pairs $D = \{q_1, a_1, ..., q_n, a_n\}$, where a_i is the answer for question q_i. A conversational question q_i can be ambiguous and its interpretation depends on the conversational context $c_i = \{q_1, a_1, ..., q_{i-1}, a_{i-1}\}$. The goal of CQR for QA is to learn a model $\mathcal{R}_\theta$, parameterized by θ, that can translate q_i associated with c_i into q'_i, so that the semantic meaning of q'_i is equivalent to q_i.

$$q'_i = \mathcal{R}_\theta(q_i, c_i) \tag{1}$$

A pretrained single-turn QA model is expected to answer q'_i better than q_i. Note that the QA model can be trained from a single-turn dataset different from D and is fixed when training the rewriter. The motivation is to explore whether the already deployed single-turn QA models can be exploited to train a rewriter and reused without further training by accepting the rewritten questions.

4 Approach

4.1 Model Overview

We show our CQR approach with a modularized design in Figure 2. There are two major components: a CQR model $\mathcal{R}_\theta$ as introduced in Section 3 and a reward function $\mathcal{F}$ that evaluates rewrite q'_i generated by $\mathcal{R}_\theta$ by producing a reward score. Then the CQR training can be formulated as a reinforcement training problem, where the objective

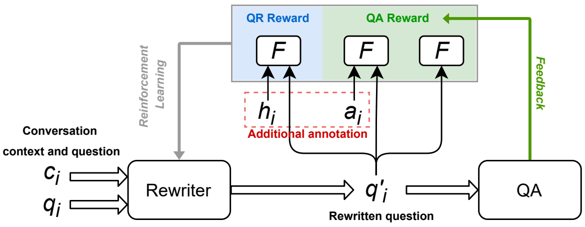


Figure 2: Overview of our CQR approach. h_i is human rewriting of q_i and a_i is the ground-truth answer of q_i.

is to maximize an expected reward or equivalently minimize the following loss function:

$$\mathcal{L}_{rl}(\theta) = -\mathbb{E}_{q'_i \sim \mathcal{R}_\theta(q_i, c_i),\, q_i \sim \mathcal{T}}(\mathcal{F}(q'_i)) \;, \tag{2}$$

where q_i comes from data distribution $\mathcal{T}$. During training, we push $\mathcal{R}_\theta$ to generate q'_i that achieves a higher reward by minimizing Equation 2. Hereinafter, we omit θ from $\mathcal{R}_\theta$ for simplicity.

We define two types of rewards: QR rewards evaluate how similar a question rewrite is to the ground truth one produced by human annotators; QA rewards evaluate how well a QA model can answer a question rewrite. We summarize the characteristics of different rewards in Table 1. By maximizing one of the QR or QA rewards, we can explicitly optimize the model to achieve the QR or QA target. Next, we describe the two types of rewards.

Reward	ROUGE	F1	Confidence	BM25
Reward Type	QR	QA	QA	QA
CQA Type	-	Extractive	Extractive	Retrieval
Need Annotated Rewrites	Y	N	N	N
Need Annotated Answers	N	Y	N	N

Table 1: Characteristics of different rewards.

4.2 QR Rewards

The rationale of maximizing QR rewards is similar to the aims of prior work: a good question rewrite should be similar to a human rewrite. We use the ROUGE-L score (Lin, 2004) between the question rewrite q'_i and the ground-truth h_i as the QR reward:

$$\mathcal{F}(q'_i, h_i) = ROUGE_L(q'_i, h_i) \tag{3}$$

This reward has been widely used by RL methods for language generation tasks. Note that Eq. 3 does not depend on the QA model and prior work can be considered as maximizing QR rewards.

4.3 QA Rewards

We define QA rewards that reflect how well the question rewrites can help a QA model obtain better answers. Since QA rewards are task/model-dependent, we introduce QA rewards for the following two sub-types.

4.3.1 Extractive CQA

Extractive CQA is a machine reading comprehension (MRC) task and an extractive QA model $\mathcal{M}$ extracts the most likely span answer given a question q and an evidence document p:

$$a_s = \arg\max_{a_s} P_{\mathcal{M}}(a_s | q, p) \tag{4}$$

We assume that $\mathcal{M}$ is trained on regular single-turn QA data, and expects the input question q to be self-contained. Therefore, CQA questions should be rewritten by $\mathcal{R}$ before being sent to $\mathcal{M}$. Next, we introduce supervised and unsupervised QA rewards.

Supervised QA rewards. A straightforward way to measure the quality of a question rewrite q'_i in terms of QA is to calculate the similarity between the predicted answer by $\mathcal{M}$ with q'_i as input and the ground-truth answer a_i. We denote a'_s as the extracted answer span by $\mathcal{M}$ using the rewritten question q'_i as input. We measure the overlap between a'_s and a_i by F1 score:

$$\mathcal{F}(q'_i, a_i) = F1(\arg\max_{a'_s} P_{\mathcal{M}}(a'_s | q'_i, p), a_i) \tag{5}$$

Intuitively, the rewrite q'_i is better if a'_s is closer to the ground-truth answer. Compared with Equation 3, Equation 5 depends on the ground-truth answers instead of human rewrites.

Unsupervised QA rewards. For a predicted span a'_s, $\mathcal{M}$ assigns a probability $r_c = P_{\mathcal{M}}(a'_s | q'_i, p)$ that reflects the model's confidence about the answer. We assume that a higher confidence score of an answer indicates that the QA model has a better question understanding. Therefore, we directly use the probability of the most likely answer as the confidence reward for a question rewrite:

$$\mathcal{F}(q'_i) = \max P_{\mathcal{M}}(a'_s | q'_i, p) \tag{6}$$

F1 rewards can be considered as judgment scores on predicted answers by humans since the ground-truth answers are used, while confidence rewards represent the model's self-judgments.

