Generating Information-Seeking Conversations from Unlabeled Documents

Gangwoo Kim[1♣*] **Sungdong Kim**[2,3*] **Kang Min Yoo**[2,4] **Jaewoo Kang**[1†]
Korea University[1] NAVER AI Lab[2] KAIST AI[3] NAVER CLOVA[4]
{gangwoo_kim, kangj}@korea.ac.kr
{sungdong.kim, kangmin.yoo}@navercorp.com

Abstract

Synthesizing datasets for conversational question answering (CQA) from unlabeled documents remains challenging due to its interactive nature. Moreover, while modeling *information needs* is an essential key, only few studies have discussed it. In this paper, we introduce a novel framework, SIMSEEK, (**Sim**ulating information-**Seek**ing conversation from unlabeled documents), and compare its two variants. In our baseline SIMSEEK-SYM, a questioner generates follow-up questions upon the predetermined answer by an answerer. On the contrary, SIMSEEK-ASYM first generates the question and then finds its corresponding answer under the conversational context. Our experiments show that they can synthesize effective training resources for CQA and conversational search tasks. As a result, conversations from SIMSEEK-ASYM not only make more improvements in our experiments but also are favorably reviewed in a human evaluation. We finally release a large-scale resource of synthetic conversations, WIKI-SIMSEEK, containing 2 million CQA pairs built upon Wikipedia documents. With the dataset, our CQA model achieves the state-of-the-art performance on a recent CQA benchmark, QuAC (Choi et al., 2018) [1].

1 Introduction

Conversational question answering (CQA) involves modeling the information-seeking process of human dialogue. In the task, systems should understand questions according to the conversational context. To build robust systems, large-scale CQA datasets (Choi et al., 2018; Reddy et al., 2019; Saeidi et al., 2018; Campos et al., 2020) have recently been introduced. Still, they are limited in scale to generalize toward real-world applications,

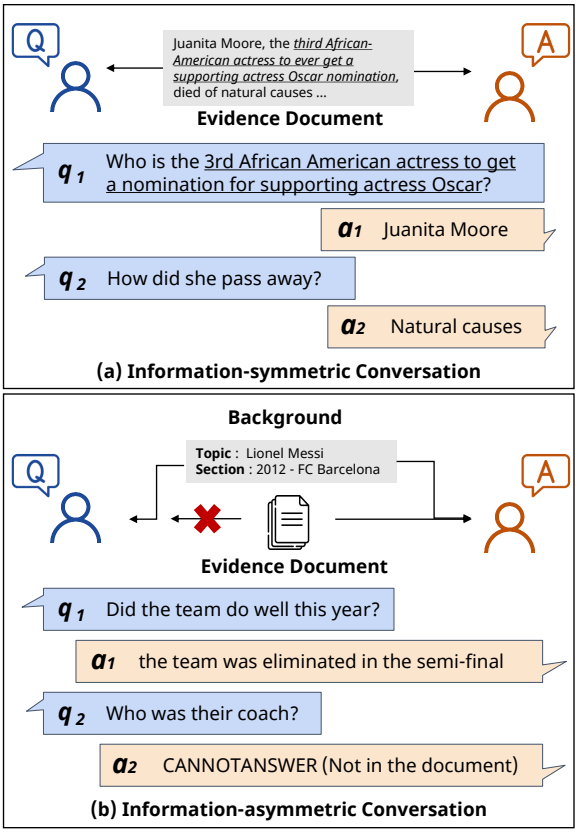


Figure 1: Examples of two conversation scenarios. In the former, questioners with access to the evidence document often ask less related things to the conversation. In the latter, questioners seek new information from the inaccessible document, leading to information-seeking behaviors, i.e., open-ended and unanswerable questions.

which motivates the development of automated methods for constructing CQA datasets.

However, generating CQA datasets is a challenging task, which requires interactions between questioner and answer. Therefore, most of the literature has discussed subparts of the overall process. One line of research in conversational question generation (CQG) aims to generate follow-up questions upon held-out conversations (Pan et al., 2019; Qi et al., 2020; Gu et al., 2021). Another line of research in CQA has greatly enhanced answer accu-

♣ Work done while interning at NAVER AI Lab
* Equal contribution † Corresponding author
[1]The code and dataset are available at github.com/naver-ai/simseek.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2362 - 2378
December 7-11, 2022

racy (Qu et al., 2019b; Kim et al., 2021; Zhao et al., 2021). Despite the recent advances, they assume all other ingredients (i.e., gold history by humans) are given.

Moreover, modeling *information needs* can facilitate simulating realistic conversations. As illustrated in Figure 1 (a), questioners with excessive information often ask questions incoherent with the conversation. On the other hand, the *information needs* drive the questioners to seek new knowledge via conversation, failing to do so sometimes, as shown in Figure 1 (b). However, only few CQA studies have focused on it (Qi et al., 2020).

