Conversational QA Dataset Generation with Answer Revision

Seonjeong Hwang
Graduate School of Artificial Intelligence
POSTECH, Pohang, South Korea
seonjeongh@postech.ac.kr

Gary Geunbae Lee*
Computer Science and Engineering
Graduate School of Artificial Intelligence
POSTECH, Pohang, South Korea
gblee@postech.ac.kr

Abstract

Conversational question–answer generation is a task that automatically generates a large-scale conversational question answering dataset based on input passages. In this paper, we introduce a novel framework that extracts question-worthy phrases from a passage and then generates corresponding questions considering previous conversations. In particular, our framework revises the extracted answers after generating questions so that answers exactly match paired questions. Experimental results show that our simple answer revision approach leads to significant improvement in the quality of synthetic data. Moreover, we prove that our framework can be effectively utilized for domain adaptation of conversational question answering.

1 Introduction

Conversational question answering (CQA) involves answering questions by considering a given text as well as previous conversations. To facilitate research on CQA, a range of datasets have been proposed in recent years (Choi et al., 2018; Reddy et al., 2019; Campos et al., 2020; Anantha et al., 2020; Adlakha et al., 2022). However, building a robust CQA system for a specific domain requires a large-scale domain-specific dataset; moreover, obtaining such a dataset is considerably expensive and time-consuming.

To resolve this issue, in our previous study, we had proposed a conversational question–answer generation (CQAG) framework that automatically creates multiturn question–answer (Q–A) pairs from given passages (Hwang and Lee, 2021). The framework is a two-stage architecture that adopts contextual answer extraction (CAE) and conversational question generation (CQG). Considering previous conversations, the CAE module extracts the next question-worthy phrase from the passage, and then the CQG module generates the conversational question corresponding to the phrase. However, the framework has the limitation that the error may propagate to the question generation stage and even to data generation for subsequent turns if improper answers are extracted by the CAE module.

In this paper, we introduce a CQAG framework with answer revision (CQAG-AR), in which the conversational question generation with answer revision (CQG-AR) module revises the extracted answer to a more suitable one immediately after generating a question. In experiments, we synthesize CQA data using CQAG-AR and then evaluate CQA models trained on these synthetic data. Results reveal that answer revision by the CQG-AR module leads to absolute improvements of up to 13.4% and 15.3% in F1 score and exact match (EM), respectively, for the CQA models. Furthermore, fine-tuning the Wikipedia-domain CQA model on different synthetic data increases EM by up to 13.1%, showing that our framework is beneficial for CQA domain adaptation.

2 Related Work

CQG aims to create conversational questions based on input text. It can be subdivided into answer-aware and answer-unaware approaches. Answer-aware CQG generates conversational questions corresponding to prepared answers (Gao et al., 2019). Gu et al. (2021) exploited accumulated representations of previous conversations to generate the current question by successively encoding answers and questions that constitute conversation history. By contrast, answer-unaware CQG synthesizes conversational questions without given answers (Wang et al., 2018; Pan et al., 2019; Qi et al., 2020). Further, Nakanishi et al. (2019) introduced a framework that first finds the location of points of interest in the passage, identifies question types, and subsequently generates conversational questions.

CQAG attempts to automatically construct CQA data for various domains. In our previous study

*Corresponding author

Proceedings of the 29th International Conference on Computational Linguistics, pages 1636–1644
October 12–17, 2022.

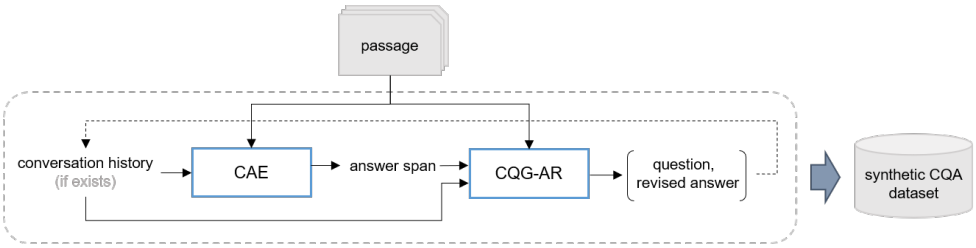

Figure 1: Overview of CQAG-AR. Synthetic Q–A pairs are used as conversation history to generate the subsequent Q–A pairs (dotted line).

