CoHS-CQG: Context and History Selection for Conversational Question Generation

Xuan Long Do[1,2,*]**, Bowei Zou**[2]**, Liangming Pan**[3]**, Nancy F. Chen**[2]**, Shafiq Joty**[1,4]**, Ai Ti Aw**[2]

[1]Nanyang Technological University, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]National University of Singapore, [4]Salesforce AI Research
{xuanlong001@e.ntu, srjoty@ntu}.edu.sg
{zou_bowei, nfychen, aaiti}@i2r.a-star.edu.sg
liangmingpan@u.nus.edu

Abstract

Conversational question generation (CQG) serves as a vital task for machines to assist humans, such as interactive reading comprehension, through conversations. Compared to traditional single-turn question generation (SQG), CQG is more challenging in the sense that the generated question is required not only to be meaningful, but also to align with the occurred conversation history. While previous studies mainly focus on how to model the flow and alignment of the conversation, there has been no thorough study to date on which parts of the context and history are necessary for the model. We argue that shortening the context and history is crucial as it can help the model to optimise more on the conversational alignment property. To this end, we propose *CoHS-CQG*, a two-stage CQG framework, which adopts a *CoHS* module to shorten the context and history of the input. In particular, *CoHS* selects contiguous sentences and history turns according to their relevance scores by a *top-p* strategy. Our model achieves state-of-the-art performances on CoQA in both the answer-aware and answer-unaware settings. Our work will be publicly available at https://github.com/dxlong2000/CoHS-CQG.

1 Introduction

One of the key goals of AI is to build systems that can understand and assist humans through conversation. In conversations, asking questions is an important dialogue act that serves as an important communication skill for AI models to better interact with humans (Allen et al., 2007). Taking it a step further, asking good questions could facilitate collecting users' intentions and feedback, starting a new topic, and enhancing the interactivity and persistence of dialogues. In NLP, this line of research is formulated as the task of *Conversational*

---

* Contribution during the internship at Institute for Infocomm Research.

If you were a pilot flying a plane...television interview.
After 104 seconds, Li was able to bring the fighter to a safe landing. The plane only suffered slight damage.
"You are a hero! Congratulations!" Li's teacher, Xu Yongling wrote in a text message to him after the landing. He said Li was a cool-headed pilot...

Q1: Who was flying the plane?
A1: Li Feng.
...
Q11: Did he jump out of the craft?
A11: No.
Q12: What did he do?
A12: Bring the fighter to a safe landing.
Q13: Was it almost completely destroyed?
A13: No.
Q14: Who taught him?
A14: Xu Yongling.
Q15: Did he say anything to Li?
A15: Yes.

*Ground-truth Q16: **What?***
*A16: **You are a hero! Congratulations!***

Figure 1: A dialog sample in CoQA validation set, which reflects not all the sentences in the context and history turns are necessary for generating Q16.

Question Generation (CQG), which aims to generate questions based on the conversation history (Pan et al., 2019a; Nakanishi et al., 2019).

Although question generation has been explored intensively (Pan et al., 2019b; Lu and Lu, 2021), most existing studies focus on single-turn question generation, which aims to generate one question from a given context. However, in the scene of conversation, it poses an additional challenge of multi-turn question generation, in which the model is required to generate multiple questions during the conversation, and the generated questions should be coherent and form a smooth conversation flow.

Despite the intensive exploration of single-turn QG, less attention has been drawn on CQG. Previous work of CQG mostly focuses on solving two main challenges: coreference alignment and conversation flow. Gao et al. (2019) proposed CFNet to model coreference alignment and conversation flow explicitly. Gu et al. (2021) proposed ChainCQG, a two-stage model with two modules: the Answer Encoder learns the representation of the context and answer in each turn, and the Question Genera-

tion learns the representation of the conversational history and generates the next turn's question. However, most previous work makes use of all the context and conversation history indiscriminately. On the contrary, we argue that not all sentences in the context, and not all previous turns in the conversation history are necessary for the model to generate the question in the next turn, and they may even harm the generation capacity of the model. Figure 1 shows such an example, where we see that only the blue parts of the context and history are necessary for generating the 16-th turn's question.

To address the above concerns, we introduce *CoHS-CQG*, a two-stage CQG model, as described in Figure 2. In the answer-aware stage, we input sentences in the context and the conversation history turns into a pretrained sentence-transformer (Reimers and Gurevych, 2019) for calculateing the relevance scores of the (`sentence, history turn`) pairs. Context and History Selection (*CoHS*) module (Section 3.1) is then employed to shorten the context and conversation history concurrently by selecting *top-p* {sentence, history turn} pairs of contiguous sentences in the context, and contiguous previous turns in the history. The shortened context and history are then fed into a T5-based (Raffel et al., 2020) question generation model to generate the questions. By training the model on the shortened context and history, we observe that generated questions are generally more aligned with the conversation, which reflects that the model is optimised better in the conversational alignment. Our model achieves state-of-the-art results in the CQG answer-aware setting on both automatic evaluation metrics and a careful human evaluation. In the answer-unaware stage, we propose a pipeline approach (Section 3.2) to leverage our model on the answer-unaware setting, which also achieves the state-of-the-art performance on human evaluation.