In this paper, we propose a novel framework, SIMSEEK (**Sim**ulating Information-**Seek**ing conversation from unlabeled documents) and compare its two variants. Both consist of two sub-modules, questioner and answerer, that converse with each other; but each variant assumes opposite scenarios, respectively. In (1) SIMSEEK-SYM, an answerer first identifies answers from the document, and then a questioner asks context-dependent questions based on the predetermined answer. On the contrary, (2) SIMSEEK-ASYM allows a questioner to ask questions without any prior knowledge about the answer. Then, an answerer provides corresponding answers to the asked questions. Either way, SIMSEEK sequentially generates QA pairs at every turn, moving the conversation forward.

We generate synthetic conversations with our frameworks and evaluate them. Despite the similarity of the two frameworks, SIMSEEK-ASYM performs better, which reveals the importance of modeling *information needs*. We first conduct experiments on a challenging CQA benchmark, QuAC (Choi et al., 2018) in the semi-supervised setup. Our experimental results demonstrate the effectiveness of the synthetic dataset from SIMSEEK-ASYM that outperforms other CQA generation baselines. Besides, it also enhances the dense retriever on a conversational search benchmark, OR-QuAC (Qu et al., 2020). We perform a human evaluation to investigate how different conversations are generated by two variants, compared to the original human ones. As a result, conversations from SIMSEEK-ASYM are more favorably reviewed in overall adequacy than others, including humans.

We finally construct a large-scale resource of synthetic conversations, WIKI-SIMSEEK, which contains 2 million CQA pairs built upon 213k Wikipedia passages. Further trained on the dataset, our CQA model achieves state-of-the-art performance on QuAC. We hope it would shed light on building more robust CQA models and identifying the factors for realistic information-seeking conversation.

Our main contributions are summarized as:

- We propose a novel framework SIMSEEK that generates synthetic conversations from unlabeled documents and compare its two variants to provide a deeper understanding of the information-seeking conversation.
- To the best of our knowledge, we are the first to demonstrate the effectiveness of synthetic datasets in two downstream tasks, CQA and conversational search.
- We construct and release a large-scale resource of synthetic conversations, WIKI-SIMSEEK. By leveraging it, we achieve state-of-the-art performance on a challenging CQA benchmark, QuAC.

2 Background

In the information-seeking conversation, two agents (i.e., questioner and answerer) converse about the specific topic. To provide accurate knowledge to the questioner, the answerer can utilize the document that consists of answer-containing passage c and its background knowledge $\mathcal{B}$ (i.e., the title and abstract). Let q_t be the current question and a_t be its corresponding answer at turn t. Formally, CQA systems are required to find correct answer a_t to the question q_t from the passage c based on the conversational history $\mathcal{H}_{<t} = [(q_1, a_1), ..., (q_{t-1}, a_{t-1})]$, i.e. $p(a_t \mid q_t, c, \mathcal{H}_{<t})$.

Typically, most CQG research assumes that the questioner can access the answer-containing passage c. Hence, they formulate the task of generating the question q_t based on the passage c and answer a_t, i.e. $p(q_t \mid c, a_t, \mathcal{H}_{<t})$ (Gao et al., 2019; Pan et al., 2019; Gu et al., 2021). The formulation can be considered a straightforward extension of the dominant paradigm in single-turn QA generation (Puri et al., 2020; Lewis et al., 2021), where models generate questions given their answers. Recently, Qi et al. (2020) suggest a new viewpoint to promote a natural scenario of information-seeking conversation. In the setup, CQG modules are blinded to the answer-containing passage and rely on background information $\mathcal{B}$ when generating conversational question q_t, i.e. $p(q_t \mid \mathcal{B}, \mathcal{H}_{<t})$.

Figure 2: Overview of our frameworks continuing the held-out conversation $\mathcal{H}_{<t}$. To generate a QA pair (q_t, a_t) at current turn t, (a) SIMSEEK-SYM first extracts the answer candidate a_t from the passage c. Then, the questioner generates question q_t that can be answered by a_t. (b) SIMSEEK-ASYM first asks a follow-up question without accessing the passage. The answerer then provides an answer to the question from the evidence passage. Finally, we append the resulting QA pairs to the history $\mathcal{H}_{<t}$ for moving on to the next turn.

3 SimSeek

We newly introduce two opposite ways to simulate synthetic conversations from unlabeled documents, SIMSEEK-SYM and SIMSEEK-ASYM, as illustrated in Figure 2.

3.1 SimSeek-sym

We propose a strong baseline, SIMSEEK-SYM, inspired by the information-symmetric scenario. It can also be viewed as a straightforward extension of QA generation frameworks that are dominant in single-turn QA tasks (Puri et al., 2020; Lewis et al., 2021). The framework is composed of the following components:

1. A *conversational answer extractor* (CAE) to detect answer candidates from the passage, considering the conversation.
2. An *answer-grounded CQG* (CQG_{answer}) to generate conversational questions that are likely to be answered by the detected candidates.
3. A *filtering* CQA model that predicts an answer to the generated question based on the conversation. If the predicted answer is not matched with the predetermined answer by CAE, the QA pair is dropped.