(Hwang and Lee, 2021), we designed a vanilla CQAG that generated multiturn Q–A pairs based on a given passage in an autoregressive manner. However, the framework has a drawback in that the validity of the extracted answer directly affects the quality of the conversation.

3 Methods

Figure 1 illustrates a CQAG-AR generation pipeline. To generate a question q_t and answer a_t for the t-th turn of conversations, our framework obtains a passage p and conversation history $h_t = ((q_1, a_1), ..., (q_{t-1}, a_{t-1}))$ as inputs. The CAE module extracts a probable answer span a_t^s considering these inputs. Next, the CQG-AR module generates a conversational question q_t and revised answer a_t^r given the inputs and the extracted answer span a_t^s. Finally, we use the revised answer a_t^r as the answer a_t. The modules do not employ the conversation history to synthesize the Q–A pair for the first turn of conversations. t is omitted in all notations in the following description.

3.1 Contextual Answer Extraction

From a given passage p, the CAE module extracts an answer span a^s that is most likely to be of interest to a questioner, considering the conversation history h, i.e., $P(a^s \mid p, h)$. We implemented the module using a pretrained BERT (Devlin et al., 2018) with two fully connected (FC) layers at the top (Hwang and Lee, 2021). Each FC layer predicts the index of start and end tokens of the potential answer span in the passage:

$$prob_i^s = \mathrm{Softmax}(\mathrm{FC}^s(\mathrm{BERT}(p, h)))[i],$$
$$prob_j^e = \mathrm{Softmax}(\mathrm{FC}^e(\mathrm{BERT}(p, h)))[j],$$

where $prob_i^s$ ($prob_j^e$) represents the probability for the i-th (j-th) token in the passage being the start (end) token of the answer span.

During generation, the top k answer candidates whose start and end indices are i and j are extracted according to the sum of probabilities $prob_i^s + prob_j^e$. The CAE module outputs the answer span with the highest sum of probabilities after deduplicating the candidates compared with the answers used in the conversation history. If the candidate set is empty after deduplication, generation is terminated. To train the module, we computed the sum of cross-entropy losses between predicted start and end indices and the ground truth indices.

3.2 Conversational Question Generation with Answer Revision

Considering the input passage and conversation history, the CQG-AR module generates a conversational question and then revises the answer span that is extracted by the CAE module. The module first considers that the extracted answer span is proper for use as an answer and modifies it if it is inappropriate. To enable this process, we collected training examples of passage p, conversation history h, answer span a^s, and revised answer a^r, which contained positive (*proper* a^s, a^r) and negative (*improper* a^s, a^r) pairs. The module can preserve proper answers extracted by the CAE module by learning positive examples. Additionally, negative examples teach the module how to correct improper answer spans with better answers to the generated questions. We devised two negative sampling techniques to collect improper answer spans from proper ones.

3.2.1 Generating Negative Samples

For the positive examples, we set ground truth answers of the CQA dataset (e.g., QuAC (Choi et al., 2018)) to both *proper* a^s and a^r. The main experiments were conducted with CoQA (Reddy et al., 2019), which contains free-form answers paired with rationales extracted from passages. To obtain proper answer spans from CoQA, answer spans with the highest F1 score compared to the free-form answer from the rationale were extracted. The

improper a^s was obtained from the *proper* a^s by using the following techniques.

Answer Span Expansion If the extracted answer contains several key phrases, it may be unsuitable as an answer for a single question. In addition, unnecessary words around the answer span should be discarded if they are extracted together. To cover these cases, we generated the *improper* a^s by additionally connecting surrounding words of random length to the front or the rear of the *proper* a^s. However, if the sample was extended to phrases that were answers of other Q–A pairs, the model could confuse the target a^r. Therefore, we ensured that the sample did not overlap with answers for other turns.