4.3.2 Retrieval CQA

We also evaluate our method's generalization on a different **retrieval CQA** task, where the goal is to return a list of documents in descending order of relevance scores produced by a retrieval CQA model:

$$rel = \mathcal{M}(q, p) \tag{7}$$

where p is a document. A retrieval CQA model usually consists of two stages. In the first stage, a lightweight ranking algorithm such as BM25 is used to retrieve top-k candidate documents. In the second stage, a more complex model such as BERT (Devlin et al., 2019) is used to rerank candidate documents. Here, we use the BM25 score between a question and a document, which is a type of QA reward that does not use annotated answers:

$$\mathcal{F}(q'_i) = BM25(q'_i, p) \tag{8}$$

We expect the rewrite q'_i can retrieve documents with higher BM25 scores in the first stage than q_i so that the performance in the re-ranking stage can also be improved.

4.4 Training

There are two steps in our training framework. The first step, the pre-training step, which has the same supervised target as prior work. The objective is to minimize the cross-entropy loss between the model's prediction q' and human ground-truth rewrites h:

$$\mathcal{L}_{sup} = -y_h \log y_{q'} \ , \tag{9}$$

where y_h is the one-hot vector of h and $y_{q'}$ is the distribution over tokens in q' predicted by the model. Supervised pre-training ensures the model has the basic ability to rewrite the original question given the conversational context.

The second step continues training $\mathcal{R}$ with RL to maximize different rewards. In this work, we use Self-Critical Sequence Training (SCST) (Rennie et al., 2017). Given a question q, we generate two question rewrites q^s and q'. q^s is generated by sampling the word distribution from $\mathcal{R}$ at each step, and q' is generated by $\mathcal{R}$ using greedy decoding. Then we minimize the following loss function:

$$\mathcal{L}_{rl} = (r' - r^s) \sum_{t=1}^{N} \log P_{\mathcal{R}}(w^s_t | w^s_{1:t-1}, q, c) \tag{10}$$

Here, $P_{\mathcal{R}}(\cdot)$, which is defined by $\mathcal{R}$, is the probability of generating the t-th word conditioning on previously generated tokens of q_s, the original question q and conversational history c. Intuitively, minimizing $\mathcal{L}_{rl}$ increases the likelihood of q^s if it obtains a higher reward than q' (i.e. $r^s > r'$), and thus maximizes the expected total reward. Given a reward function, we can obtain $r' = \mathcal{F}(q')$ ($\mathcal{F}$ can be one of Equation 3,5,6,8) and $r^s = \mathcal{F}(q^s)$.

We only choose one of the reward functions to obtain the reward for a question. We leave the combination of different rewards as future work. Additional training procedure details are described in Appendix A.

5 Data and Experimental Setup

5.1 Datasets

Similar to Vakulenko et al. (2021), we experiment with CANARD (Elgohary et al., 2019) for extractive CQA and CAsT-19 (Dalton et al., 2020) for retrieval CQA. As CAsT-19 is small compared to CANARD, prior work (Vakulenko et al., 2021) uses the same model trained on CANARD to evaluate the rewriting performance on the test set of CAsT-19. Similarly, we start with the modelnon CANARD, and continue RL training with the BM25 reward on the training set without using any human annotations provided by CAsT-19.

5.2 Evaluation Metrics

We use BLEU-1, BLEU-4, ROUGE-1 and ROUGE-L for automatic evaluation. We also evaluate the performance of rewrites on downstream QA tasks. For CANARD, we use F1 and Exact Match (EM). For CAsT-19, we report MAP, MRR and NDCG@3 as in Vakulenko et al. (2021).

5.3 Baselines

We consider the following baselines:

Origin uses the original conversational question as input of QA.

BART$_{CQR}$ We fine-tune BART (Lewis et al., 2020) as a supervised baseline which has the same training procedure as the pre-training step of our method.

Co-reference (Vakulenko et al., 2021) is a rule-based method. We replace anaphoric expressions in original questions with their antecedents from the previous conversation turns. A public neural co-reference model (Lee et al., 2018) is used.

QR Method	QA Metrics		QR Metrics			
	EM	F1	B-1	B-4	R-1	R-L
Human	42.41	54.53	-	-	-	-
Original	38.41	48.95	61.06	30.98	69.91	69.71
Co-reference	38.17	48.99	54.95	30.84	74.11	73.40
BART$_{CQR}$	41.26	53.60	64.20	39.33	**76.70**	74.00
RL-QR	41.33	53.74	**64.25**	**39.52**	**76.70**	**74.01**
RL-F1	**41.91**	54.27†	62.32†	37.79†	74.93†	72.09†
RL-C	**41.91**	**54.61**†	57.47†	34.18†	71.12†	68.21†

Table 2: Overall QR and QA performance (%) on CANARD. **Bold** indicates the best results except "Human". We denote BLEU-n as B-n and ROUGE-n as R-n. † denotes statistically significant difference from BART$_{CQR}$ ($p < 0.05$ with t-test).

QR Method	QA Metrics		QR Metrics			
	EM	F1	B-1	B-4	R-1	R-L
BART$_{CQR}$ (50%)	41.37	53.52	**63.83**	**38.88**	**76.57**	**73.79**
RL-C (50%)	**42.09**	54.76†	62.13	37.52	75.03	72.10
RL-C (50%+100%)	42.05	**54.84**†	57.86†	34.44†	71.67†	68.54†

Table 3: QR and QA performance (%) of BART$_{CQR}$ and RL-C when using 50% of ground-truth rewriting. † denotes statistically significant difference from BART$_{CQR}$ (50%) ($p < 0.05$ with t-test).

Human uses the human rewrites and can be considered as an upper bound. However, we later show that the human baseline is the upper bound for QR target but not QA target.