In summary, our main contributions are: *(1) CoHS-CQG*, a two-stage CQG framework for both answer-aware and answer-unaware settings, which adopts a novel module, *CoHS*, to shorten the context and history before inputting them to the QG model. *CoHS* can be plugged into any CQG model, which makes it easily reproducible, *(2)* new strong state-of-the-art performances on answer-aware and answer-unaware CQG, and *(3)* a thorough analysis and evaluation about the selection capacity of *CoHS*.

2 Related Work

2.1 Single-turn Question Generation

Single-turn Question Generation (SQG) has been focused extensively through the years. Early studies relied on syntactic transformation to convert declarative sentences to questions (Heilman and Smith, 2010; Khullar et al., 2018). Recently, Du et al. (2017) showed the limitations of such rule-based methods and formulated the question generation problem as a sequence-to-sequence task. The task is generally cast into two main streams: answer-aware and answer-unaware.

In the answer-aware setting, the target answer is revealed to SQG models. The models then have to solve the task by either treating the answer as an extra input feature or encoding the answer by a separate network (Pan et al., 2019b). However, the answer is not available in the answer-unaware case. Traditional approaches in this setting include two main steps: answer-span selection and answer-aware question generation (Du and Cardie, 2017; Subramanian et al., 2018). Recent state-of-the-art systems in answer-aware setting (Dong et al., 2019; Qi et al., 2020b; Lelkes et al., 2021; Murakhovs'ka et al., 2021) and in answer-unaware setting (Scialom et al., 2019; Lopez et al., 2020) all rely on transformer-based architectures, and they are commonly evaluated on SQuAD (Rajpurkar et al., 2016).

2.2 Conversational Question Generation

Despite the intensive exploration in both settings of the single-turn QG task, there is much less exploration in Conversational Question Generation (CQG). Most of the previous studies focus on the answer-unaware setting (Pan et al., 2019a; Nakanishi et al., 2019; Qi et al., 2020a), but a limited number of works are in the answer-aware setting. In general, there are two main challenges in CQG: coreference alignment and conversation flow. Models in the answer-aware setting then have been proposed to solve those problems such as CFNet (Gao et al., 2019), by which the coreference alignment and conversation flow are modeled explicitly, and ChainCQG (Gu et al., 2021), which contains two modules: the Answer Encoder (AE) module learns the representation of the context and answer span in each turn, and the Question Generation (QG) module learns the representation of the conversational history and generates the next turn's question.
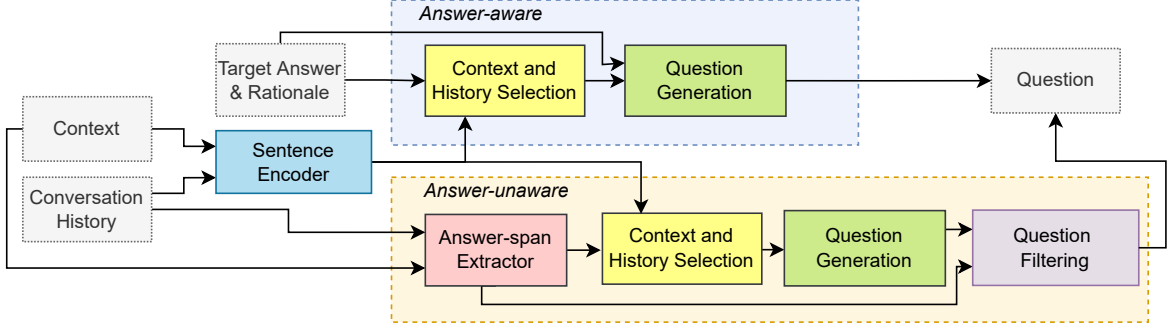
Figure 2: An overview of our proposed framework *CoHS-CQG*. The modules with the same color have the same functionality.

3 CoHS-CQG

We formulate the conversational question generation (CQG) task in two different settings: *answer-aware* and *answer-unaware*. For the *answer-aware* CQG, given the referential context $C = \{c_1, c_2, ..., c_m\}$ where c_i is the i-th sentence in context, the conversation history $H_n = \{(q_1, a_1), (q_2, a_2), ..., (q_{n-1}, a_{n-1})\}$, where (q_i, a_i) is the i-th turn question-answer pair in conversation, the target answer a_n, and the rationale r_n, as input $\mathcal{D}_n^a = \{C, H_n, a_n, r_n\}$, the model then learns to generate the question q_n. The rationale r_n is an associated text span from the context which contains or explains the given answer a_n. For the *answer-unaware* CQG, however, given $\mathcal{D}_n^u = \{C, H_n\}$, the model learns to generate the current question q_n without a_n and r_n.

Our proposed CoHS-CQG framework is shown in Figure 2. The context C and conversation history H_n are first fed into a Sentence Encoder (SE) to compute the relevance scores. In the answer-aware setting, the relevance scores, together with a_n and r_n are input to the Context and History Selection (CoHS) for selecting the parts of C and H_n that are most relevant to the current generation turn, and they are then input to the Question Generation (QG) module. In the answer-unaware case, since a_n is unavailable, C and H_n are first fed into the Answer-span Extractor (AE) to extract a_n, and a_n is later verified by the Question Filtering (QF) module.

