Integrating Question Rewriting in Conversational Question Answering: A Reinforcement Learning Approach

Etsuko Ishii[*], **Bryan Wilie**[*], **Yan Xu**[*], **Samuel Cahyawijaya**[*], **Pascale Fung**
The Hong Kong University of Science and Technology
{eishii, bwilie, yxucb, scahyawijaya}@connect.ust.hk

Abstract

Resolving dependencies among dialogue history is one of the main obstacles in the research on conversational question answering (CQA). The conversational question rewrites (QR) task has been shown to be effective to solve this problem by reformulating questions in a self-contained form. However, QR datasets are limited and existing methods tend to depend on the assumption of the existence of corresponding QR datasets for every CQA dataset. This paper proposes a reinforcement learning approach that integrates QR and CQA tasks without corresponding labeled QR datasets. We train a QR model based on the reward signal obtained from the CQA, and the experimental results show that our approach can bring improvement over the pipeline approaches. The code is available at https://github.com/HLTCHKUST/cqr4cqa.

1 Introduction

Conversational Question Rewrites (QR) systems paraphrase a question into a self-contained format using its dialogue history so as to make it easier to understand by the Conversational Question Answering (CQA) system. Prior works (Elgohary et al., 2019a; Anantha et al., 2021a; Kim et al., 2021) have shown that explicit guidance of QR benefits the performance of the CQA models in multiple questions answering datasets.

However, the existing works on QR in the context of CQA are often ignorant of two critical issues. Firstly, they are dependent on the assumption that QR datasets exist on target CQA datasets, although existing QR datasets only cover a small amount of CQA. It is also notable that building a novel QR dataset is expensive. Current works mainly focus on QuAC (Choi et al., 2018) datasets thanks to QR datasets constructed from it (Elgohary et al., 2019a; Anantha et al., 2021a), however, the other popular CQA datasets such as CoQA (Reddy et al., 2019) remain less explored. Secondly, although QR task evaluation is mainly done by automatic metrics that compute n-gram overlaps with BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), and by human evaluation, there is no correlation guaranteed between those metrics and the performance in CQA. In fact, Petrén Bach Hansen and Søgaard (2020) and Buck et al. (2018) suggest "better" rewrites in the human eye are not necessarily better for machines.

To this end, we propose to alleviate the limitation of the QR system by introducing a reinforcement learning framework that utilizes QR to overcome the aforementioned two obstacles. In this framework, a QR model plays the role of "the agent" which receives rewards from a CQA model which acts as "the environment." During training, a QR model aims to maximize the performance on the CQA task by generating better rewrites of the questions. Exploiting the reinforcement learning nature, we can benefit CQA regardless of the existence of QR annotation, and we can ensure that QR contributes to the final objective of improving CQA. Experimental results show that our framework successfully improves the CQA performance by 4.1 to 8.6 F1 score on CoQA and 4.7 to 9.2 F1 on QuAC compared to the pipeline baselines that combine a QR model and a QA model.

Our contributions in this paper can be summarized three-fold as follow:

- We propose a reinforcement learning framework for CQA which can be applied regardless of the existence of corresponding QR datasets.

- Our experimental results on two popular CQA datasets show that our approach improves the performance over the simple combination baselines of a QA and QR model.

[*] Equal Contribution

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Student Research Workshop, pages 55 - 66
May 22-27, 2022 ©2022 Association for Computational Linguistics