Answer Span Reduction Some important words that constitute a complete answer may be omitted when extracting the answer span. This phenomenon may risk creating invalid Q–A pairs. To create these types of *improper* a^s, we removed a random number of words from both ends of the *proper* a^s. Examples of negative sampling are included in Appendix A.

3.2.2 Modeling

When the passage p, conversation history h, and answer span a^s are given, the CQG-AR module sequentially generates the conversational question q for the input answer span a^s and then revises the answer span a^s based on the generated question q:

$$P(q, a^r|p, h, a^s) = \prod_{i=1}^{<q>} P(q_i|p, h, a^s, q_{1:i-1}) \times \prod_{j=1}^{<a^r>} P(a^r_j|p, h, a^s, q, a^r_{1:j-1}),$$

where a^r denotes the revised answer and $< \cdot >$ indicates the length of the element.

We implemented the CQG-AR module using T5 (Raffel et al., 2019). We focused on the masked self-attention mechanism of Transformer (Vaswani et al., 2017), where the decoder utilizes knowledge of previously generated tokens to predict the current token. To revise the answer in a form that is more natural to the question, the module outputs the modified answer immediately after the question is generated. The answer span was highlighted using a special token so that the content of interest could be effectively recognized in question generation and answer revision. To train our module, we computed the cross-entropy loss between the question and answer of the ground truth and the module's prediction.

4 Experiments

4.1 Experimental Setup

We employed CoQA (Reddy et al., 2019), which comprised 8k passages collected from seven different domains and human-annotated conversations. To investigate whether CQAG-AR could be effectively utilized to construct a CQA system in a new domain, we split the data into *in-domain* (Wikipedia) and *out-of-domain* (children's stories, literature, news, and middle and high school English exams). The *in-domain* data were used to train CQAG frameworks, and the quality of synthetic data generated by the trained CQAG frameworks was evaluated using the *out-of-domain* data.

In addition, we used QuAC (Choi et al., 2018) and DoQA (Campos et al., 2020) to evaluate our framework. QuAC is based on 13k Wikipedia passages, and DoQA comprises passages collected from FQAs of three practical domains. Because the other two domains constituted only the test set, we used only the *cooking* domain of DoQA in our experiment. The CQAG frameworks used in experiments could generate only open-ended types of data. Therefore, closed-ended (yes and no) and unanswerable types of examples were excluded from the datasets; these were denoted by CoQA^D, QuAC^D, and DoQA^D.

4.2 Baseline Frameworks

To evaluate the quality of the synthetic data generated by our framework, we used two baseline CQAG frameworks:

CQAG-Chain ChainCQG[1] (Gu et al., 2021) is a state-of-the-art answer-aware CQG model, and CQAG-Chain combines the CAE module of CQAG-AR and ChainCQG as a CQG module.

Vanilla CQAG (Hwang and Lee, 2021) This framework shares the same CAE module with CQAG-AR but adopts a simple T5-based CQG module. The CQG module accepts the same input elements as CQG-AR but generates only conversational questions.

[1]The original ChainCQG accepted a rationale-highlighted passage as an input element but we highlighted an answer span in the passage in our experiment. In addition, we implemented the model based on the original source code: https://github.com/searchableai/ChainCQG.

Training data		In-domain	Out-of-domain			
		Wikipedia	News	Mid/High Sch.	Literature	Children's Sto.
Synthetic	CQAG-Chain	71.3 / 59.6	69.2 / 56.9	64.1 / 51.4	59.4 / 47.6	63.1 / 47.6
	Vanilla CQAG	71.3 / 59.9	67.6 / 55.7	65.8 / 52.7	60.3 / 48.3	66.6 / 50.5
	CQAG-AR (ours)	**83.1 / 73.8**	**81.0 / 71.0**	**74.4 / 63.2**	**71.7 / 61.0**	**75.2 / 61.9**
Human-annotated		85.8 / 76.4	86.3 / 75.9	79.0 / 67.6	79.0 / 67.8	82.5 / 70.1

Table 1: F1 (%) and EM (%) scores of CQA models on the CoQA^D test set for each domain (The highest performances among results for synthetic data are shown in bold.)