Conversational Answer Extractor The component identifies spans that are likely to become an answer to the probable questions from the passage. The selected span should also be natural to keep the conversational flow. Specifically, the CAE model $p_a^{sym}(a_t \mid \mathcal{H}_{<t}, c)$ calculates the likelihood of answer span a_t and predicts the most likely prediction $\hat{a}_t$ without taking the current question q_t. By the likelihood values, we obtain the set of top-k answer candidates $\hat{A}_t = \{\hat{a}_t^1, \hat{a}_t^2, \ldots, \hat{a}_t^k\}$. By jointly encoding the history $\mathcal{H}_{<t}$ with the passage c, the component could consider conversational flow when extracting $\hat{A}_t$. We adapt 2D span extraction head upon the backbone architecture as Lewis et al. (2021) propose.

Answer-grounded CQG Grounded on each extracted span, the CQG_{answer} generates a follow-up question on the held-out conversation. Thus, it should satisfy multiple objectives at once; generating proper questions for the answer and coherent with the history. Formally speaking, the CQG_{answer} synthesizes the conversational question based on the history, passage, and extracted answer, i.e. $p_q^{sym}(q_t \mid c, a_t, \mathcal{H}_{<t})$. We employ a T5-based sequence-to-sequence model as a backbone of the component (Raffel et al., 2020). In particular, we highlight target answer a_t as rationale span in the passage c with a special token suggested by Gu et al. (2021). In addition, we adopt a mask prediction scheme that aligns its objective with that of the pre-training phase following Chada and Natarajan (2021).

Roundtrip Filtration for CQA The filtering model ensures the quality of generated questions, by checking the roundtrip consistency (Alberti et al., 2019; Puri et al., 2020; Lewis et al., 2021). When the predictions of the filtering model are not matched with the predetermined answer by CAE, the question-answer pairs are discarded. We ease the filtering rule, from exact match to word-level similarity (i.e., F1 score) since the answers in CQA

are often lengthy, compared to those in the single-turn QA. We employ the fine-tuned CQA model as our filtered, i.e., $p_f^{sym}(a_t \mid q_t, c, \mathcal{H}_{<t})$.

3.2 SimSeek-asym

To simulate the information-asymmetric conversation effectively, we introduce a novel framework, SIMSEEK-ASYM (Figure 2 (b)). The framework consists of the following components:

1. A *prior-grounded CQG* (CQG$_{prior}$) for generating conversational questions relying solely on prior information (i.e., background information relevant to the topic).

2. A *conversational answer finder* (CAF) to comprehend the generated question and provides the most acceptable answer to the question from the evidence passage.

Prior-grounded CQG CQG$_{prior}$ asks questions from insufficient information. Hence, the component requires neither the answer at the current turn nor the answer-containing passage. Instead, it generates questions solely based on the background information about the topic, $\mathcal{B}$. Specifically, it models conversational question q_t from the given history $\mathcal{H}_{<t}$ and background $\mathcal{B}$, i.e. $p_q^{asym}(q_t \mid \mathcal{H}_{<t}, \mathcal{B})$.

For a fair comparison of two CQG components, T5-based sequence generator is adopted to implement the CQG$_{prior}$, same with the CQG$_{answer}$. They share the same architecture but their designs differ from each other. We restrict CQG$_{prior}$ from accessing answer-relevant information, encouraging it to learn information-seeking behavior. Although it slightly sacrifices QG performance in the automatic metric (i.e., BLEU), CQG$_{prior}$ plays a crucial role in simulating realistic information-seeking conversations. We also demonstrate a one-to-one comparison by performing an intrinsic evaluation in Appendix B.1.

Conversational Answer Finder The conversational answer finder (CAF) provides an answer to the generated question based on the evidence passage. Its objective is modeling $p_a^{asym}(a_t \mid q_t, c, \mathcal{H}_{<t})$. CAF plays the answerer's role in the information-seeking scenario by providing the requested information from the passage c. Note that any CQA model can be adopted as the CAF component and hence trained on the existing CQA datasets in the same way. The design choice enables SIMSEEK-ASYM to generalize toward other advanced CQA approaches effectively.

3.3 Synthetic CQA from Documents

Our two SIMSEEK frameworks can generate synthetic conversations from unlabeled documents. To train all modules in our frameworks, we use finite amount of human-labeled dataset $\mathcal{D} = \{(\mathcal{B}^i, c^i, \mathbf{q}^i, \mathbf{a}^i)\}_{i=0}^{|\mathcal{D}|}$, where the $\mathbf{q}^i$ and $\mathbf{a}^i$ denote all questions and answers, respectively, up to the maximum turn T in i-th conversation.