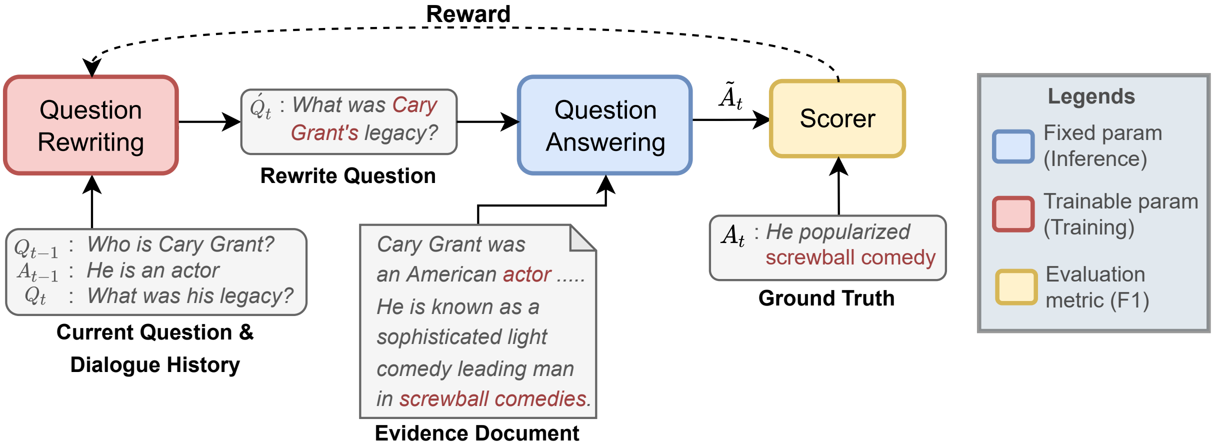
Reward
Question Rewriting
$\acute{Q}_t$: What was Cary Grant's legacy?
Rewrite Question
Question Answering
$\tilde{A}_t$
Scorer
Legends
Fixed param (Inference)
Trainable param (Training)
Evaluation metric (F1)
Q_{t-1} : Who is Cary Grant?
A_{t-1} : He is an actor
Q_t : What was his legacy?
Current Question & Dialogue History
Cary Grant was an American actor ..... He is known as a sophisticated light comedy leading man in screwball comedies.
Evidence Document
A_t : He popularized screwball comedy
Ground Truth

Figure 1: Overview of our reinforcement learning approach for CQA task that involves a QR model and a QA model. The current question Q_t and its dialogue history are reformulated into a self-contained question $\acute{Q}_t$ by the QR model. Then, $\acute{Q}_t$ and optionally, its dialogue history is passed to the QA model to extract an answer span $\tilde{A}_t$ from the provided evidence document. We train the QR model by maximizing the reward signal (F1 score) obtained by the comparison between the predicted answer span $\tilde{A}_t$ and the gold span A_t.

- We provide extensive analysis on suitable settings for our approach, such as training algorithms for the QR model and existing QR datasets for the QR model initialization.

2 Related Work

2.1 Conversational Question Answering

Recently, along with the raised popularity of works on dialogue systems (Madotto et al., 2020b,a; Ishii et al., 2021; Lin et al., 2021; Xu et al., 2021a; Liu et al., 2019b) and question answering(Su et al., 2020, 2019, 2022), conversational question answering (CQA) has gained more attention. CQA task aims to assist users for information-seeking purpose. It has been widely studied in the recent years and many CQA datasets have been made publicly available, such as CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), ShARC(Saeidi et al., 2018), Doc2Dial (Feng et al., 2020), and DoQA (Campos et al., 2020). Existing works focus on improving the model structure (Zhu et al., 2018; Huang et al., 2018; Yeh and Chen, 2019; Ohsugi et al., 2019; Zhang et al., 2021; Zhao et al., 2021) to deal with the dialogue history, leveraging different training techniques, such as adversarial training (Ju et al., 2019), or utilizing data augmentation via multi-task learning (Xu et al., 2021b) to improve the model performance in an end-to-end fashion. Unlike the aforementioned works, we propose to increase performance on CQA tasks by improving the readability of the questions via reinforcement learning. It may not align with the question readability for human beings. Furthermore, Our proposed approach does not contradict with the above methods, but can co-exist with them in the CQA system instead.

2.2 Question Rewrites

As the key challenge in CQA is to understand a highly-contextualized question, several QR datasets are proposed to offer a subtask in CQA which is to paraphrase a question in a self-contained style (Elgohary et al., 2019a; Petrén Bach Hansen and Søgaard, 2020; Anantha et al., 2021a). While many of the existing works put more effort on generating high-quality rewrites (Lin et al., 2020; Vakulenko et al., 2021), Kim et al. (2021) recently introduced a framework to leverage QR to improve the performance of CQA models by additional a consistency-based regularization. Their EXCORD feeds the original questions together with the rewritten questions, whereas we only use the rewritten questions. Similar to our work, Buck et al. (2018) train a question rewriting model in reinforcement learning framework interacting with the question answering environment to transform a given query from a declarative sentence into an interrogative one. It is noteworthy that QR in CQA requires more effort than in their setting, since we seek a QR model to elaborate dialogue history so as to resolve anaphora or ellipsis in a question to rewrite, rather than simple grammatical transformation of the given query.