4.3 CQA with Synthetic Datasets

In this section, we evaluate synthetic data generated by CQAG-AR and baseline frameworks by conducting the CQA task. In the first experiment, CQAG frameworks learned the *in-domain* data of CoQA^D and then generated synthetic CQA data based on the passages extracted from *in-domain* and *out-of-domain* data. Note that we constructed synthetic training and validation sets from original splits of CoQA. We implemented a simple CQA model using T5 (Raffel et al., 2019) and trained several CQA models using human-annotated data (CoQA^D) and synthetic datasets. The training details, an example of synthetic conversations, and statistics of the synthetic data can be found in Appendix B and C.

Table 1 presents F1 and EM scores of CQA models, which learned the synthetic data, on the test set[2] of CoQA^D. The CQA models derived from CQAG-AR outperformed other models, showing significant margins of 11.8% and 13.9% for in-domain data, and average margins of 10.5% and 12.5% for out-of-domain data in F1 and EM, respectively, compared with those derived from the vanilla CQAG. These results demonstrate that the answer revision approach is considerably beneficial in terms of generating valid CQA datasets. However, we additionally found that the out-of-domain CQA models showed lower performances than the in-domain models across all CQAG frameworks. This result implies that CQAG frameworks are less robust when extracting valid CQA data from out-of-domain passages.

Training data	#Training examples	F1
Human-annotated	3.7k	45.1
Synthetic	3.7k	51.5
(CQAG-AR)	4.7k	53.1

Table 2: CQA performances on the test set of DoQA^D.

[2] https://github.com/google/BIG-bench

In Table 2, we compare the evaluation results on the test set of DoQA^D (cooking) for CQA models trained on human-annotated data (DoQA^D) and synthetic data. We obtained the synthetic data by training CQAG-AR using QuAC^D and then generating the data from the passages of DoQA training and validation sets. According to the results, the CQA model trained on synthetic data, which has the same number of examples with the human-annotated data, significantly outperformed the model trained on human-annotated data. Moreover, our framework generated a larger number of examples than the ones in the original DoQA^D, which improved the F1 score of the CQA model by 1.6%. In particular, examples in QuAC^D, which were used to train CQAG-AR, were irrelevant to the cooking domain. This result indicates that our framework effectively creates synthetic CQA data for an unseen cooking domain.

4.4 Human Study

In Table 3, we classify the synthetic examples according to revision types. The distribution shows that 65.2% of answer spans extracted by the CAE module were preserved without any modification, while the other 34.8% of answers were revised. This demonstrates that the CQG-AR module could recognize invalid answer spans and selectively modify the answers. Furthermore, we found that the module could perform more complex revisions such as "multiple revision" and "complete change" though the CQG-AR module learned only examples for "reduction" and "expansion" obtained by negative sampling.

Further, we conducted human evaluation to compare the quality of synthetic data generated by the vanilla CQAG and our CQAG-AR. From the two synthetic datasets presented in Table 1, 120 examples (30 examples from each out-of-domain dataset) were sampled and three volunteers were asked to rate 80 examples according to the criteria listed in Appendix D.

Revision type	Passage	Q-A	Percentage
Preservation	... It covers and has a population of 2.06 million. It is a parliamentary republic and a member of the United Nations, European Union, and NATO. ...	Q: How many people live in Slovenia? A: 2.06 million	65.2%
Reduction	Buckinghamshire (or), abbreviated Bucks, is a county in South East England which borders Greater London to the south east, Berkshire to the ...	Q: Where is it located? A: South East England	15.3%
Expansion	... The group hired Frederick G. Kilgour, a former Yale University medical school librarian, to design the shared cataloging system. ...	Q: Who was he? A: a former Yale University medical school librarian	14.2%
Multiple revision	... Discogs currently contains over 8 million releases, by nearly 4.9 million artists, across over 1 million labels, contributed from nearly 346,000 contributor ...	Q: And how many labels? A: over 1 million	2.0%
Complete change	... Selective breeding for fast growth, egg-laying ability, conformation, plumage and docility took place over the centuries, and modern breeds ...	Q: How did breeds change over time? A: selective breeding	3.4%

Table 3: Distribution of the answer-revision types in the CQG-AR module. (The answer spans extracted by the CAE module are highlighted.)