3.1 Answer-aware CQG

Sentence Encoder (SE) Given the context C and conversation history H_n, we employ a pretrained sentence-transformer (Reimers and Gurevych, 2019) to embed each sentence c_i in C, and each question-answer pair (q_j, a_j) in H_n (i.e. the concatenation of q_j and a_j), respectively. We then

compute a relevance matrix $\mathbf{T} \in \mathbb{R}^{|C| \times |H_n|}$ as

$$\mathbf{T}[i][j] = rel(c_i, (q_j, a_j)) = \frac{\mathbf{a}_i \mathbf{b}_j}{|\mathbf{a}_i||\mathbf{b}_j|}, \quad (1)$$

where $\mathbf{a}_i$ and $\mathbf{b}_j$ are the embeddings of c_i and $concat(q_j, a_j)$, respectively, the relevance score $rel(\cdot)$ is defined as the *cosine* similarity, and $1 \leq i \leq m, 1 \leq j \leq n-1$.

Context and History Selection (CoHS) To generate the current question q_n, existing CQG models (Gao et al., 2019; Gu et al., 2021) commonly take the full context C and all the previous question-answer pairs H_n as input. Moreover, in leveraging conversation history, some studies (Ohsugi et al., 2019; Zhao et al., 2021) have begun to consider how to select historical information related to the current utterance, but only simply selected the last k turns. We argue that not all parts of the context and conversation history are necessary for the model to generate the current question since the topic in a conversation may shift. On the contrary, introducing irrelevant parts worsens the performance of the model (See Table 1 and 4). To address this problem, we propose a *top-p* CoHS strategy that dynamically selects the most relevant sentences in the context concurrently with the most relevant preceding conversation utterances.

Given the input $\mathcal{D}_n^a = \{C, H_n, a_n, r_n\}$ and the relevance matrix $\mathbf{T}$, CoHS aims to select the *top-p* of sentences and QA pairs from C and H_n, respectively. Inspired by Holtzman et al. (2020), we formulate our *top-p* CoHS strategy as, finding the subset $C_{sub} = \{c_{v-u}, c_{v-u+1}, ..., c_{v-1}\}$ and $H_{sub} = \{(q_{n-k}, a_{n-k}), (q_{n-k+1}, a_{n-k+1}), ..., (q_{n-1}, a_{n-1})\}$,

to satisfy

$$minimize(u + k) \tag{2}$$

$$\sum_{i=v-u}^{v-1} \sum_{j=n-k}^{n-1} \mathbf{T}[i][j] \geq p \tag{3}$$

$$(q_{n-1}, a_{n-1}) \in H_{sub}, c_s \in C_{sub} \tag{4}$$

where p is a given threshold, and c_s is the sentence that contains r_n. First, the optimizing goal is to minimize the sum of $u + k$, where u and k are the numbers of the contiguous sentences from C and contiguous preceding conversation turns from H_n, respectively (Eq.(2)). Then, the sentences and conversation turns with higher similarity than the threshold p are selected as the candidates for building C_{sub} and H_{sub} (Eq.(3)). In addition, since the sentence containing the ground-truth rationale c_s and the last previous conversation turn (a_{n-1}, q_{n-1}) are intuitively relevant for generating the current question, we set two constraints in Eq.(4). Note that the contiguity of C_{sub} and H_{sub} is necessary due to the integrity and coherence of input. The advantage of the heuristic *top-p* CoHS strategy is that CQG models can dynamically select the most relevant C_{sub} and H_{sub} according to different conversation progress, which well adapts when topic shifting. When $H_n = \varnothing$, we select five sentences around c_s (see Appendix A.1).

Question Generation (QG) We employ a T5 (Raffel et al., 2020) as our question generation model. To fine-turn the T5 on the shortened context and history, we concatenate the input $\mathcal{D}_n^a = \{C, H_n, a_n, r_n\}$ in format: `Answer:` a_n, r_n `Context:` C_{sub} `[SEP]` H_{sub}. The model then learns to generate the target question q_n.

3.2 Answer-unaware CQG

In Section 3.1, we utilize 1) the ground-truth previous conversation history H_n, and 2) the ground-truth current answer a_n and rationale r_n, to verify how well the model performs in generating the current question q_n. However, in a more realistic scenario such as a dialogue system, it is necessary to verify whether the model has a good ability to generate questions continuously, that is, the coherence and fluency of the generated questions. To this end, we propose an *answer-unaware* process as shown in Figure 2, including *Answer-span Extractor*, *CoHS* (depicted in Section 3.1), *QG* (depicted in Section 3.1), and *Question Filtering*.

Answer-span Extractor (AE) First, we treat the earliest sentence in the context as the current rationale r_n such that r_n does not contain any rationales of previous turns. Then, a T5 model is trained on SQuAD (Rajpurkar et al., 2016) to predict the target answer span (a) given its original sentence in context (r). We use the model to extract a_n from r_n. Note that we remove the answer spans that are the same as those of previous turns, to ensure that the generated questions are informative enough. Finally, we obtain a set of selected candidate answer spans $A_n = \{a_1^*, a_2^*, ..., a_t^*\}$. Each $a_i^* \in A_n$, together with r_n, and $\mathcal{D}_n^u$ are fed into the CoHS and QG modules to generate the candidate question q_i^*.

Question Filtering (QF) Under the answer-unaware setting, since the conversation history is not manually-labeled, we observe that one type of the common errors is that the generated question may not be answerable by the given context, or its answer may not the provided target answer a_i^*. To address this issue, we train a T5 model on CoQA (Reddy et al., 2019) to answer the generated question q_i^*, and only accept q_i^* if the predicted answer is the same as a_i^*.