2.3 Reinforcement Learning in Natural Language Generation

One of the most common reasons to adopt reinforcement learning (RL) methods in natural language generation (NLG) is due to inconsistency between train/test measurement (Keneshloo et al., 2019). If we apply deep neural network, we often train with differentiable loss function such as token-wise cross-entropy; but in test time, we use BLEU or ROUGE which we cannot directly use as a loss function. Motivated as such, RL approaches have been investigated in various NLG tasks, for example, in machine translation (Ranzato et al., 2016; Wu et al., 2016; He et al., 2017; Bahdanau et al., 2017), abstractive summarization (Ranzato et al., 2016; Paulus et al., 2018; Böhm et al., 2019), or dialogue generation (Li et al., 2016). If we train an NLG model from scratch with reinforcement learning, however, a model frequently suffers from too large exploration space (Ranzato et al., 2016). Thanks to the recent advance in large pretrained language models, RL approaches are investigated as an alternative fine-tuning approach which can reflect human preferences (Ziegler et al., 2019; Stiennon et al., 2020; Jaques et al., 2020) or reduce non-normative text (Peng et al., 2020). Our work also utilize large pretrained language models and use RL approach for fine-tuning.

3 Methodology

In this section, we present our reinforcement learning framework and the training algorithm. Firstly, we offer several preliminary definitions used throughout the paper, and secondly, we describe the strategy to train the whole framework.

3.1 Preliminary Definition

We denote a CQA dataset as $\{\mathcal{D}^n\}_{n=1}^{N}$ and the dialogue history at turn t as $\mathcal{D}_t = \{(Q_i, A_i)\}_{i=1}^{t}$, where Q_t is the question and A_t is the answer. Along with the QA pairs, the corresponding evidence document Y_t is also given.

In our proposed framework, a QA model and a QR model are involved. In CQA tasks, the answers to the questions are composed as pairs of start indexes and end indexes in the given paragraphs, where we denote as $A_t = \{{a_t}^s, {a_t}^e\}$. Let's denote a generated rewrite question sequence of Q_t as $\acute{Q}_t = \{\acute{q}_l\}_{l=1}^{L}$. The objective of the QR model is to rewrite the question Q_t at turn t into a self-contained version, based on the current question and the dialogue history $\mathcal{D}_{t-1}$.

Algorithm 1 RL training process of our QR agent

```
Require: $\{\mathcal{D}^n\}$: CQA dataset
Require: $\pi_{\theta_0}$: Pretrained language model
Train an environment $f_\phi$ on $\{\mathcal{D}^n\}$
Initialize an agent $\pi_\theta$ with $\pi_{\theta_0}$
while not done do
    Sample an input state from
      the CQA dataset $X_t \sim \{\mathcal{D}^n\}$
    Construct a rewrite sequence $\acute{Q}_t$
      which maximize $\pi_\theta(\acute{Q}_t|X_t)$
    Calculate F1-score $r$ via $r(f_\phi(\acute{X}_t))$
    Update $\pi_\theta$ using an RL algorithm with
      state $X_t$, action $Q_t$, and reward $R_t$
end while
```

As shown in Figure 1, we consider in a reinforcement learning framework. An agent takes an input state $X_t = (\mathcal{D}_{t-1}, Q_t)$ and generates a paraphrase $\acute{Q}_t$. Then, $\acute{X}_t = (\mathcal{D}_{t-1}, \acute{Q}_t)$ and an evidence document Y_t are provided to an environment, namely, a QA model f_ϕ, which extracts an answer span $\tilde{A}_t = f_\phi(\acute{X}_t, Y_t)$. We aim the agent, a QR model π_θ, to learn to generate a high-quality paraphrase of given question based on the reward received from the environment.

The policy, in our case the QR model, assigns the probability

$$\pi_\theta(\acute{Q}_t|X_t) = \prod_{l=1}^{L} p(\acute{q}_l|\acute{q}_1, \ldots, \acute{q}_{l-1}, X_t). \quad (1)$$

Our goal is to maximize the expected reward of the answer returned under the policy, namely,

$$\mathrm{E}_{\acute{q}_t \sim \pi_\theta(\cdot|q_t)}[r(f_\phi(\acute{X}_t))], \quad (2)$$

where r is a reward function. We apply the token-level F1-score between the predicted answer span $\tilde{A}_t$ and the gold span A_t as the reward r. We can directly optimize the expected reward in Eq. 2 using reinforcement learning algorithms.