5.4 Implementation Details

For all the QA models, we simulate the scenario where they are trained on single-turn QA data and cannot be updated when interacting with the rewriting component. The goal is to improve single-turn QA models for CQA, which means the input for QA models does not include any previous context.

Single-turn Extractive QA Model. To simulate a single-turn extractive QA model, we fine-tune ALBERT-XXLarge-v2 (Lan et al., 2020) on the CANARD training set.

Single-turn Retrieval QA Model. Same as in Vakulenko et al. (2021), we use Anserini's implementation of BM25 (Robertson et al., 2009) for the first-stage retrieval to obtain the top 1000 passages. In the second stage, we use BERT-large for passage re-ranking. Both components are fine-tuned on the MS MARCO dataset so that the two-stage pipeline resembles a single-turn retrieval QA model.

Rewriting Models. Our RL-based methods and the supervised BART baseline (BART$_{CQR}$) use BART-base model (Lewis et al., 2020).[1] We use the official CANARD validation set for early stopping. **RL-QR** denotes the model when QR rewards are used. **RL-F1**, **RL-C** and **RL-BM25** denote models where the F1, confidence and BM25 rewards are used, respectively.

6 Results

Here, we study the following research questions:
RQ1: Can our proposed QR and QA rewards improve the overall CQA performance? In particular, how effective are unsupervised rewards?
RQ2: Does achieving the best QR target mean achieving the best QA target?
RQ3: What is the quality, as judged by humans, of the reward-guided question rewrites?

6.1 Evaluation on Extractive CQA

We list the results on CANARD in Table 2. EM and F1 are QA metrics while others are QR metrics. We observe several trends.

First, RL-based methods achieve the best results on both QA or QR metrics over other non-human baselines. Compared with BART$_{CQR}$, our proposed RL methods can further improve the per-

[1] The max sequence length is set to 284, with batch size 24. An Adam weight decay optimizer with an initial learning rate of 1e-5 is used to train those models for 10 epochs.

QR Method	QA Metrics			QR Metrics			
	MAP	MRR	NDCG@3	B-1	B-4	R-1	R-L
Origin	17.85	46.44	27.86	71.63	51.54	82.65	81.24
Human	39.23	87.06	58.19	-	-	-	-
BART$_{CQR}$	28.02	61.49	44.04	**75.12**	**55.54**	84.82	83.84
Co-reference	26.82	59.74	43.05	71.19	51.79	**88.06**	**87.69**
RL-BM25	**28.41**	**63.20**	**45.54**	71.92	52.01	82.92	81.59

Table 4: QR and retrieval performance (%) on CAsT-19.

formance on QA target and QR target. Specifically, RL-C outperforms BART$_{CQR}$ by 1.88% and 1.58% in terms of F1 and EM, respectively. RL-QR achieves marginally better scores on BLEU-1, BLEU-4 and ROUGE-L than BART$_{CQR}$. RL-F1 achieves better F1 and EM scores than RL-QR and BART$_{CQR}$ but does not outperform RL-C. We notice that the F1 reward is less sensitive to question rewrites than the confidence reward. A small change in a question can lead to the same answer and F1 score. However, the confidence score can be different. In this aspect, RL-C seems to differentiate the fluctuations on rewrites better than RL-F1. In answer to **RQ1**, the confidence reward is the most effective for CQA performance. As an unsupervised reward which does not require either human rewrites or gold answers to a question, the confidence reward is even more effective than the F1 reward. However, we do not claim or target state-of-the-art performance in our work. The goal is to verify whether our RL framework for CQR with different rewards can further improve the performance of a single-turn QA system for CQA.

Second, using QR rewards (RL-QR) leads to limited performance improvement under both QA and QR metrics compared with BART$_{CQR}$. Maximizing the ROUGE rewards (Eq. 3) and minimizing the cross-entropy loss (Eq. 9) share the similar intuition that a good reformulation from the model should be similar to human reformulated questions. The two objectives are very close and therefore lead to similar results. It is important to note that the best scores of QR metrics and QA metrics are not achieved by the same method. Moreover, using QA rewards even lead to a large decrease in QR metrics. Therefore, in response to **RQ2**, achieving the best QR target does not mean achieving the best QA target, and vice versa.

Third, **RL-C achieves higher F1 scores than the human baseline**. Previous work (e.g. Vakulenko et al., 2021) treats human annotations as an upper bound. However, we argue that more human-like rewrites do not guarantee better QA performance. The results verify our hypothesis that QA target does not necessarily align with QR target. In §6.4, we qualitatively analyze if rewrites generated by RL-C are better than the ground-truth.

6.2 Training with Fewer Samples

For a real-world CQA system, we can obtain a large number of user questions with no corresponding ground-truth rewrites or answers. Since the confidence reward can be obtained easily from the downstream QA models without requiring human annotations, we can use RL-C to continue training the rewriting model. We first train a baseline using 50% of training data from CANARD (denoted as BART$_{CQR}$ (50%)). Then we continue RL training with the confidence reward using either the same 50% data used in pre-training (denoted as RL-C (50%)) or all questions in CANARD training set (denoted as RL-C (50%+100%)). The results are summarized in Table 3. We can see that RL-C (50%+100%) benefits from the large amount of questions during RL training and achieves better F1 and EM scores than RL-C (50%). Interestingly, RL-C (50%+100%) outperforms the human baseline in Table 2 by 0.31% in terms of F1. We also experimented with other ratios of data for supervised pre-training and continually RL training. In the experiments, we had similar observations that continual RL training with confidence rewards can further improve the downstream CQA performance.