In the inference phase, we suppose a unseen corpus $\mathcal{C} = \{(\mathcal{B}^j, c^j)\}_{j=0}^{M}$, which contains total M number of unlabeled documents. Each SIMSEEK framework sequentially synthesizes questions and answers at every turn t, starting from empty history $\mathcal{H}_{<0}$. Specifically, SIMSEEK-SYM is formulated as:

$$p^{sym}(q_t, a_t \mid \mathcal{H}_{<t}, \mathcal{B}, c) \approx p_q^{sym}(q_t \mid a_t, \mathcal{H}_{<t}, c) \cdot p_a^{sym}(a_t \mid \mathcal{H}_{<t}, c)$$

It first narrows down the potential target of the question, constraining the question distribution. Note that it does not consider the filtering process while generating the conversation. Instead, we discard unqualified (q, a) pairs after all conversations are terminated. On the other hand, SIMSEEK-ASYM decomposes the process into:

$$p^{asym}(q_t, a_t \mid \mathcal{H}_{<t}, \mathcal{B}, c) \approx p_a^{asym}(a_t \mid q_t, \mathcal{H}_{<t}, c) \cdot p_q^{asym}(q_t \mid \mathcal{H}_{<t}, \mathcal{B})$$

Contrary to SIMSEEK-SYM, it allows the question distribution to approximate any questions relevant to the topic. Finally, our frameworks generate question and answer at every turn t as:

$$\hat{q}_t, \hat{a}_t = \arg\max_{q_t, a_t} p(q_t, a_t \mid \hat{\mathcal{H}}_{<t}, \mathcal{B}, c)$$

where $\hat{\mathcal{H}}_{<t}$ is a sequence of the generated (q, a) pairs at previous turns and $p(\cdot)$ can be modeled as either $p^{sym}(\cdot)$ or $p^{asym}(\cdot)$. The generated pair $(\hat{q}_t, \hat{a}_t)$ is appended to $\hat{\mathcal{H}}_{<t}$, resulting in $\hat{\mathcal{H}}_{<(t+1)}$. The conversation progresses until it reaches the maximum turn T or satisfies several termination rules [2]. Finally, we obtain sequences of the generated questions $\hat{\mathbf{q}}_j$ and answers $\hat{\mathbf{a}}_j$ by iterating the generation process, which results in the synthetic CQA dataset $\hat{\mathcal{D}} = \{(\mathcal{B}^j, c^j, \hat{\mathbf{q}}^j, \hat{\mathbf{a}}^j)\}_{j=0}^{M}$.

[2] See each termination rule in Sec 4.1 and 6.1, respectively

4 Evaluating Synthetic Conversations

We evaluate our SIMSEEK in the semi-supervised setup. To this end, we train all components on the existing CQA dataset $\mathcal{D}$ first. Then, synthetic conversations are generated upon unseen documents $\mathcal{C}$ by our frameworks. The resulting conversations are used as an additional training resource for downstream tasks. We train task-specific backbones on the synthetic datasets and test their performances in two downstream tasks, CQA and conversational search. See more details in Appendix A, C.

4.1 Experimental Setup

Datasets All baselines and our frameworks are trained on a recent CQA benchmark, QuAC (Choi et al., 2018), which consists of 100k QA pairs for information-seeking conversation. CANARD (Elgohary et al., 2019) convert questions in QuAC into self-contained questions such that they could be understood without the conversation. We construct a single-turn QA dataset by replacing questions in QuAC with them, which is called CANARD in below. For evaluating the quality of synthetic conversations in the conversational search, we use OR-QuAC (Qu et al., 2020). It extends QuAC to the open-domain setup and measures the performance in the passage retrieval task. Further details are described in Appendix A.1.

Semi-supervised Setup We split the original training set of QuAC into three subsets, QuAC$_{seen}$, QuAC$_{unseen}$, and the validation set, following prior works (Elgohary et al., 2019; Qu et al., 2020)[3]. We train all components on QuAC$_{seen}$, considering it as the given dataset $\mathcal{D}$. Then, assuming unlabeled documents of QuAC$_{unseen}$ as an unseen corpus $\mathcal{C}$, we construct synthetic dataset $\hat{\mathcal{D}}$ with the CQA generation frameworks.