3.2 Training Algorithm

Prior to the training process, the QA model f_ϕ is fine-tuned on $\{\mathcal{D}^n\}$ and the QR model is initialized with $\pi_\theta = \pi_{\theta_0}$, where π_{θ_0} is a pretrained language model. We apply Proximal Policy Optimization (PPO) (Schulman et al., 2017; Ziegler et al., 2019) to train π_θ. PPO is a policy gradient method which alternate between sampling data through interaction with the environment and optimizing a surrogate objective function via stochastic gradient ascent. PPO makes use of learned state-value function to compute the variance-reduced advantage-function. The overview of our training process is shown in Algorithm 1.

Following Ziegler et al. (2019), we penalize the reward r with a KL-penalty so as to prevent the policy π_θ from drifting too far away from π_{θ_0}:

$$R_t = R(\acute{X}_t) = r(f_\phi(\acute{X}_t)) - \beta \mathrm{KL}(\pi_\theta, \pi_{\theta_0}),$$

where β represents a weight factor. We perform reinforce learning on the modified reward of R_t instead of r.

Inspired by MIXER (Ranzato et al., 2016), we apply a cross-entropy loss on the first m tokens of the generated sequence $\acute{Q}_t$ by using the tokens from the original question Q_t as the label to enhance training stability in addition to the KL-penalty. We decrease the number of tokens where we apply the cross-entropy loss over time steps, allowing the policy π_θ to explore more. By applying the cross-entropy constraint, the PPO objective function $\mathcal{L}(\theta)$ is modified into:

$$\mathcal{L}_l = \begin{cases} -\sum_{i=1}^{|\mathcal{V}|} q_{i,l} \log(\acute{q}_{i,l}) & (l \leq m) \\ \mathcal{L}^{\mathrm{CLIP}}(\acute{q}_l) + c\mathcal{L}^{\mathrm{VF}}(\acute{q}_l) & (l > m) \end{cases} \quad (3)$$

$$\mathcal{L}(\theta) = \frac{1}{m}\sum_{l=1}^{m} \mathcal{L}_l + \sum_{l=m+1}^{L} \mathcal{L}_l, \quad (4)$$

where $|\mathcal{V}|$ is the vocabulary size, $\mathcal{L}^{\mathrm{CLIP}}$ is the clipped surrogate loss (see Eq. 7 in Schulman et al. (2017)), c is a value loss coefficient, and $\mathcal{L}^{\mathrm{VF}}$ is the value function loss (see Eq. 9 in Schulman et al. (2017)).

In addition to MIXER, we introduce another strategy to improve exploration (denoted as EXPLORE) that comes along with beam-search decoding. The strategy of using beam-search is to search for top-k sequences with the highest likelihood during the generation process and take the one with the highest likelihood over k sequences. Our approach is utilizing the top-k' ($1 \leq k' \leq k$) sequences collected during beam search for PPO training. With this approach, we can improve the exploration capability of the QR model without requiring additional exploration steps.

4 Experiments

4.1 Datasets

We conduct our experiments on two CQA datasets, CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018). Since the test set is not publicly available for both CoQA and QuAC, we follow the splitting introduced by (Kim et al., 2021). We leverage the train/dev/test split provided by Kim et al. (2021) for the QuAC experiments. We randomly sample 5% of data samples in the training set in units of dialogues and adopt them as our validation set for CoQA since there is no public split available.

CoQA CoQA dataset contains 127K questions with answers in the conversation form (8K conversations in total). The questions are highly contextualized with the dialogues, and the answers are free-form text with their corresponding evidence in the passage for reference. In this paper, following the settings of the other existing works (Huang et al., 2018; Yeh and Chen, 2019; Ju et al., 2019), we still construct the answers as spans extracted from the passages, where the gold labels used in training are the snippets with the highest F1 score compared to the annotated answers.

QuAC QuAC is also a crowd-sourced CQA dataset that contains 14K information-seeking QA conversations. In contrast to CoQA, QuAC is designed as a span-extraction dataset with dialogue acts. Moreover, 20% of the questions in QuAC are unanswerable questions, whereas those in CoQA take up ∼1.3%.

We initialize the QR model with a pre-trained language model that is fine-tuned on QR datasets. We apply two QR datasets, i.e., QReCC and CANARD, for the fine-tuning to obtain more insights on the influence of the QR model initialization with different data sources.