6.3 Evaluation on Retrieval CQA

For RL-BM25, we use RL-C trained on CANARD as the pretrained model, then train it to maximize the BM25 reward, which can be readily obtained from the retrieval model. Results on CAsT-19 are shown in Table 4. As with extractive CQA, RL-BM25 achieves lower scores on QR metrics than

	RL-C vs. BART$_{CQR}$ (%)	RL-C vs. Human (%)
(1) RL is better	121 (60.5%)	105 (52.5%)
(2) RL is worse	39 (19.5%)	58 (29.0%)
(3) Both are good	28 (14.0%)	33 (18.5%)
(4) Both are bad	12 (6.0%)	4 (2.0%)
Total	200	200

Table 5: Results of user study comparing two groups of rewrites using four preference options.

Example 1	Original	What happened after **he** was fired?
	Human	What happened after **Aynsley Dunbar** was fired?
	BART$_{CQR}$	What happened after **Aynsley Dunbar** was fired?
	RL-C	What happened after **Aynsley Dunbar** was fired **by Herbie Herbert in late 1978?**
Example 2	Original	What position did **he** play?
	Human	What position did **Red Schoendienst** play?
	BART$_{CQR}$	What position did Ernie **Schoendienst** play?
	RL-C	What position did Don **Schoendienst** play **in the Majors**?

Table 6: Qualitative comparison of question rewrites. More examples are shown in Appendix C.

baselines. However, it improves the NDCG@3 of BART$_{CQR}$ by relatively 3.4%, which shows our framework also generalizes to retrieval CQA. Note that we do not use any supervised signals in CAsT-19 training set for RL training.

6.4 Human Evaluation

In addition to CQA performance, generating user-friendly rewrites is also important for real-world applications, since the rewrites sometimes will be displayed to users. To answer **RQ3**, we perform a user study to evaluate the quality of model generated rewrites. Specifically, two groups are compared: (1) The first group contains the rewrites generated by RL-C and human rewrites; (2) The second group contains rewrites from RL-C and BART$_{CQR}$, respectively. For each group, we randomly choose 200 questions from CANARD testing set. For each pair, we collect human's judgments on which rewrite contains more accurate context and details from conversation history.

The results are shown in Table 5. The study suggests that RL-C significantly performs better than Human and BART$_{CQR}$ (p-value < 0.001, see details in Appendix B.2). Remarkably, annotators prefer the rewrites from RL-C than humans in more than 50% cases. We show two examples in Table 6. In the first example, both RL-C and BART$_{CQR}$ correctly replace the pronoun with the referred person name. However, the rewrite generated by RL-C includes more accurate details which appear in conversation history. In the second example, both RL-C (same as RL-F1 and RL-QR) and BART$_{CQR}$ fail to generate the correct person's name. This error might be due to the prior knowledge of BART. To answer **RQ3**, we find that our reward-guided model can generate rewrites preferred by humans. However, all rewriting models can suffer from the coreference resolution problem.

7 Conclusion

We proposed a conversational question rewriting (CQR) approach using reinforcement learning. Such rewriting approaches are an emerging solution in real-world settings where QA systems with many existing answering backends trained on standalone questions must be adapted to work in conversational settings.

After assessing various QA and QR rewards, we showed that optimizing QR rewards is limited in improving CQA performance. In contrast, QA rewards that do not require ground-truth annotations consistently achieve the best CQA performance over baselines. For extractive CQA, using confidence rewards improved F1 by 2% over BART-based baseline on CANARD; and for retrieval CQA, using BM25 rewards improved the NDCG@3 of the baseline by 3.4% on CAsT-19. A human evaluation also demonstrated that our approach can generate higher-quality rewrites with more accurate and detailed context information.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2018. Ask the right questions: Active question reformulation with reinforcement learning. In *International Conference on Learning Representations*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec cast 2019: The conversational assistance track overview. *ArXiv*, abs/2003.13624.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6130–6141, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-oriented query reformulation with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. 2018. Conversational query understanding using sequence to sequence modeling. In *Proceedings of the 2018 World Wide Web Conference*, pages 1715–1724.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Min Tang, Jiaran Cai, and Hankz Hankui Zhuo. 2019. Multi-matching network for multiple choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7088–7095.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Hyperbolic representation learning for fast and efficient neural question answering. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 583–591.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 355–363.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, and Gaurav Singh Tomar. 2021. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. *arXiv preprint arXiv:2112.08558*.

Appendix

A Algorithm

There are two steps for training the rewriting model.

1. The 1st step pre-training (line 1-6) is to minimize the cross-entropy loss between the model's prediction q' and human ground-truth rewrite h. This objective is used in most of prior work (e.g., Vakulenko et al. (2021)).

2. The 2nd step (line 8-16) continues training the model with a reinforcement learning method (Self-Critical Sequence Training). In line 10 and 11, we only chose one of the reward functions to obtain the reward for a question. We leave the combination of different rewards as future work.

Algorithm 1: CQR training

```
   Input : Initialized rewriter $\mathcal{R}$, human question
           rewrites $H$, conversations $\mathcal{T}_{pre}$ for
           pre-training, conversations $\mathcal{T}_{rl}$ for RL
           training, selected reward function $\mathcal{F}$
   Output: Trained rewriter $\mathcal{R}$
   /* Step 1: pre-training $\mathcal{R}$ */
1  for $D \in \mathcal{T}_{pre}$ do
2      for question q, context $c \in D$ and $h \in H$ do
3          generate $q' = \mathcal{R}(q, c)$ (greedy decoding);
4          minimize loss in Equation 9;
5      end
6  end
7  /* Step 2: self-critical training
       */
8  for $D \in \mathcal{T}_{rl}$ do
9      for question q, context $c \in D$ and $h \in H$ do
10         generate $q^s$ from $\mathcal{R}(q, c)$ by sampling;
11         generate $q'$ from $\mathcal{R}(q, c)$ by greedy
             decoding;
12         obtain $r^s = \mathcal{F}(q^s)$;
13         obtain $r' = \mathcal{F}(q')$;
14         minimize loss in Equation 10;
15     end
16 end
```

B Human Study Design

For each annotation, an annotator is presented with the evidence document, conversation history, the original question and two rewrites. The annotator is required to select one from four options as listed in Table 5. The source of rewrite is anonymized. For each pair of rewrite, we randomly assign them to two options so that the judgments are not biased by the position of choices. We collect two judgments per rewrite pair. If there is a tie, we collect additional judgments. The final judgments are based on majority vote.