4 Experimentation

4.1 Experimental Settings

Dataset We conduct experiments on CoQA (Reddy et al., 2019), a large-scale CQA dataset including 8k conversations. Each conversation contains a referential context and multiple question-answer pairs. In total, there are 127k question-answer pairs collected via Amazon Mechanical Turk. The key characteristics of this dataset are its factoid questions (*i.e.* What? Where? When? How long?) and free-form answers. Since the test set of CoQA is unavailable, we randomly sample 10% of the original training set as our new *validation set*, and keep the original validation set as our *test set* so that future works can be compared with us.

Baseline Models We use a T5$_{base}$ (220M) as our *CoHS-CQG*'s backbone. For the answer-aware baselines, we reimplement CFNet (Gao et al., 2019), an effective CQG framework. We also fine-tune a T5$_{base}$ (Raffel et al., 2020) and a BART$_{base}$ (Lewis et al., 2020), the SOTA transformer-based generation models, on CoQA. For the answer-unaware baseline, we compare with the SOTA framework ReDR (Pan et al., 2019a).

Model	ROUGE-L	B1	B2	B3	B4	METEOR	BERTScore
CFNet	41.25	34.24	22.71	16.57	12.39	27.76	91.43
ChainCQG*	42.22	35.54	26.03	19.63	14.54	30.97	92.54
BART$_{base}$	44.77	35.86	26.32	19.84	15.09	31.60	92.95
T5$_{base}$	45.80	39.09	29.04	22.17	17.03	34.09	93.07
T5$_{base}$ + dyn-HS (p = 0.5)	48.64	40.83	30.74	23.64	18.18	36.49	93.43
T5$_{base}$ + dyn-CS (p = 1)	49.69	41.62	31.44	24.29	18.72	37.42	93.61
CoHS-CQG (Ours, p = 5)	**49.91**	**42.10**	**31.86**	**24.65**	**19.11**	**37.76**	**93.65**

Table 1: Automated evaluation results on our test set (i.e. CoQA validation set). *dyn-HS* and *dyn-CS* are *dynamic History Selection* and *dynamic Context Selection* respectively (Section 4.4). B1 to B4 denotes BLEU 1-4.

Implementation Details We initialise *CoHS-CQG* with pretrained checkpoints from Huggingface (Wolf et al., 2020). We use AdamW (Loshchilov and Hutter, 2019) with the warmup ratio of 0.1 and the initial learning rate of 1e-4. We train the model for 100k iterations with standard window size of 512, and use a Beam search decoding strategy with beam size of 4.

Evaluation Metrics We compute the standard n-gram-based similarity metrics, which are commonly used for text generation, including ROUGE-L (Lin, 2004), BLEU (1-4) (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). We compute BLEU 1-4 by *corpus_bleu* function from NLTK library.[1] We compute ROUGE-F scores in our evaluations by Python implementation of *rouge-score* library.[2] We also calculate BERTScore (Zhang et al., 2020), a similarity score between the generated and ground-truth texts by using deep contextualized embeddings.

Human evaluation is also important to the CQG task since the CQG model may generate the question for the following turn in multiple ways, given the target answer. As such, we conduct human evaluation on both the answer-aware setting and answer-unaware setting.

4.2 Automatic Evaluation

Table 1 shows the automatic evaluation results. We observe that *CoHS-CQG (p = 5)* achieves state-of-the-art performance on all the automatic evaluation metrics. In particular, we derive 3 observations. First, *CoHS-CQG* improves its original baseline T5$_{base}$ significantly. Second, comparing to only dynamically selecting previous turns (T5$_{base}$ + dyn-HS) or sentences in the context (T5$_{base}$ + dyn-CS),

[1] https://www.nltk.org/
[2] https://github.com/google-research/google-research/tree/master/rouge

Model	Flu.	C-Align	Ans.
Answer-aware			
BART$_{base}$	2.38	2.23	2.34
T5$_{base}$	2.72	2.44	2.54
CoHS-CQG (Ours)	**2.74**	**2.58**	**2.60**
Krippendorff's α	0.80	0.71	0.77
Answer-unaware			
ReDR	1.06	1.06	1.05
CoHS-CQG (Ours)	**2.70**	**2.41**	**2.73**
Krippendorff's α	0.84	0.85	0.82

Table 2: Human evaluation results on the validation set of CoQA. **Top**: answer-aware, on 100 random generated samples; **Bottom**: answer-unaware, on 20 random conversations. "Krippendorff's α" shows the inter-annotator agreement. *Flu.*: Fluency, *C-Align*: Conversational Alignment, *Ans.*: Answerability.

p value	Avg. #S	Avg. #P	ROUGE-L	BLEU-4
1	3.18	1.89	48.96	18.19
2	3.73	2.49	49.38	18.61
3	4.27	2.95	49.41	18.65
5	5.29	3.67	**49.91**	**19.11**
7	6.13	4.25	49.35	18.69
10	8.12	4.92	49.18	18.52
∞	16.09	7.97	45.80	17.03

Table 3: Average #Sentences, #Prev. Turns at different p values of preprocessed contexts and histories by *CoHS* of CoQA. $p = \infty$ means selecting the full context and history. *Avg. #S*: average number of sentences, *Avg. #P*: average number of previous turns.