CANARD CANARD dataset (Elgohary et al., 2019b) is a question-rewriting (QR) dataset which aims at conducting question-in-context rewriting to convert the questions with long conversation histories into short and self-contained questions. The questions are generated by rewriting a subset of the original questions in QuAC dataset. The dataset is split in to training, development, and test sets in size of 31K, 3.4K, and 5.6K.

QReCC QReCC dataset (Anantha et al., 2021b) is another QR dataset. In contrast to CANARD,

Models		CoQA						QuAC		
		Overall F1	Child.	Liter.	M&H	News	Wiki.	F1	HEQ-Q	HEQ-D
	end-to-end	84.5	**84.4**	82.4	**82.9**	86.0	86.9	**67.8**	63.5	7.9
QReCC	pipeline-eval	80.6	80.8	78.6	78.3	81.7	83.7	62.9	58.5	5.2
	pipeline-train	82.9	82.9	80.9	81.5	84.4	84.8	66.3	62.0	6.6
	ours	**84.7**	84.3	**83.1**	82.7	**86.3**	86.8	67.6	63.2	7.8
CANARD	pipeline-eval	75.9	75.5	74.8	73.1	76.3	79.8	58.2	54.5	5.2
	pipeline-train	82.8	83.4	80.1	80.8	84.4	85.6	66.5	62.5	7.4
	EXCORD†	83.4	**84.4**	81.2	79.8	84.6	**87.0**	67.7	**64.0**	**9.3**
	ours	84.4	84.1	82.7	82.6	86.0	86.7	67.4	62.7	8.1

Table 1: Evaluation results of our approach and baselines on the test set. EXCORD† follows the results reported in Kim et al. (2021). **Bold** are the best results amongst all. Underlined represents the best score on each combination of the CQA and QR datasets.

QReCC dataset is built upon three publicly available datasets: QuAC, TREC Conversational Assistant Track (CAsT) (Dalton et al., 2020) and Natural Questions (NQ) (Kwiatkowski et al., 2019), where QreCC contains 14K dialogues with 80K questions in total, and 9.3K dialogues are from QuAC. The sampled data are further extended to the open-domain CQA setting. It also supports the passage retrieval and reading comprehension tasks. The dataset is split into training, development, and test sets. However, in the released version, the training and development set are merged. In the experiments, we directly train the model with the merged training set and use the test set for evaluation.

4.2 Evaluation Metrics

To automatically evaluate the performance of the QA models, following Reddy (2020), we leverage the unigram F1 score. In CoQA evaluation, the models are also evaluated per domain on all six domains as listed in the official validation set (our test set instead), i.e., Children Stories (Child.), Literature (Liter.), Mid-High School (M&H), News, and Wikipedia (Wiki.). Following the leaderboard, for the QuAC dataset, we incorporate the human equivalence score HEQ-Q and HEQ-D for QuAC evaluation. HEQ-Q indicates the percentage of questions on which the model outperforms human beings and HEQ-D represents the percentage of dialogues on which the model outperforms human beings for every question in the dialogue.

4.3 Models

QA model In all the experiments, we leverage pre-trained RoBERTa (Liu et al., 2019a) model as the initial model and adapt it to different CQA tasks (see Table A1 in Appendix for more details). The RoBERTa model is the leading pre-trained model according to different leaderboards and it has shown its effectiveness on QA tasks (Ju et al., 2019; Zhao et al., 2021; Yasunaga et al., 2021; Zhu et al., 2021). The model is trained to predict the start positions and the end positions of the given contexts with respect to the questions. Since our proposed method is model-agnostic, the QA component in the framework can be replaced with any existing QA models.

QR model We use GPT-2 (Radford et al., 2019) as the base model to train the QR models (see Table A2 in Appendix for more details). In the QR training process, we provide the dialogue history and the current question as the inputs and train the model to rewrite the current question into a self-contained version that is able to be answered without considering the dialogue history.