B.1 Appen Interface

Figure 3 shows the interface for annotators. Figure 4 contains the instruction which is visible for each annotator. In the instruction, we show several annotation examples in Figure 5.

B.2 Significance Tests

Here we describe how we conduct the Wilcoxon signed-rank test on the annotation results. When comparing RL-C with Human, for each sample, if annotators think RL-C is better, RL-C obtains score 1 and Human obtains score -1. Similarly, if annotators think Human is better, then Human obtains score 1 and RL-C obtains score -1. For other cases (i.e. both are good or both are bad), each of them obtains score 0. Then we use the method "scipy.stats.wilcoxon" in scipy library[2] to do the test. About the study annotator agreement rates, 48% samples have 100% agreement and the overall agreement rate is around 80%.

C Rewriting Examples

In Table 7, we show examples where the rewrites generated by RL-C are preferred by human annotators over the baseline method and ground truth. Compared with ground-truth rewrites, RL-C tends to generate rewrites with more factual details, which can help the user and also downstream QA systems to understand the question without conversation history. To some degree, it explains why the CQA performance is improved with RL-C, while the corresponding scores of QR metrics (i.e., BLEU-1, BLEU-4, ROUGE-1 and ROUGE-L) are very low. It also indicates that the human ground-truth in existing CQR datasets is not perfect and only evaluating CQR model with QR metrics can be biased.

The cases where both RL-C and the baseline generate incorrect rewrites are shown in Table 8. We can see that both methods make mistakes in coreference resolution. However, RL-C still has the tendency to include more conversational context in the rewrites.