We use synthetic datasets detailed in Section 5. In particular, we set the maximum turn T as 6 and do not end the conversation early than that. Models for downstream tasks are trained on either the synthetic set only ($\hat{\mathcal{D}}$) or the merged set ($\mathcal{D}+\hat{\mathcal{D}}$), and tested on the original evaluation set. We report and compare their performances to measure the quality of synthetic conversations. More details are in Appendix A.2

Baselines for Synthetic CQA Generation We introduce solid baselines for synthesizing CQA datasets and compare them with our methods. Since there doesn't exist any prior work available, we simply extend PAQ (Lewis et al., 2021), one of the dominant frameworks in single-turn QA tasks, and then adopt them as our baselines. For PAQ-CANARD baseline, we train its all components on CANARD, regarding it as the single-turn QA task. PAQ-QuAC extends it by adopting CQG$_{answer}$ that can consider the conversation history. The questioners in all baselines also require the target answer when generating questions. But all answerers cannot consider the held-out conversation. More details are in Appendix A.3

Baselines in Downstream Tasks After building synthetic conversations, we train and test the baseline models in downstream tasks. For the CQA task, we choose three backbone architectures, RoBERTa (Liu et al., 2019) in base and large size, and Longformer-large (Beltagy et al., 2020). Longformer architecture has been shown to be effective for encoding much longer history, which achieves competitive performance with the previous state-of-the-art approach (Zhao et al., 2021). In addition, we test synthetic datasets in one of the document retrieval tasks, conversational search, where systems are required to retrieve relevant documents to conversational queries. We employ the off-the-shelf dense retriever, DPR (Karpukhin et al., 2020), as our baseline for conversational search. Further details are in Appendix A.4

4.2 Experimental Results

Semi-supervised CQA We evaluate the effectiveness of our synthetic datasets on a recent CQA benchmark, QuAC. Table 1 shows the end performance of CQA models trained on the resulting datasets. When using the synthetic dataset alone ($\hat{\mathcal{D}}$), PAQ-CANARD shows the lowest performance in all CQA backbones. It implies the difficulty of directly extending single-turn QA methods to CQA. Adopting CQG module that can consider the conversational context (PAQ-QuAC) advances CQA performance with a huge gap of over 15 F1 scores in all backbones. It indicates that learning to comprehend conversational questions is crucial to improving CQA performance. SIMSEEK-SYM, where all components consider the history, shows comparable scores with PAQ-QuAC. Despite the small gap, SIMSEEK-SYM largely outperforms PAQ-QuAC without the filtering process that often saturates the end-CQA performances.

[3] Table 8 shows detailed statistic

CQA Backbone Synthetic CQA Generation	Trained on $\hat{\mathcal{D}}$	 $\mathcal{D}+\hat{\mathcal{D}}$
RoBERTa-base		
None ($\hat{\mathcal{D}}$ = empty)	-	64.4
PAQ-CANARD	38.2	64.3
PAQ-QuAC	55.9	64.6
SIMSEEK-SYM	55.5	64.4
SIMSEEK-ASYM	**62.5**	**65.3**
Human Annot. ($\hat{\mathcal{D}}$ = $\text{QuAC}_{\text{unseen}}$)	65.3	67.5
RoBERTa-large		
None ($\hat{\mathcal{D}}$ = empty)	-	65.6
PAQ-CANARD	38.8	66.5
PAQ-QuAC	51.5	66.6
SIMSEEK-SYM	54.3	66.3
SIMSEEK-ASYM	**64.8**	**67.5**
Human Annot. ($\hat{\mathcal{D}}$ = $\text{QuAC}_{\text{unseen}}$)	65.0	70.3
Longformer-large		
None ($\hat{\mathcal{D}}$ = empty)	-	72.0
PAQ-CANARD	37.5	71.5
PAQ-QuAC	61.7	71.7
SIMSEEK-SYM	60.8	71.7
SIMSEEK-ASYM	**71.5**	**73.1**
Human Annot. ($\hat{\mathcal{D}}$ = $\text{QuAC}_{\text{unseen}}$)	72.3	73.8

Table 1: Comparison over synthetic CQA generation methods. We report F1 scores for the end CQA performance on the development set of QuAC. Frameworks for synthetic CQA generation are trained on the original dataset $\mathcal{D}$ and generate the snythetic dataset $\hat{\mathcal{D}}$. Finally, student CQA baselines are fine-tuned on either $\hat{\mathcal{D}}$ or $\mathcal{D}+\hat{\mathcal{D}}$. "Human Annot." indicates human-labeled conversations from the original QuAC, i.e., $\text{QuAC}_{\text{unseen}}$.

[4] SIMSEEK-ASYM shows dominant performance compared to other baselines. Moreover, the performance gap increases as the size of CQA models gets larger. It implies that SIMSEEK-ASYM could generate finer quality of synthetic conversations when leveraging better CQA models.

Results of ($\mathcal{D}$ + $\hat{\mathcal{D}}$) show augmentation effect of generated datasets. Most of the baselines fail to improve the performance, which implies the difficulty of generating realistic CQA examples. On the other hand, our proposed framework SIMSEEK-ASYM consistently improves CQA performance over all CQA backbones. Specifically, it improves RoBERTa-large and Longformer-large by 1.9 and 1.1 F1 scores compared to the main baseline (None), respectively. Surprisingly, Longformer-large with SIMSEEK-ASYM achieves competitive performance as when trained on the human-labeled dataset, by a gap of only 0.7. It shows SIMSEEK-ASYM succeeds in simulating human-like conversations.