Model selection and initialization in RL Before applying our methods, the QA and QR models are initialized with the best QA and QR baseline models. For both QA models that are trained on CoQA and QuAC datasets, the models with the highest F1 score on the validation set are selected. We use different metrics for the QR model selection on two datasets, following the original metrics that are used for model evaluation. We select the best QR model checkpoint on the CANARD dataset and

	... Far on in the hot days of June the Excommunication, for some weeks arrived from Rome, was solemnly published in the Duomo. Romola went to witness the scene, that the resistance it inspired might invigorate that sympathy with Savonarola ...			... Jenny loves singing. But her baby sister is crying so loud that Jenny can't hear herself, so she was angry! Her Mom said she could try to play with her sister, but that only made ...	
	Utterance	F1 Score		Utterance	F1 Score
Q_{t-1}	Where **was the Excommunication published?**		Q_{t-1}	How **is she feeling?**	
A_{t-1}	in the Duomo		A_{t-1}	Angry.	
Q_t	When?	0.61	Q_t	Why?	0.82
$\acute{Q}_t$	When **was the Excommunication published?**	1.0	$\acute{Q}_t$	Why **is she feeling?**	0.98

Table 2: Examples of rewritten questions by the trained QR model initialized with QReCC. We can see that the model learns how to recover the abbreviated contents from the dialogue history to get a better score on CoQA.

QReCC dataset based on the BLEU [1] score and the unigram recall (ROUGE-1 R) score respectively.

4.4 Baselines

We compare our proposed approach with three different settings: (i) directly finetuning the QA model on the CQA tasks without the QR model (**end-to-end**), (ii) inferencing the QA model with questions rewritten by the QR model (**pipeline-eval**), and (iii) finetuning the QA model with questions rewritten by the QR model (**pipeline-train**).

4.5 Experimental Setup

Our implementation is based on Wolf et al. (2020). We conduct all of the experiments with GeForce RTX 2080 Ti. To obtain the models for initialization, GPT-2 is trained on QReCC and CANARD dataset as the QR model, and RoBERTa is trained on CoQA and QuAC datasets as the QA model with Adam optimizer (Kingma and Ba, 2015) with a learning rate of $3e-5$. We report other hyperparameters used in the model initialization in Table A1 and Table A2 in Appendix.

For PPO training, we train the QR model with Adam optimizer with a learning rate of $1e-7$. Further, we use beam search with beam size of 5, preventing generation repetition (Keskar et al., 2019) with using repetition penalty of 1.1, and set the maximum input sequence length to 512. On the MIXER settings, we initialize cross entropy length as 3 and limits its minimum to 1. We then run the PPO with value function coefficient of 1.0, while ensuring the sequence length of question rewriting model input to be 150 tokens maximum and the generations length to be 50 tokens at maximum. To ensure that the learned policy does not deviate too much, we apply an additional reward signal, adaptive KL factor β according to the magnitude of the KL-penalty with a KL-coefficient $K_\beta = 0.1$. Other hyperparameters are listed in Table A3 in Appendix.

4.6 Results

We report our experimental results in Table 1. Our approach achieves 84.7 and 67.6 F1 scores on the CoQA and QuAC datasets, respectively, and constantly outperforms the pipeline baselines. As shown in the examples listed in Table 2, our approach successfully teaches the QR model to refer to the dialogue history and recover the abbreviated contents if necessary. Comparing to the pipeline-eval, our approach scores at least 4.1 F1 score better, which indicates that our reinforcement learning approach successfully trains the QR model to paraphrase the questions into more preferred format of the QA model from the QR model initialization. Our approach performs better at least by 1.0 overall F1 score than EXCORD (Kim et al., 2021) on CoQA, and comparably on QuAC. However, our approach could not contribute to the considerable improvement over the end-to-end baseline, which poses the need for further investigation.

5 Discussion

In this section, we provide our findings regarding the most suitable settings of our approach, including the comparison of the QR datasets for initialization, the QR model architectures, the training algorithms, and the effect of the decoding strategy. For automatic evaluation, we report Exact Match (EM) in addition to the unigram F1 score. EM in-

[1] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu-detok.perl

Models	# Params	Evaluation	
		F1	EM
GPT-2	243M	**84.7**	**76.6**
BART-base	210M	83.6	75.7

Table 3: Comparison of the QR model architectures initialized with QReCC and further trained on CQA. GPT-2 achieves a higher F1 score and EM than BART-base. Note that we have twice as many parameters as GPT-2 and 1.5 times as BART-base since we copy the decoders and train them for estimating the value function.

dicates that the percentage of the predictions is the same as the gold answers, while the F1 score evaluates the performance with uni-gram overlapping.