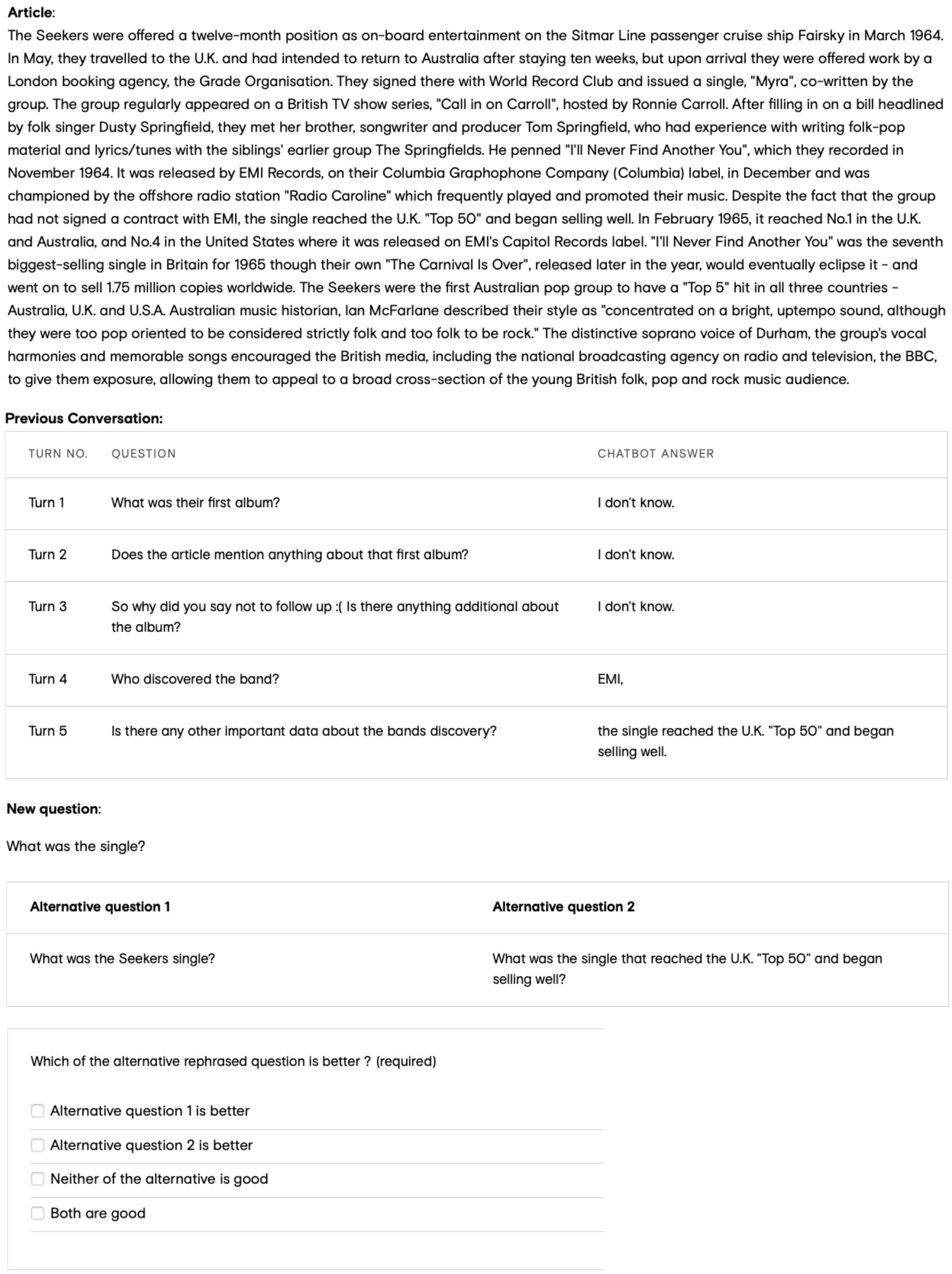
[2] https://docs.scipy.org

Original	What kind of things did **she** write about?
Human	What kind of things did **Le Guin** write about?
BART$_{CQR}$	What kind of things did **Le Guin** write about?
RL-C	What kind of things did **Le Guin** write about **in Sociology**?
Original	What did **he** study at the university?
Human	What did **Chinua Achebe** study at the university?
BART$_{CQR}$	What did **Chinua Achebe** study at the university?
RL-C	What did **Chinua Achebe** study at **the University of Ibadan**?
Original	What did **he** do after leaving Arrowverse?
Human	What did **John Barrowman** do after leaving Arrowverse?
BART$_{CQR}$	What did **John Barrowman** do after leaving Arrowverse?
RL-C	What did **John Barrowman** do after leaving the Arrowverse **television franchise?**
Original	What kind of topics did **the show** cover?
Human	What kind of topics did **the Rush Limbaugh Show** cover?
BART$_{CQR}$	What kind of topics did **Rush Limbaugh's show** cover?
RL-C	What kind of topics did **Rush Limbaugh's radio show** cover **in the 1970s**?
Original	What did he do after he landed?
Human	What did **Lindbergh** do after he landed?
BART$_{CQR}$	What did **Charles Lindbergh** do after he landed **at Le Bourget Aerodrome?**
RL-C	What did **Charles Lindbergh** do after he landed **at Le Bourget Aerodrome at 10:22 p.m. on Saturday, May 21, 1927**?

Table 7: Examples of rewrites where the reformulated questions from RL-C are the best judged by human annotators.

Original	Did others agree with **him**?
Human	Did others agree with **Gottfried Wilhelm** on the idea that the truth of religion and philosophy cannot contradict with each other?
BART$_{CQR}$	Did others agree with Leibniz's Theodicy?
RL-C	Did others agree with Leibniz's ideas that the truths of theology (religion) and philosophy cannot contradict each other, since reason and faith are both "gifts of God" in the Theodicy?
Original	What did **he** rejoin as?
Human	What did **Eddie Collins** rejoin as?
BART$_{CQR}$	What did Chris Hedges rejoin as?
RL-C	What did Chris Hedges rejoin as in August?
Original	What year did **his** first film debut?
Human	What year did was **Paul Verhoeven**'s first film debut?
BART$_{CQR}$	What year did Steven Seagal's first film debut?
RL-C	What year did James Cameron's first film debut? Flesh and Blood (1985)?
Original	Did **he** go into acting then?
Human	Did **Coogan** go into acting after college?
BART$_{CQR}$	Did Charlie Chaplin go into acting after A Day's Pleasure?
RL-C	Did Charlie Chaplin go into acting after college?
Original	Did **they** do a second album?
Human	Did **Gerry Mulligan and Chet Baker's quartet** do a second album?
BART$_{CQR}$	Did Pacific Jazz do a second album?
RL-C	Did Pacific Jazz do a second album **after PJ-8?**

Table 8: Examples question rewrites where both RL-C and BART$_{CQR}$ make mistakes.



Article:

The Seekers were offered a twelve-month position as on-board entertainment on the Sitmar Line passenger cruise ship Fairsky in March 1964. In May, they travelled to the U.K. and had intended to return to Australia after staying ten weeks, but upon arrival they were offered work by a London booking agency, the Grade Organisation. They signed there with World Record Club and issued a single, "Myra", co-written by the group. The group regularly appeared on a British TV show series, "Call in on Carroll", hosted by Ronnie Carroll. After filling in on a bill headlined by folk singer Dusty Springfield, they met her brother, songwriter and producer Tom Springfield, who had experience with writing folk-pop material and lyrics/tunes with the siblings' earlier group The Springfields. He penned "I'll Never Find Another You", which they recorded in November 1964. It was released by EMI Records, on their Columbia Graphophone Company (Columbia) label, in December and was championed by the offshore radio station "Radio Caroline" which frequently played and promoted their music. Despite the fact that the group had not signed a contract with EMI, the single reached the U.K. "Top 50" and began selling well. In February 1965, it reached No.1 in the U.K. and Australia, and No.4 in the United States where it was released on EMI's Capitol Records label. "I'll Never Find Another You" was the seventh biggest-selling single in Britain for 1965 though their own "The Carnival Is Over", released later in the year, would eventually eclipse it - and went on to sell 1.75 million copies worldwide. The Seekers were the first Australian pop group to have a "Top 5" hit in all three countries - Australia, U.K. and U.S.A. Australian music historian, Ian McFarlane described their style as "concentrated on a bright, uptempo sound, although they were too pop oriented to be considered strictly folk and too folk to be rock." The distinctive soprano voice of Durham, the group's vocal harmonies and memorable songs encouraged the British media, including the national broadcasting agency on radio and television, the BBC, to give them exposure, allowing them to appeal to a broad cross-section of the young British folk, pop and rock music audience.

Previous Conversation:

TURN NO.	QUESTION	CHATBOT ANSWER
Turn 1	What was their first album?	I don't know.
Turn 2	Does the article mention anything about that first album?	I don't know.
Turn 3	So why did you say not to follow up :(Is there anything additional about the album?	I don't know.
Turn 4	Who discovered the band?	EMI,
Turn 5	Is there any other important data about the bands discovery?	the single reached the U.K. "Top 50" and began selling well.

New question:

What was the single?

Alternative question 1	**Alternative question 2**
What was the Seekers single?	What was the single that reached the U.K. "Top 50" and began selling well?