		Vanilla CQAG	CQAG-AR
Question Connectivity	Dependent	67.9%	66.7%
	Independent	27.7%	30.6%
	Unnatural	4.5%	2.8%
Answer Correctness	Correct	64.2%	85.8%
	Partially correct	23.3%	4.2%
	Incorrect	12.5%	10.0%

Table 4: Results of human evaluation for synthetic data generated by vanilla CQAG and CQAG-AR.

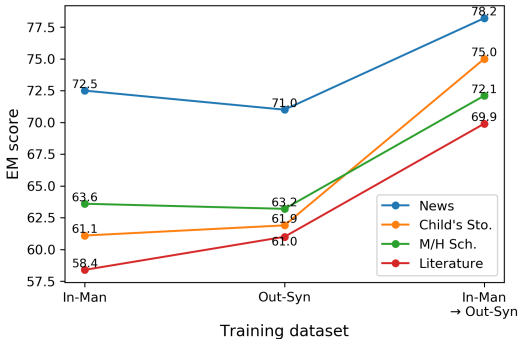

Figure 2: EM scores of several CQA models on the CoQA^D test set for each domain.

Although 2.9% more synthetic questions of CQAG-AR are independent of the previous conversations than the questions of the vanilla CQAG in Table 4, the synthetic questions of the two frameworks show almost similar evaluation results. These results prove that the vanilla CQAG, which performs only question generation, and CQAG-AR, which performs question generation and answer revision in an end-to-end manner, can generate questions of similar quality. However, 21.6% more synthetic answers of CQAG-AR were judged as correct answers compared with those of the vanilla CQAG. Furthermore, the number of partially correct answers was considerably reduced through answer revision. This reveals that answer revision is effective in correcting inappropriate answer spans extracted by the CAE module into correct answers that match well with the question.

4.5 Domain Adaptation

In this experiment, we tested the effectiveness of the synthetic data generated by CQAG-AR in adapting the CQA model from the Wikipedia domain to new domains (*out-of-domain*). We trained CQA models, which were initialized with parameters of T5-Large, in three training settings. In the **In-Man** setting, we trained the CQA model on the Wikipedia data of CoQA^D. In the **Out-Syn** setting, CQA models learned out-of-domain synthetic data. Finally, the model of In-Man setting was fine-tuned with synthetic data of each out-of-domain case in the **In-Man** $\rightarrow$ **Out-Syn** setting.

As shown in Figure 2, the models in the Out-Syn setting yielded results similar to those of the model in the In-Man setting while exhibiting better EM scores in the two domains. Notably, fine-tuning the Wikipedia CQA model using our synthetic data (In-Man $\rightarrow$ Out-Syn) improved the EM scores of the model by an average of 9.7% across all domains. This result indicates that our framework can be effectively utilized for domain adaptation in CQA.

5 Conclusion

In this paper, we propose CQAG-AR, which automatically synthesizes high-quality CQA data. Our framework comprises CAE and CQG-AR modules. Considering the conversation history, the CAE module extracts a question-worthy phrase from a given passage, and then the CQG-AR module generates a conversational question while revising the extracted answer to make it more suitable. Experimental results show that CQAG-AR outperforms baseline frameworks in terms of generating high-quality CQA data. In addition, fine-tuning a Wikipedia-domain CQA system on our synthetic data for out-of-domain cases improves the model performances by significant margins.

Acknowledgements

We would like to thank Professor Yunsu Kim (POSTECH) for his expert advice. This work was supported by SAMSUNG Research, Samsung Electronics Co.,Ltd., and also supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2020-0-01789) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation)

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2020. Open-domain question answering goes conversational via question rewriting. *arXiv preprint arXiv:2010.04898*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa–accessing domain-specific faqs via conversational qa. *arXiv preprint arXiv:2005.01328*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yifan Gao, Piji Li, Irwin King, and Michael R Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. *arXiv preprint arXiv:1906.06893*.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. Chaincqg: Flow-aware conversational question generation. *arXiv preprint arXiv:2102.02864*.