5.1 Comparison of QR datasets for Initialization

We compare the effect of QR dataset initialization, i.e., QReCC and CANARD, to the evaluation performance of the CQA task. As shown in Table 1, QR models trained with QReCC dataset almost always performs better than the ones trained with CANARD dataset, with one exception on the pipeline-train approach on the QuAC dataset. We assume this is because QReCC gives more generalization ability to the QR models since QReCC is composed of several QA datasets, whereas CANARD is more devoted to QuAC as it is a subset of QuAC.

5.2 Comparison of QR Model Architectures

In the search for suitable architecture, we compare the architectures of the QR models in the RL training both using GPT-2 and BART-base (Lewis et al., 2020). First, we train BART with QReCC in the same way as GPT-2 and then fine-tune it with CoQA using our reinforcement learning approach. It is noteworthy that utilizing the BART-base serves the system worse fits compared with using GPT-2 as reported in Table 3. However, this performance gap could be due to our implementation. For BART, we only copy the decoder to estimate the value function, resulting in 70M parameters for estimating the value function, but GPT-2 uses all 117M parameters for it. In future, we plan to attempt to use the whole parameters of BART for estimating the value function.

Algorithm	CoQA		QuAC	
	F1	EM	F1	EM
PPO	**84.7**	**76.6**	**67.6**	**51.3**
REINFORCE	84.2	76.1	64.8	49.3

Table 4: Comparison between training algorithms of the QR model. PPO scores constantly better than the REINFORCE algorithm.

5.3 Comparison of RL algorithms

In addition to the PPO approach, we explore the REINFORCE algorithm (Williams, 1992) to train π_θ. We use self-critical sequence training (Rennie et al., 2017) instead of MIXER in REINFORCE experiments. In self-critical sequence training, we normalize the reward r derived from the sampled rewrites $\acute{X}_t$ with the reward derived from another rewrite which is generated by greedy decoding. We use Adam optimizer with a learning rate of $1\mathrm{e}-7$ and keep the other hyperparameters the same as the PPO training.

We report the evaluation results of CoQA and QuAC with the initialization of QReCC in Table 4. REINFORCE could not outperform PPO, although bringing some improvement over the majority of the pipeline baselines. This observation that PPO is better than the REINFORCE supports the experimental results reported in Andrychowicz et al. (2021).

5.4 Ablation Study

We examine the effects of different exploration strategies, namely, EXPLORE and MIXER, on our approach and report it in Table 5. The QR models in the experiments are initialized with QReCC. Both EXPLORE and MIXER improve performance, although the combination does not outperform MIXER-only settings. We assume this is because the benefit of MIXER offsets the contribution of EXPLORE.

EXPLORE helps to explore more by sampling multiple candidates of question rewrites, and improves F1 scores by 2.5 for CoQA and 1.2 for QuAC. On the other hand, MIXER teaches the QR model to copy the first m tokens from the original question to limit the exploration space and stabilize the training process, resulting in a 3.6 F1 score gain in CoQA and 2.9 F1 score gain in QuAC. Combined, MIXER and EXPLORE offset the benefits of each other. To further improve the

Algorithm	CoQA		QuAC	
	F1	EM	F1	EM
PPO	81.1	73.3	64.7	49.4
+ EXPLORE	83.8	75.9	65.9	50.3
+ MIXER	**84.7**	**76.6**	**67.6**	**51.3**
+ MIXER + EXPLORE	84.2	76.2	67.5	51.2

Table 5: Effects of EXPLORE and MIXER in our framework. Both EXPLORE and MIXER benefit performance, while the combination does not outperform MIXER-only.

performance, we plan to seek an adequate balance of exploration and exploitation since we believe more exploration can boost the performance but stabilizing the training is challenging according to our observation.

5.5 Effects of Decoding Strategies

We explore the effect of decoding strategies for generating question rewrites for inference. We find the greedy search significantly worsens the performance by around 3 to 6 F1 score loss. While the performance steadily improves along with the increase of the beam size, we can not see non-trivial improvement when the beam size is equal or larger than three. We also examine the different sets of hyperparameters, for example, repetition penalty, sampling (combination of temperature, top-k, and top-p) approaches. As reported in Table 6, using smaller or no repetition penalty yields better results. We observe that sampling methods only alter the results marginally. We assume that beam search works satisfactorily because the optimal rewritten questions are more or less predictable similar to machine translation (Yang et al., 2018; Murray and Chiang, 2018).