Which of the alternative rephrased question is better ? (required)

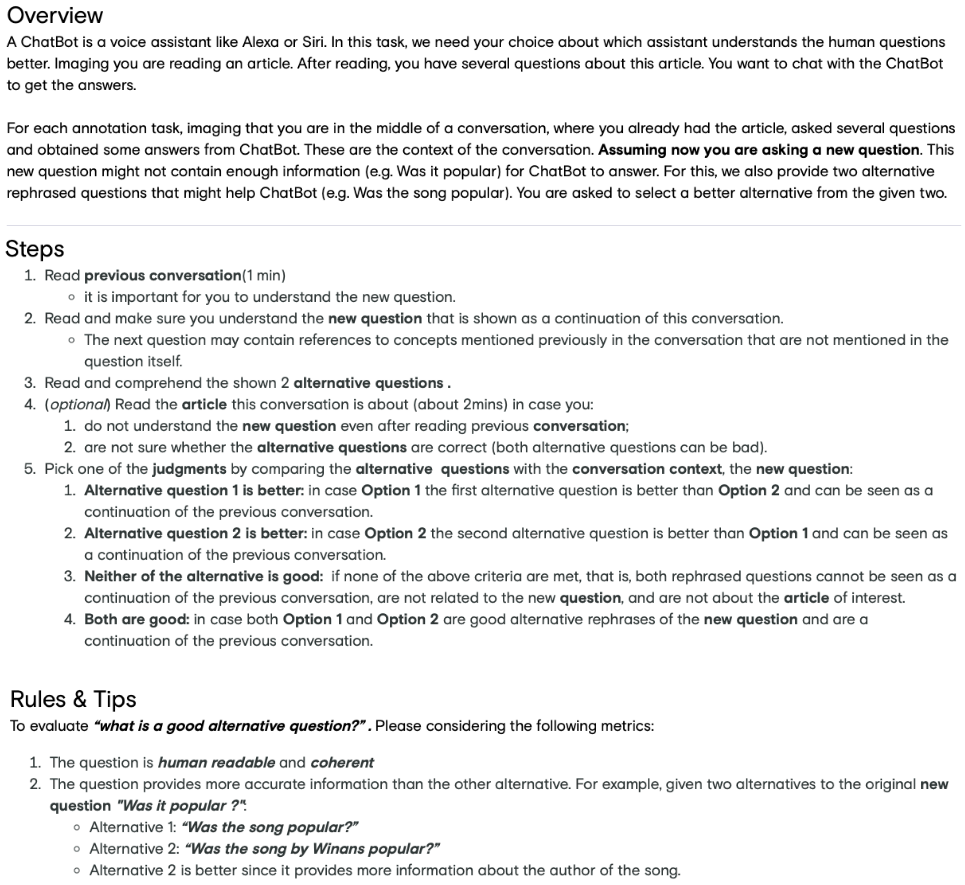
□ Alternative question 1 is better

□ Alternative question 2 is better

□ Neither of the alternative is good

□ Both are good

Figure 3: Interface on Appen.



Overview

A ChatBot is a voice assistant like Alexa or Siri. In this task, we need your choice about which assistant understands the human questions better. Imaging you are reading an article. After reading, you have several questions about this article. You want to chat with the ChatBot to get the answers.

For each annotation task, imaging that you are in the middle of a conversation, where you already had the article, asked several questions and obtained some answers from ChatBot. These are the context of the conversation. **Assuming now you are asking a new question**. This new question might not contain enough information (e.g. Was it popular) for ChatBot to answer. For this, we also provide two alternative rephrased questions that might help ChatBot (e.g. Was the song popular). You are asked to select a better alternative from the given two.