[4] Table 6 provides an ablation study on it

Retrieval Model Synthetic CQA	OR-QuAC MRR	 R@5	 R@20
DPR trained on $\mathcal{D}$ + $\hat{\mathcal{D}}$			
None ($\hat{\mathcal{D}}$ = empty)	53.3	64.8	73.8
SIMSEEK-SYM	50.4	62.4	72.3
SIMSEEK-ASYM			
w/ RoBERTa-base	51.5	63.3	73.6
w/ RoBERTa-large	53.4	64.4	73.6
w/ Longformer-large	**54.4**	**66.1**	**75.3**

Table 2: Evaluation results of conversational passage retrieval on OR-QuAC test set. Longformer-large architecture is used for SIMSEEK-ASYM.

Utility in Conversational Search Table 2 shows the retrieval performances of baseline retrieval model, DPR (Karpukhin et al., 2020) on OR-QuAC dataset (Qu et al., 2020). The resulting conversations from SIMSEEK-SYM degrade the retrieval performance, indicating it fails to model questions in the information-seeking conversation. We report the results of SIMSEEK-ASYM combined with different answerers. Among them, the framework with Longformer-large only succeeds in boosting the retrieval performance of DPR. Although dense retrievers do not encode any answers by the task setup, retrieval performances vary depending on the capabilities of answerer model. It implies that interacting with a better answerer allows the questioner to ask more diverse and adequate questions, leading to a more realistic information-seeking conversation.

5 Analysis

We report detailed statistics of the generated datasets and perform a human evaluation to analyze the quality of conversations and compare them.

5.1 Qualitative Analysis

Table 3 summarizes statistics of synthetic conversations from two frameworks. All datasets show similar overlap score of question-answer, word-level F1 of (q_t, a_t). In contrast, we observe a meaningful gap in the overlap between the question and previous responses, word-level F1 of (q_t, $a_{0:(t-1)}$), which measures how many words from the opponent's responses are reused in the current question. On the other hand, SIMSEEK-SYM more frequently exploits one of the tricks for seeking new informa-

	QuAC	SIMSEEK SYM	SIMSEEK ASYM
tokens / question	6.5	7.3	7.5
tokens / answer	15.1	17.6	16.8
F1 of (q_t, a_t)	7.0	10.2	9.7
F1 of $(q_t, a_{0:(t-1)})$	17.1	19.0	26.9
% Anything else?	18.4	23.8	17.0
% Unanswerable Qs	17.3	1.0	19.7

Table 3: Comparison over the original QuAC and synthetic datasets from our frameworks. SIMSEEK-ASYM uses Longformer-large as the answerer in the table. For scalable analysis, we automatically count "Anything else?" questions with certain strings (e.g., "other" and "else")

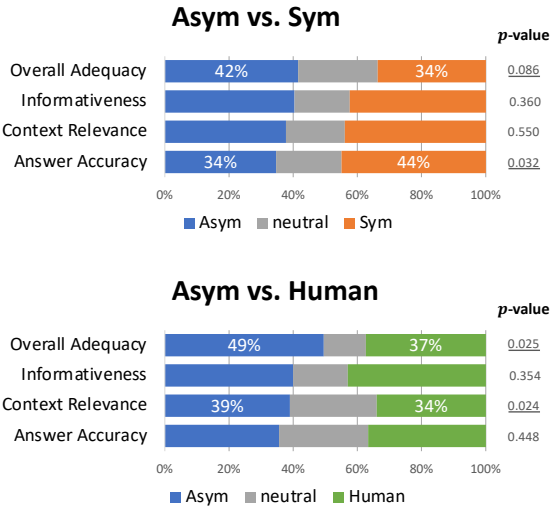


Figure 3: Pairwise human evaluation results for the original QuAC (human) and the synthetic conversations from SIMSEEK (sym and asym). The annotators pick the better one from two random QA pairs in terms of each criterion. We report the overall proportion of majority votes for each instance. We conduct a bootstrap test with 10^5 samples for the difference between pairs. We add labels only when the difference is statistically significant (p-value lower than 0.1).

tion, "Anything else?" questions[5], which shifts the current topic and requests any new information. The questioners ask these questions effortlessly without considering conversational context much. SIMSEEK-SYM rarely asks unanswerable questions. On the other hand, SIMSEEK-ASYM often fails to acquire answers as humans do and their frequencies are similar. SIMSEEK-ASYM generates conversations that have similar statistics to the original QuAC, overall.

5.2 Human Evaluation

We perform a human evaluation to examine the quality of synthetic conversations. Specifically, we conduct a pairwise judgment with Amazon Mechanical Turk, asking the workers to assess the relative quality of follow-up QA pairs. More details are described in Appendix E.