6 Conclusion and Future Work

In this paper, we propose a reinforcement learning framework for CQA that a QR model that acts as an agent and a QA model as an environment. Our experiments show that the QR model learned to paraphrase questions into a more suitable format for the QA model by reward signal obtained from the CQA performance. Since our exploration is conducted with limited combinations of QA/QR model structures and datasets, we plan to explore the other combinations to justify our approach. Moreover, it would be beneficial to train the QR model without the dialogue history to enforce the QR model and make the question more self-contained. If we can minimize the contribution of the dialogue history in CQA, we can treat the CQA task as a single-turn QA task, and it enormously expands possible solutions for the CQA.

Repetition penalty	Evaluation	
	F1	EM
1.0	67.37	51.46
1.1	**67.47**	**51.40**
1.3	67.12	51.12

Table 6: Using smaller or no repetition penalty tends to yield better results.

Ethical Considerations

This work is not related to any specific real-world application. All the datasets used in our experiments are collected by crowdsourcing (Anantha et al., 2021a), especially through Amazon Mechanical Turk (Reddy et al., 2019; Choi et al., 2018; Elgohary et al., 2019a), and they are publicly available. As the nature of the task, the data collection of CQA and QR is done anonymously and does not involve any privacy or intellectual property concern.

Acknowledgement

We thank the anonymous reviewers for their valuable comments. This work has been partially supported by the China NSFC Project (No. NSFC21EG14), Hong Kong PhD Fellowship Scheme, Research Grant Council, Hong Kong (PF18-25016, PF20-43679), and School of Engineering PhD Fellowship Award, the Hong Kong University of Science and Technology.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021a. 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.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021b. 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.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. 2021. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. 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 3110–3120, Hong Kong, China. 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*.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

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

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019a. 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.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019b. 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.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Di He, Hanqing Lu, Yingce Xia, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2017. Decoding with value networks for neural machine translation. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. In *International Conference on Learning Representations*.

Etsuko Ishii, Genta Indra Winata, Samuel Cahyawijaya, Divesh Lala, Tatsuya Kawahara, and Pascale Fung. 2021. ERICA: An empathetic android companion for covid-19 quarantine. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 257–260, Singapore and Online. Association for Computational Linguistics.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3985–4003, Online. Association for Computational Linguistics.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *arXiv preprint arXiv:1909.10772*.

Y Keneshloo, T Shi, N Ramakrishnan, and CK Reddy. 2019. Deep reinforcement learning for sequence-to-sequence models. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2469–2489.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *arXiv preprint arXiv:2004.01909*.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. XPersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019b. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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 1297–1303, Hong Kong, China. Association for Computational Linguistics.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020a. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2372–2394, Online. Association for Computational Linguistics.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020b. Plug-and-play conversational models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2422–2433, Online. Association for Computational Linguistics.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A simple but effective method to incorporate multi-turn context with bert for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. 2020. Reducing non-normative text generation from language models. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 374–383, Dublin, Ireland. Association for Computational Linguistics.

Victor Petrén Bach Hansen and Anders Søgaard. 2020. What do you mean 'why?': Resolving sluices in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7887–7894.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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*.

Saichethan Reddy. 2020. Detecting tweets reporting birth defect pregnancy outcome using two-view CNN RNN based architecture. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 125–127, Barcelona, Spain (Online). Association for Computational Linguistics.

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.

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, Los Alamitos, CA, USA. IEEE Computer Society.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading. *arXiv preprint arXiv:2203.00343*.

Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. 2019. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 203–211, Hong Kong, China. Association for Computational Linguistics.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, and Pascale Fung. 2020. CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

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.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2021a. [link].

Yan Xu, Etsuko Ishii, Genta Indra Winata, Zhaojiang Lin, Andrea Madotto, Zihan Liu, Peng Xu, and Pascale Fung. 2021b. Caire in dialdoc21: Data augmentation for information seeking dialogue system. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 46–51.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546.

Yi-Ting Yeh and Yun-Nung Chen. 2019. Flowdelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 86–90.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14506–14514.

Jing Zhao, Junwei Bao, Yifan Wang, Yongwei Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. RoR: Read-over-read for long document machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1862–1872, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

Hongyin Zhu, Prayag Tiwari, Ahmed Ghoneim, and M Shamim Hossain. 2021. A collaborative ai-enabled pretrained language model for aiot domain question answering. *IEEE Transactions on Industrial Informatics*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.