Steps

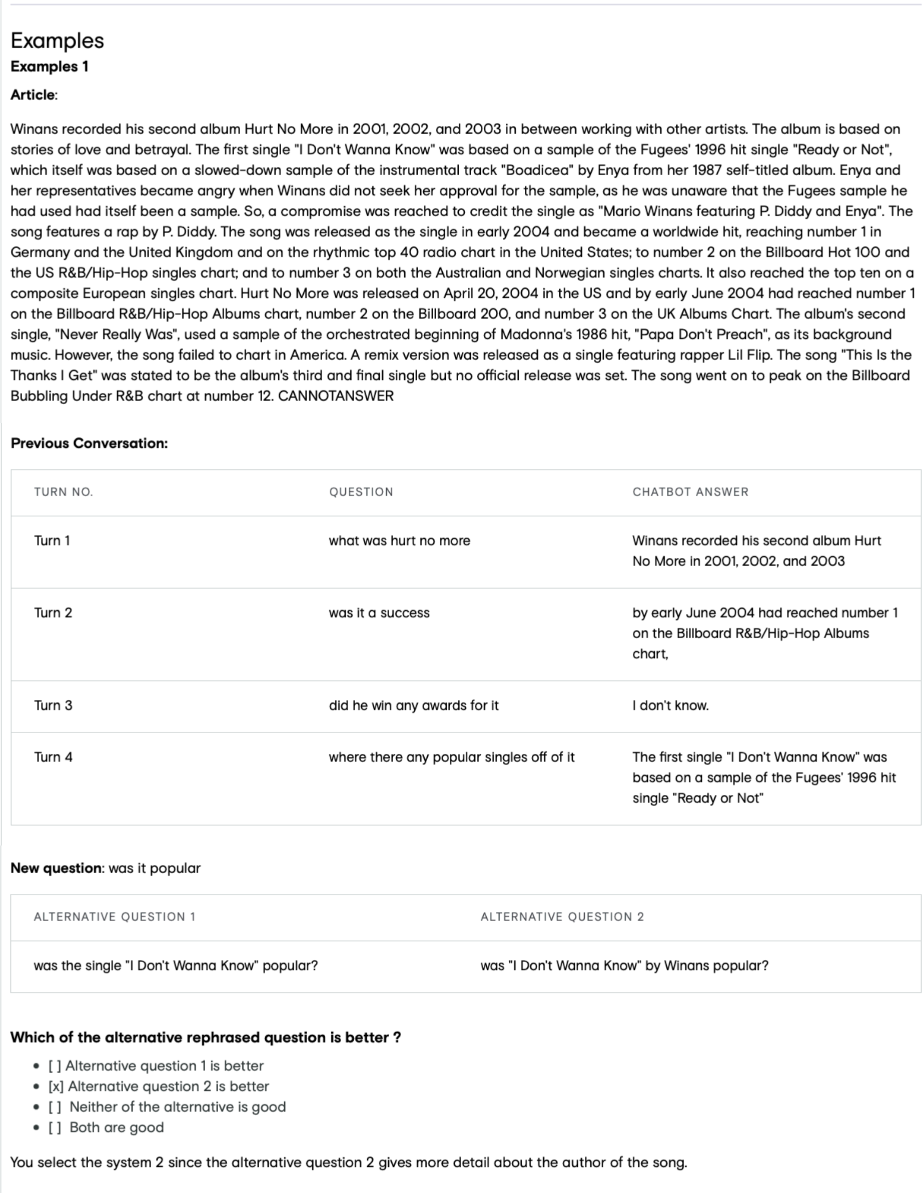
1. Read **previous conversation**(1 min)
   - it is important for you to understand the new question.
2. Read and make sure you understand the **new question** that is shown as a continuation of this conversation.
   - The next question may contain references to concepts mentioned previously in the conversation that are not mentioned in the question itself.
3. Read and comprehend the shown 2 **alternative questions .**
4. (*optional*) Read the **article** this conversation is about (about 2mins) in case you:
   1. do not understand the **new question** even after reading previous **conversation**;
   2. are not sure whether the **alternative questions** are correct (both alternative questions can be bad).
5. Pick one of the **judgments** by comparing the **alternative questions** with the **conversation context**, the **new question**:
   1. **Alternative question 1 is better:** in case **Option 1** the first alternative question is better than **Option 2** and can be seen as a continuation of the previous conversation.
   2. **Alternative question 2 is better:** in case **Option 2** the second alternative question is better than **Option 1** and can be seen as a continuation of the previous conversation.
   3. **Neither of the alternative is good:** if none of the above criteria are met, that is, both rephrased questions cannot be seen as a continuation of the previous conversation, are not related to the new **question**, and are not about the **article** of interest.
   4. **Both are good:** in case both **Option 1** and **Option 2** are good alternative rephrases of the **new question** and are a continuation of the previous conversation.

Rules & Tips

To evaluate ***"what is a good alternative question?" .*** Please considering the following metrics:

1. The question is ***human readable*** and ***coherent***
2. The question provides more accurate information than the other alternative. For example, given two alternatives to the original **new question *"Was it popular ?"***:
   - Alternative 1: ***"Was the song popular?"***
   - Alternative 2: ***"Was the song by Winans popular?"***
   - Alternative 2 is better since it provides more information about the author of the song.

Figure 4: Instruction for annotators.



Examples

Examples 1

Article:

Winans recorded his second album Hurt No More in 2001, 2002, and 2003 in between working with other artists. The album is based on stories of love and betrayal. The first single "I Don't Wanna Know" was based on a sample of the Fugees' 1996 hit single "Ready or Not", which itself was based on a slowed-down sample of the instrumental track "Boadicea" by Enya from her 1987 self-titled album. Enya and her representatives became angry when Winans did not seek her approval for the sample, as he was unaware that the Fugees sample he had used had itself been a sample. So, a compromise was reached to credit the single as "Mario Winans featuring P. Diddy and Enya". The song features a rap by P. Diddy. The song was released as the single in early 2004 and became a worldwide hit, reaching number 1 in Germany and the United Kingdom and on the rhythmic top 40 radio chart in the United States; to number 2 on the Billboard Hot 100 and the US R&B/Hip-Hop singles chart; and to number 3 on both the Australian and Norwegian singles charts. It also reached the top ten on a composite European singles chart. Hurt No More was released on April 20, 2004 in the US and by early June 2004 had reached number 1 on the Billboard R&B/Hip-Hop Albums chart, number 2 on the Billboard 200, and number 3 on the UK Albums Chart. The album's second single, "Never Really Was", used a sample of the orchestrated beginning of Madonna's 1986 hit, "Papa Don't Preach", as its background music. However, the song failed to chart in America. A remix version was released as a single featuring rapper Lil Flip. The song "This Is the Thanks I Get" was stated to be the album's third and final single but no official release was set. The song went on to peak on the Billboard Bubbling Under R&B chart at number 12. CANNOTANSWER

Previous Conversation:

TURN NO.	QUESTION	CHATBOT ANSWER
Turn 1	what was hurt no more	Winans recorded his second album Hurt No More in 2001, 2002, and 2003
Turn 2	was it a success	by early June 2004 had reached number 1 on the Billboard R&B/Hip-Hop Albums chart,
Turn 3	did he win any awards for it	I don't know.
Turn 4	where there any popular singles off of it	The first single "I Don't Wanna Know" was based on a sample of the Fugees' 1996 hit single "Ready or Not"

New question: was it popular

ALTERNATIVE QUESTION 1	ALTERNATIVE QUESTION 2
was the single "I Don't Wanna Know" popular?	was "I Don't Wanna Know" by Winans popular?

Which of the alternative rephrased question is better ?

- [ ] Alternative question 1 is better
- [x] Alternative question 2 is better
- [ ] Neither of the alternative is good
- [ ] Both are good

You select the system 2 since the alternative question 2 gives more detail about the author of the song.

Figure 5: An annotation example in the instruction.