Seonjeong Hwang and Gary Geunbae Lee. 2021. A study on the automatic generation of conversational qa corpora. In *Annual Conference on Human and Language Technology*, pages 133–138. Human and Language Technology.

Mao Nakanishi, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2019. Towards answer-unaware conversational question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 63–71.

Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019. Reinforced dynamic reasoning for conversational question generation. *arXiv preprint arXiv:1907.12667*.

Peng Qi, Yuhao Zhang, and Christopher D Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. *arXiv preprint arXiv:2004.14530*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. *arXiv preprint arXiv:1805.04843*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Examples of Negative Sampling

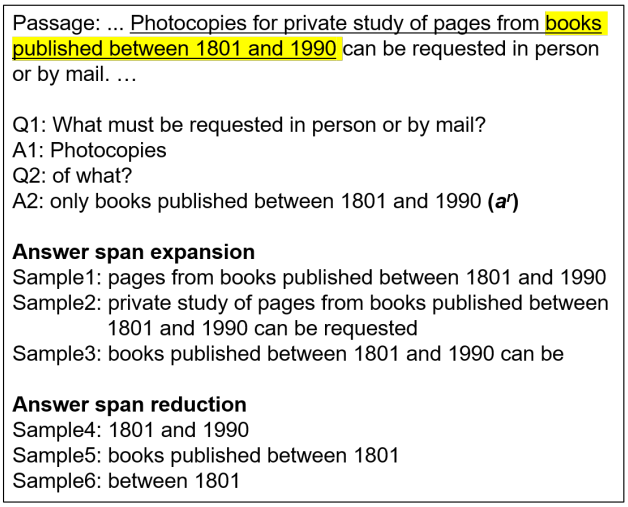

Figure 3: Example of negative sampling applied to one passage in CoQA. The rationale for A2 is underlined, and *proper* a^s corresponding to A2 is highlighted in the passage. A2 is a^r for both the *proper* and *improper* a^s samples.

B Training Details

To implement the CAE module, we used parameters of BERT-large-uncased. Only the previous two pairs of Q–A were used as the conversation history to extract the i-th answer span and passages p longer than 512 tokens were truncated with a stride of 128 tokens:

$$\textbf{Input:}\ [\text{CLS}]\ q_{i-2}\ a_{i-2}\ q_{i-1}\ a_{i-1}\ [\text{SEP}]\ \text{truncated}\ p\ [\text{SEP}]$$

For the CQG-AR module, we initialized the module with parameters of T5-large. The input and target sequences for generation of the i-th Q–A pair are as follows:

$$\textbf{Input:}\ a_i^s\ \text{highlighted}\ p\ [\text{SEP}]\ [\text{A}]\ a_{i-4}\ [\text{Q}]\ q_{i-4}\ \ldots\ [\text{A}]\ a_{i-1}\ [\text{Q}]\ q_{i-1}[\text{A}]\ a_i^s$$
$$\textbf{Target:}\ [\text{Q}]\ q_i\ [\text{A}]\ a_i^r\ [\text{EOS}]$$

Because the input sequence length of the T5 encoder was limited, we truncated the passage p at the 32nd token after the location of a_i^s. Regarding the conversation history, only the previous four Q–A pairs were used. Special tokens ([Q] and [A]) were added before each question and answer, and [A] and a_i^s were appended to the end of the input sequence. [Q] was used as a bos token, and answer generation started when [A] was returned after predicting the question.

We utilized the Transformers library and pre-trained parameters from HuggingFace[3] and conducted experiments using A100 GPUs. Further, AdamW was used as the optimization algorithm with a batch size of 4 and a learning rate of 3e-5. In addition, a learning rate scheduling algorithm was applied and the warm-up period was set to the initial 10% of the total steps. For CAE, we optimize the module based on F1 score between predicted answer span and the ground truth. For CQG-AR, beam search with a beam size of 4 was used during data generation. The best module was selected based on METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang et al., 2019) on the synthetic development set.