CoHS-CQG achieves better performances, which indicates that dynamically selecting both is more effective. Third, with the threshold of relevance $p = 5$ (Eq.(3)), the *CoHS* module shortens the context to around 5 sentences and the history to 3 previous turns on average (Table 3), by which it achieves the best performance.

We also compare our *CoHS-CQG* with the current SOTA answer-aware CQG model, ChainCQG (Gu et al., 2021) which contains two GPT-2 (Rad-

ford et al., 2019) blocks. Since the provided codes from the authors are incomplete, and the reported results of ChainQCG in (Gu et al., 2021) are on the authors' own test set (they splitted 10% of the training set to become their own test set), we were not able to reproduce the results. Thus, we reimplement ChainCQG (denote it as ChainCQG*). We can see that *CoHS-CQG* outperforms ChainCQG* on all automatic evaluation metrics significantly.

4.3 Human Evaluation

Evaluation Setup We further conduct human evaluation to validate the results. In answer-aware case, we randomly select 100 generated questions associated with the context and conversation history. In answer-unaware case, however, since there is no ground-truth history, simply evaluating 100 random generated samples may not be a fair comparison. Thus, we first select 20 random contexts in our test set. For each context, since the number of turns generated by our model and the competing one, ReDR (Pan et al., 2019a), may not be the same, we heuristically select the first five generated turns from each model's output to compare, resulting in 100 samples in total. We hire three annotators who are English native speakers. Each annotator was instructed to rate the generated questions on a 1-3 scale (3 for the best) based on three criteria: **(1) Fluency** measures not only the grammatical correctness but also the meaning, and factual correctness of generated questions, **(2) Conversational Alignment** measures the alignment of generated questions with the given conversation, **(3) Answerability** measures whether the generated questions are well answerable or not. We measure the annotators' agreement by Krippendorff's alpha (Krippendorff, 2011). Our rating system is described in Appendix A.2.

Observations The top of Table 2 shows the averages of human scores over three annotators in the answer-aware setting. We derive two main observations. First, there is a significant improvement in the *Conversational Alignment* of *CoHS-CQG* compared to T5, which indicates that with the shortened context and history as input, the model learns to focus and align with the given conversation history much better. Second, compared to T5, there is also a slight increase in the *Answerability*, which further shows that the quality of the generated questions is improved. There is also a minor improvement in the *Fluency*, which is reasonable because T5

#Pre. turns	ROUGE-L	BLEU-4
1	48.14	17.43
2	**48.34**	17.66
3	48.21	**17.68**
4	47.77	17.64
5	47.15	17.59
6	46.90	17.12
Full history	45.33	16.73

Table 4: Performance of $T5_{base}$ with different fixed number of previous turns on our validation set.

commonly generates fluently, grammatically and meaningfully correct questions. Our annotators have a good overall agreement with an alpha coefficient of 0.76.

The bottom of Table 2 shows the human evaluation for the answer-unaware setting. First, we observe that ReDR has low *Fluency* score due to most of the generated questions are factually wrong or have no meaning associated with the given context. It also has low scores on the other two metrics as the generated questions are frequently repetitive. Second, the generated questions by *CoHS-CQG* are generally high-quality, fluent and answerable as they already passed the *Question Filtering* module. The annotators achieve a good overall inter-agreement with an alpha coefficient of 0.83.

4.4 Effects of Context and History Selection

We further conduct the studies about the performance of T5 when we dynamically select the context sentences or the previous turns but not both of them concurrently. In this section, we formulate these two problems as below.

Dynamic Context Selection In this setting, we follow the previous studies on CQA (Ohsugi et al., 2019; Zhao et al., 2021) to select the last k previous turns. The results are shown in Table 4. Since $T5_{base}$ achieves the best performance on BLEU-4 by using the last 3 previous turns, we adopt this setting in the following independent context selection experiments. Given the context C, answer a_n, rationale r_n, and the last k previous turns $H_{sub} = \{h_{n-k}, h_{n-k+1}, ..., h_{n-1}\}$, $h_i = concat(q_i, a_i)$, we formulate the context selection problem in this section as finding the smallest sub-

p value	Avg. #Sentences	ROUGE-L	BLEU-4
0	1	48.15	17.88
1	4.35	**49.69**	**18.72**
2	8.28	49.47	18.43
3	11.51	48.45	18.30
4	13.42	48.80	18.16
∞	16.09	48.31	18.01

Table 5: Average number of selected sentences from context (fixing the last 3 previous turns) of each *p* value (left) and performance of "T5$_{base}$ + *dyn-CS*" on our test set (right). $p = \infty$ and $p = 0$ denote utilizing the full context and only the sentence that contains the rationale respectively.

p value	Avg. #History Turns	ROUGE-L	BLEU-4
0	0	43.91	15.32
0.5	2.57	**48.64**	**18.18**
1	4.11	48.27	18.07
2	6.32	47.14	17.70
3	7.45	46.74	17.32
∞	7.97	45.80	17.03

Table 6: Average number of selected history turns from conversation (with the full context) of each *p* value (left) and performance of "T5$_{base}$ + *dyn-HS*" on our test set (right). $p = \infty$ and $p = 0$ denote utilizing the full history and none of the history respectively.

set $C_{sub} = \{c_{v-u}, c_{v-u+1}, ..., c_{v-1}\}$, to satisfy:

$$\sum_{x=v-u}^{v-1} \sum_{y=n-k}^{n-1} T[x][y] \geq p \qquad (5)$$

$$c_s \in C_{sub} \qquad (6)$$

where p is a given threshold, and c_s is the sentence that contains r_n. We name this model as T5$_{base}$ + *dyn-CS* in Table 1 where *dyn-CS* stands for dynamic Context Selection.