We first ask the workers to judge (1) the overall adequacy of generated QA pairs to the given history. It represents how adequate the QA pair is for continuing the given conversation. Additionally, we also ask (2) informativeness (i.e., does the question try to gather new information), (3) context relevance (i.e., how relevant or specific is the question to the given context), and (4) answer accuracy (i.e., whether the answer is a correct one to the question), inspired by the metrics of Qi et al. (2020); Li et al. (2021); Thoppilan et al. (2022). Figure 3 summarizes the results. We find that there are no significant differences in informativeness. Hence, SIMSEEK-ASYM show similar informativeness scores compared to humans.

SIMSEEK-SYM asks questions closely related to context and answer; however they are rarely adequate. The annotators conclude that conversations from SIMSEEK-SYM are more relevant to the document than SIMSEEK-ASYM. Moreover, answers to their corresponding questions are more accurate than SIMSEEK-ASYM since the questioner asks questions with having their answers. However, SIMSEEK-SYM are less frequently chosen as adequate. In other words, it succeeds in generating coherent QA pairs at each turn, but it is inadequate in simulating information-seeking behaviors.

Conversations from SIMSEEK-ASYM are reviewed as more adequate even than humans, overall. We observe that repeating the opponents' responses often leads to high adequacy[6]. As shown in Table 3, SIMSEEK-ASYM asks questions highly overlapped with previous responses $a_{0:(t-1)}$, which makes it perceived as more helpful and communicative by the annotators. On the contrary, original questions in QuAC often request new knowledge concisely, which seems relatively less enthusiastic. Thus, this leads to the ironic result that human-annotated conversations are less frequently chosen in terms of overall adequacy.

[5] See the example in Table 9.

[6] We further report the qualitative case study in Table 9

Dataset	Domain	Dialogs	Ques.
Single-turn QA			
SQuAD	Wikipedia		107K
Natural Questions	Wikipedia		307K
PAQ	Wikipedia		65M
CQA			
QuAC	Wikipedia (People)	13K	98K
CoQA	7 sub-domains	13K	127K
DoQA	Stack Exchange	2K	10K
Open-Domain CQA			
OR-QuAC	Wikipedia (People)	5.6K	40.5K
QReCC	Wikipedia	14K	81K
TopiOCQA	Wikipedia	4K	50K
Ours			
WIKI-SIMSEEK	Wikipedia	213K	2.1M

Table 4: Comparison over CQA datasets with WIKI-SIMSEEK

CQA Model	QuAC		
	F1	HEQ-Q	HEQ-D
HAE (Qu et al., 2019a)	63.1	58.6	6.0
GraphFlow (Chen et al., 2019)	64.9	-	-
HAM (Qu et al., 2019b)	66.7	63.3	9.5
ExCorD (Kim et al., 2021)	67.7	64.0	9.3
RoR (Zhao et al., 2021)*	75.7	**73.4**	**17.8**
Ours			
Longformer-large	74.0	71.0	13.7
+ WIKI-SIMSEEK	75.0	72.5	13.2
+ CoQA	69.5	63.3	7.7
+ CoQA + WIKI-SIMSEEK	**76.1**	**73.4**	16.4

Table 5: Comparsion over baseline CQA models on the development set of QuAC. By using SIMSEEK-ASYM, WIKI-SIMSEEK is generated from unlabeled documents of Wikipedia . Models with asterisk (*) are additionally trained on CoQA datatset. GraphFlow (Chen et al., 2019) does not report HEQ scores.

6 Wiki-SimSeek

Based on our results, we newly construct synthetic conversations on the larger scale of corpus, Wikipedia, by applying SIMSEEK-ASYM. We finally release a large-scale resource of information-seeking conversations, WIKI-SIMSEEK, which consists of 2.1 million questions and answer pairs upon 213k Wikipedia passages.

6.1 Dataset Construction

First, we crawl the documents from Wikipedia by using KILT tools (Petroni et al., 2021) [7]. Following Choi et al. (2018), we collect Wikipedia articles from a list of category keywords leveraging a web interface, Wikipedia foundation[8]. We use the abstract of Wikipedia documents as background information, while sections between 250 and 550 words are picked as evidence documents. Then, we use SIMSEEK-ASYM with the answerer (Longformer-large) to simulate synthetic conversations upon the crawled Wikipedia documents. In particular, SIMSEEK-ASYM generates the conversations until they reach the twelfth turn or more than three unanswerable questions are asked. Table 4 shows overall statistics of WIKI-SIMSEEK and compares it to other QA datasets. Each dialog contains 10.0 question-answer pairs on average, which shows that the framework can carry on long conversations and bring out new information.