For CQA, we designed a simple T5-based model that accepted the concatenation of the passage, conversation history, and question as input and then generated the answer to the input question. We initialized our model with T5-Large and trained the model with AdamW, setting the batch size between 4 and 8 and the learning rate to 3e-5. When fine-tuning the Wikipedia-domain CQA model with synthetic data in Section 4.5, we fine-tuned the model for one epoch, with a batch size of 1 and the learning rate between 1e-7 and 1e-6. We employed the same training and decoding strategies used for the CQG-AR module.

[3] https://huggingface.co/

C Synthetic Data

C.1 Example of Synthetic Conversation

Passage: CHAPTER IV. Signor Andrea D'Arbino, searching vainly through the various rooms in the palace for Count Fabio d'Ascoli, and trying as a last resource, the corridor leading to the ballroom and grand staircase, discovered his friend lying on the floor in a swoon, without any living creature near him. Determining to avoid alarming the guests, if possible, D'Arbino first sought help in the antechamber. He found there the marquis's valet, assisting the Cavaliere Finello (who was just taking his departure) to put on his cloak. While Finello and his friend carried Fabio to an open window in the antechamber, the valet procured some iced water. This simple remedy, and the change of atmosphere, proved enough to restore the fainting man to his senses, but hardly–as it seemed to his friends–to his former self. They noticed a change to blankness and stillness in his face, and when he spoke, an indescribable alteration in the tone of his voice. "I found you in a room in the corridor," said D'Arbino. "What made you faint? Don't you remember? Was it the heat?" Fabio waited for a moment, painfully collecting his ideas. He looked at the valet, and Finello signed to the man to withdraw. "Was it the heat?" repeated D'Arbino. "No," answered Fabio, in strangely hushed, steady tones. "I have seen the face that was behind the yellow mask." "Well?" "It was the face of my dead wife." "Your dead wife!" "When the mask was removed I saw her face. Not as I remember it in the pride of her youth and beauty–not even as I remember her on her sick-bed–but as I remember her in her coffin."

Conversation

Q1: Who was searching for Fabio d'Ascoli?
A1: Signor Andrea D'Arbino
Q2: Where was he searching?
A2: the palace
Q3: What was his last resort?
A3: the corridor leading to the ballroom and grand staircase
Q4: What did he find?
A4: lying on the floor in a swoon
Q5: Where did he seek help first?
A5: the antechamber
Q6: Who helped him?
A6: the marquis's valet
Q7: Who helped him put on his cloak?
A7: the Cavaliere Finello
Q8: What did the valet give him?
A8: some iced water
Q9: What change did his friends notice?
A9: a change to blankness and stillness in his face
Q10: What did D'Arbino say was the cause?
A10: heat
Q11: What was behind the mask?
A11: the face of my dead wife
Q12: What did I see?
A12: her face
Q13: How did I remember her?
A13: in her coffin

Table 5: Samples of generated Q-A pairs using CQAG-AR from a Wikipedia passage in CoQA. Answer spans before revision are highlighted in the passage in order.

C.2 Statistics of Synthetic Data

	Synthetic dataset	CoQA
#Words in question	5.6	5.4
#Words in answer	3.0	2.6
#Turns per passage	12.1	15.1

Table 6: Average number of words in the questions and answers, and the average number of conversation turns in CoQA and our synthetic data extracted from CoQA passages.

Domain	#Passages		#Q–A pairs	
	Train	Dev	Train	Dev
Wikipedia	1.6k	0.1k	23.6k	1.5k
News	1.7k	0.1k	21.6k	1.2k
Mid/High Sch.	1.7k	0.1k	22.2k	1.3k
Literature	1.6k	0.1k	17.7k	1.1k
Children's Sto.	0.6k	0.1k	6.3k	1.2k

Table 7: Statistics summarizing the synthetic datasets generated from CoQA passages.

D Criteria for Human Evaluation

Question Connectivity	
Dependent	The current question refers to previous conversations (e.g., via pronoun usage or ellipses).
Independent	The current question is not dependent on previous conversations.
Unnatural	The current question has grammatical errors or overlaps with previous conversations.
Answer Correctness	
Correct	Questions are paired with correct answers.
Partially correct	Answers are incomplete or contain unnecessary information.
Incorrect	Not the correct answer to the question.

Table 8: Criteria for human evaluation of synthetic CQA data.