Table 5 shows how different values of threshold p (Eq.(5)) affects the selection and the performance of the model. We observe that with a fixed number of previous turns $k = 3$, $p = 1$ gives us the best performance on ROUGE-L and BLEU-4. By setting threshold $p = 1$, and fixed 3 previous turns, the *CoHS* module selects around 4 sentences in each context sample on average. The result indicates that selecting more contexts does not lead to better performance, which is consistent with our motivation.

Dynamic History Selection In this setting, we follow most previous works on CQA (Ohsugi et al., 2019; Zhao et al., 2021) and CQG (Gao et al., 2019; Gu et al., 2021) to use the whole context C, and then we dynamically select different numbers of previous turns. We formulate the history selection problem as finding the smallest subset $H_{sub} = \{(q_{n-k}, a_{n-k}), (q_{n-k+1}, a_{n-k+1}), ..., (q_{n-1}, a_{n-1})\}$, to satisfy:

$$\sum_{x=1}^{m} \sum_{y=n-k}^{n-1} T[x][y] \geq p \qquad (7)$$

$$(q_{n-1}, a_{n-1}) \in H_{sub} \qquad (8)$$

where p is a given threshold, and c_s is the sentence that contains r_n. We name this experiment as T5$_{base}$ + *dyn-HS* in Table 1 where *dyn-HS* stands for dynamic History Selection.

Table 6 shows how different values of threshold p (Eq.(7)) affects the selection and performance of the model. We can observe that with the full context, $p = 0.5$ achieves the best performance on both ROUGE-L and BLEU-4. By setting the threshold $p = 0.5$, the *CoHS* module then selects around 3 previous turns on average. This observation is in line with our following results in Table 4 (see Section 4.5), by which we conclude that with different values of fixed number of previous turns, $k = 2$ and $k = 3$ achieve the best results.

4.5 Discussion

Effects of Relevance Threshold p To further understand how the threshold p (Eq.(3)) controls the selection of context and conversation history, we conduct experiments with different values of p. Table 3 shows the average number of the selected sentences and the selected previous turns, together with the performances of T5$_{base}$ on ROUGE-L and BLEU-4. First, we can observe that on average, the difference between the #Sentences and #Pre. Turns is not large for all values of p, which reflects that our *top-p* algorithm does not prioritise selecting long context over short history and vice-versa. This indicates that the relevance scores assist the algorithm to select the context sentences, together with the history turns in a reasonable way. Second, with $p = 5$, T5$_{base}$ yields the best performance, as we discussed in Section 4.2.

Effects of Different Fixed Previous Turns In Table 4, we study with the full context, how the number of previous history turns affect the performance of the model on our validation set. We can observe that with the full context, the settings of previous history turns $k = 2$ and $k = 3$ achieve the best performances on ROUGE-L and BLEU-4, respectively. Compared to the performances in Table 6, it indicates that dynamically selecting instead of fixing the number of previous turns indeed

ID	Context & History and the Selection of CoHS-CQG (p = 5)

Answer: his owner, **Rationale:** the cat had been abandoned by his owner

Context: When my father was dying, I traveled a thousand miles from home to be with him in his last days. It was far more heartbreaking than I'd expected, one of the most difficult and painful times in my life. After he passed away I stayed alone in his apartment. There were so many things to deal with. It all seemed endless. I was lonely. I hated the silence of the apartment. But one evening the silence was broken: I heard crying outside. I opened the door to find a little cat on the steps. He was thin and poor. He looked the way I felt. I brought him inside and gave him a can of fish. He ate it and then almost immediately fell sound asleep. The next morning I checked with neighbors and learned that the cat had been abandoned by his owner who's moved out. So the little cat was there all alone, just like I was. As I walked back to the apartment, I tried to figure out what to do with him. Having something else to take care of seemed _ But as soon as I opened the apartment door he came running and jumped into my arms. It was clear from that moment that he had no intention of going anywhere. I started calling him Willis, in honor of my father's best friend. From then on, things grew easier. With Willis in my lap time seemed to pass much more quickly. When the time finally came for me to return home I had to decide what to do about Willis. There was absolutely no way I would leave without him. It's now been five years since my father died. Over the years, several people have commented on how nice it was of me to rescue the cat. But I know that we rescued each other. I may have given him a home but he gave me something greater.

History: <s> What was crying? <s> a cat <s> what did the author feed it? <s> yes <s> what did the author feed it? <s> fish <s> then what did the cat do? <s> fell asleep <s> where was the author? <s> his father's apartment. <s> was the father alive by then? <s> no <s> how far did the author travel? <s> a thousand miles <s> who did the author check about the cat with? <s> neighbors <s> what did he find out? <s> the cat was abandoned

1

Ground-truth: by who?	*CoHS-CQG:* by who?	*T5:* who abandoned the cat?

Answer: in the morning, **Rationale:** It was here Paul found his brother on the morning of his arrival in London.

Context: CHAPTER VIII ""I AM WEARY OF A HOPELESS LOVE"" Paul and Arthur shared a bachelor residence in Mayfair; shared it, that is to say, insomuch as Paul had purchased it, and was the sole proprietor, and Arthur used it whenever he could get leave from his regiment. It was here Paul found his brother on the morning of his arrival in London. They shook hands in silence; Paul did not wish to say anything for a moment. His brother's appearance had choked him. It was one o'clock, but he was still in his dressing-gown; with sunken, pale cheeks, save for one bright spot, and with faint, dark rims underneath his eyes. There were a pile of blue papers and some ominous-looking envelopes on the table before him, and Paul could not help noticing the intense pallor of the hand which rested upon them. ""I wish you would let a fellow know what time you were coming,"" Arthur said, rather peevishly, but with an attempt at a smile. ""I didn't expect you till evening, so I was having a shack before dressing. I was late last night!"" Paul banished his gravity, as far as possible, and stood with his hands in his pockets, leaning against the mantel-piece. He heartily disliked the part of mentor, and he did not wish to play it, unless he were obliged. ""It was beastly early to get up,"" he said, ""but the connection at Normanton is so much better. One has to wait two hours by the late train, and Normanton is such a hole. I don't know that I should have come up to town at all, just yet,"" he continued after a slight pause, ""only that I'm on the committee at the club this term, you know, and I haven't attended a single meeting yet. Besides, I promised Westover to put him up this time, and the half-yearly meeting's to-morrow, you know. Got any engagement? If not, you might dine with me there. Always a full night election time, you know!""

History: <s> Where was the joint residence? <s> in Mayfair <s> who owns it? <s> Paul <s> who else stayed there? <s> Arthur <s> how often? <s> whenever he was on leave <s> from what? <s> from his regiment. <s> who is his brother? <s> Paul and Arthur were brothers <s> where did they meet up? <s> in the bachelor residence"

2

Ground-truth: when?	*CoHS-CQG:* when?	*T5:* When did Paul find his brother?

Figure 3: Case studies on the CoQA validation set and the results of T5 (with full context and history) and CoHS-CQG (with shorten context and history). The texts of rationales are underlined. The selected texts in context and history by our *CoHS (p = 5)* module are highlighted in blue.

Model	Flu.	C-Align	Ans.
Answer-unaware			
CoHS-CQG w/o AE	2.13	1.76	1.64
CoHS-CQG w/o QF	2.16	1.98	2.02
CoHS-CQG (Ours)	**2.74**	**2.58**	**2.60**
Krippendorff's α	0.82	0.79	0.77

Table 7: Human evaluation results for the ablation studies of AE and QF modules on the validation set of CoQA. "Krippendorff's α" shows the inter-annotator agreement. *Flu.*: Fluency, *C-Align*: Conversational Alignment, *Ans.*: Answerability.

improves the performance.

4.6 Ablation Studies

Ablation of Answer-span Extractor (AE) We conduct an ablation study for the *Answer-span Extractor (AE)* (*CoHS-CQG w/o AE*), in which we replace the predicted answer span a_n with the rationale r_n (a sentence in the context C). The results are shown in Table 7. Note that in this experiment, we also remove the *Question Filtering (QF)*. As expected, the *Answerability* and *Conversational Alignment* drop significantly, which is explainable since the rationale r_n may contain redundant information, thus it is not suitable to be r_n.

Ablation of Question Filtering (QF) We study the ablation of the Question Filtering (*CoHS-CQG w/o QF*), in which we use all the generated questions. The results are shown in Table 7. As we can see, the *Answerability* and *Conversational Alignment* decrease significantly. We observe that without QF, there may have been turns in which the questions are similar with the same answers, which further proves the necessity of this module.

4.7 Case study: Effectiveness of *CoHS*

When carefully studying the performances of T5 and BART, we observe that the key for these models to gain high scores on n-gram automatic metrics, such as BLEU and ROUGE, is *focusing on the history to optimise the conversational alignment*. We argue that, with a long context and the whole conversation history, the input likely tends to distract the attention of the models on the given conversation history. The models in these cases mostly focus on the given answer and rationale to generate the question, rather than highly focusing on the history. Figure 3 lists some of the examples whereby we draw the above conclusion.

Considering the first example in Figure 3, we observe that with the full context and history, the T5 model mostly relies on the rationale "*the cat had been abandoned by his owner*" to generate the question "*who abandoned the cat?*". Although the question is somehow aligned with the given conversation history, it is not close enough to the gold question "*by who?*". The other two examples in Figure 3 are also the same, and we observe a lot of cases that are similar to them. We argue that in order to generate such questions like "*by who?*", intuitively, the model should pay significant attention to the conversation history to optimise the conversational alignment. By inputting to the model the shortened context and history, we can see that the generated questions in Figure 3 by *CoHS-CQG* indeed change, and they are exactly the same as the ground-truth questions. This improvement reflects that training the model such as T5 with the shortened context and history samples indeed guides the model to optimising more on the conversational alignment property instead of just heavily focusing on the target answer and rationale.

5 Conclusion

This paper presents *CoHS-CQG*, a two-stage framework for CQG, which adopts a *CoHS* module to dynamically select relevant context and conversation history for generating the question in the current turn. Experimental results on CoQA demonstrate that the proposed *CoHS-CQG* achieves state-of-the-art performances in both answer-aware and answer-unaware settings. Our extensive analysis and studies show the effectiveness of *CoHS* in improving the CQG models. In future work, we will focus on how to select the contiguous question-worthy content from the paragraph by reasoning.

Acknowledgements

References

James Allen, Nathanael Chambers, George Ferguson, Lucian Galescu, Hyuckchul Jung, Mary Swift, and William Taysom. 2007. Plow: A collaborative task learning agent. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, page 1514–1519. AAAI Press.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Copenhagen, Denmark. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Yifan Gao, Piji Li, Irwin King, and Michael R. Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4853–4862, Florence, Italy. Association for Computational Linguistics.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. ChainCQG: Flow-aware conversational question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2061–2070, Online. Association for Computational Linguistics.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, Los Angeles, California. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Payal Khullar, Konigari Rachna, Mukul Hase, and Manish Shrivastava. 2018. Automatic question generation using relative pronouns and adverbs. In *Proceedings of ACL 2018, Student Research Workshop*, pages 153–158, Melbourne, Australia. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Computing*, 1.

Adam D. Lelkes, Vinh Q. Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. In *Proceedings of the Web Conference 2021*, WWW '21, page 2501–2511, New York, NY, USA. Association for Computing Machinery.

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.

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.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2020. Simplifying paragraph-level question generation via transformer language models.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chao-Yi Lu and Sin-En Lu. 2021. A survey of approaches to automatic question generation:from 2019 to early 2021. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 151–162, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Lidiya Murakhovs'ka, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2021. Mixqg: Neural question generation with mixed answer types.

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, Hong Kong, China. 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, Florence, Italy. Association for Computational Linguistics.

Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019a. Reinforced dynamic reasoning for conversational question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2124, Florence, Italy. Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019b. Recent advances in neural question generation.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, and Christopher D. Manning. 2020a. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 25–40, Online. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020b. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6027–6032, Florence, Italy. Association for Computational Linguistics.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 78–88, Melbourne, Australia. Association for Computational Linguistics.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

A Appendix

A.1 Comparison with Static Context Selection

To compare our dynamic context selection strategy with a static way, we simply select *five* context sentences around c_s in the context $C = \{c_1, c_2, ..., c_m\}$ since Table 5 shows that around more than four sentences on average achieves the best performance. To this end, we consider a simple heuristic method as below. If $3 \leq s \leq m - 2$, $C_{sub} = \{c_{s-2}, c_{s-1}, c_s, c_{s+1}, c_{s+2}\}$; else if $s \leq 2$, $C_{sub} = \{c_1, c_2, c_3, c_4, c_5\}$; and if $s \geq m - 1$, $C_{sub} = \{c_{m-4}, c_{m-3}, c_{m-2}, c_{m-1}, c_m\}$. We then select five sentences in the context by this way and use the whole conversation history. The result shows that the $T5_{base}$ model yields 17.24 of BLEU-4 and 47.02 of ROUGE-L, which slightly outperforms the $T5_{base}$ baseline using the full context in Table 1.

A.2 Human Rating System

In this section, we describe how our annotators are instructed to give the points in three criteria *Fluency*, *Conversational Alignment*, and *Answerability*. There are three main notes. First, *Fluency* measures not only the grammatical correctness, but also measures the meaning and factual correctness of the question with the given context. Second, in the answer-unaware setting, as there is no golden history, we do not define the *Score 2* in the *Conversational Alignment* as in the answer-aware setting. Third, for the *Answerability* criterion in the answer-unaware setting, the target answer and target rationale are unavailable. However, since our approach first selects the rationale, and then extracts the candidate answers from it to generate the questions, we still evaluate the quality of our questions (*Score 2, 3*) with the selected candidate answers by the *Question Filtering* module (Section 3). For the details, see Figure 4.

Criterion	Human Rating System	
Fluency	Answer-aware	**Score 1:** The generated question has no meaning/factually wrong with the information from the context. **Score 2:** The generated question is good, but has a small grammatical error. **Score 3:** The generated question is gramatically correct and factually correct with the information from the context.
	Answer-unaware	**Score 1:** The generated question has no meaning/factually wrong with the information from the context. **Score 2:** The generated question is good, but has a small grammatical error. **Score 3:** The generated question is gramatically correct and factually correct with the information from the context.
Conversational Alignment	Answer-aware	**Score 1:** Generated question is totally irrelevant to the conversation history. **Score 2:** Generated question is aligned to the conversation history, however, it has a differennt meaning with the golden. **Score 3:** Perfect, generated question is aligned to the conversation history and asks about the same as the golden.
	Answer-unaware	**Score 1:** Generated question is totally irrelevant to the conversation history. **Score 3:** Generated question is aligned to the conversation history.
Answerability	Answer-aware	**Score 1:** Generated question is not answerable by the context. **Score 2:** Generated question is answerable by the context, but does not have the answer as the the target answer. **Score 3:** Perfect, gennerated question is answerable by the context and its answer is target answer.
	Answer-unaware	**Score 1:** Generated question is not answerable by the context. **Score 2:** Generated question is answerable by the context, but does not have the answer as the the target answer (target answer is available since we first extract the target answers). **Score 3:** Gennerated question is answerable by the context and its answer is target answer (target answer is available since we first extract the target answers).

Figure 4: Human Rating System