6.2 Further Improvement of CQA Models

To investigate the effect of WIKI-SIMSEEK, we compare our models with previous approaches on QuAC as shown in Table 5. Unlike the semi-supervised setup, all components of our frameworks are trained on the training set of QuAC. Further training Longformer-large on WIKI-SIMSEEK improves the performance by 1.0 of F1 and 1.5 of HEQ-Q. To compare fairly with the previous best performing model RoR (Zhao et al., 2021)[9], we also employ another CQA dataset, CoQA (Reddy et al., 2019) as an additional training resource. Baseline performances are significantly degraded when we use only CoQA for data augmentation. It shows that simply combining two different datasets could cause a distribution shift, leading to a performance drop. However, when WIKI-SIMSEEK is additionally used, it boosts the CQA model performance by a large gap of 1.7 F1 score. It achieves state-of-the-art performance in F1 score on QuAC. As a result, our dataset could be one of the solutions to mitigate the shift, providing further improvement. Note that other baseline approaches could be further improved by our synthetic datasets. Moreover, since SIMSEEK-ASYM is a model-agnostic framework, other CQA models can be adopted as the answerer for generating synthetic datasets.

7 Related work

Conversational Question Answering With the advent of recent large-scale CQA datasets (Choi et al., 2018; Reddy et al., 2019), numerous studies proposed methods to resolve the challenging task. Most works focused on developing model structures (Zhu et al., 2018; Qu et al., 2019a,b) that are specialized in the CQA task. Several works demonstrated the effectiveness of the flow mechanism in

[7] github.com/facebookresearch/KILT
[8] petscan.wmflabs.org
[9] Zhao et al. (2021) also use CoQA to achieve the scores.

CQA (Huang et al., 2018; Chen et al., 2019). Most recently, leveraging self-contained questions (Kim et al., 2021) or encoding longer context (Zhao et al., 2021) have been shown to be effective in the task.

Synthetic QA Generation Many question generation (QG) researches have sparked advances in various QA tasks (Dhingra et al., 2018; Dong et al., 2019; Lewis et al., 2019; Alberti et al., 2019; Puri et al., 2020; Lewis et al., 2021). Most of early studies propose to generate them in a cloze-style (Dhingra et al., 2018; Lewis et al., 2019) or by using pre-defined templates (Fabbri et al., 2020). Recent studies for the synthetic QA generation propose the pipeline strategies composed of three sub-phases, answer extraction, question generation, and various filtering steps such as round-trip filtration (Alberti et al., 2019; Puri et al., 2020; Lewis et al., 2021).

Conversational Question Generation Many works attempt to generate human-like conversational questions. Pan et al. (2019); Gao et al. (2019) introduce the challenge of CQG and successfully extend the single-turn question generation to consider conversational input. Most prior works are based on the information-symmetric assumption (Pan et al., 2019; Gao et al., 2019; Nakanishi et al., 2019; Gu et al., 2021). Recently, Qi et al. (2020) investigate information-asymmetric conversations. They first attempt to generate the conversational questions without evidence document. Concurrently, Dai et al. (2022) propose a method to turn document into dialogue and release a large-scale dataset of synthetic dialogue. However, they report improvements only in the conversational search task.

8 Conclusion

In this work, we propose a novel framework, SIMSEEK, simulating information-seeking conversation from given unlabeled documents. Our frameworks assume two scenarios and compare them to provide a deeper understanding of information-seeking conversation. Experimental result shows that our SIMSEEK-ASYM generates human-like conversation. Moreover, we provide insightful analyses to help understand the information-seeking conversation better. We finally release the large-scale resources of synthetic conversations, WIKI-SIMSEEK. We hope it could be a stepping stone for building robust CQA models that can be generalized toward the real-world scenario. Furthermore, it could be beneficial to identify the factors in realistic information-seeking conversations.

Limitations

We tested various methods for automatically filtering the generated conversations from SIMSEEK-ASYM. However, since it already simulates human-like questions as shown in our evaluation, we failed to significantly improve the performance in downstream tasks. Thus, we only adopt several filtering rules to discard deviated conversations. One might propose a novel filtration method for SIMSEEK-ASYM by investigating our resulting dataset, WIKI-SIMSEEK. WIKI-SIMSEEK, is limited in the specific language (i.e., English). We use machines with 8 V100 GPUs and training Longformer-large on WIKI-SIMSEEK takes a few days, which requires relatively high computational costs.

Acknowledgements

We would like to appreciate Jinhyuk Lee, Hwanhee Lee, Miyoung Ko, Hyunjae Kim, Jaehyo Yoo, Hwaran Lee, Jungwoo Ha, and anonymous reviewers for providing constructive feedback. This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2022-2020-0-01819) and the High-Potential Individuals Global Training Program (RS-2022-00155958) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation). This work was funded by the National Research Foundation of Korea (NRF-2020R1A2C3010638). This work was supported by the Hyundai Motor Chung Mong-Koo Foundation.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Milan 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.

Rakesh Chada and Pradeep Natarajan. 2021. Fewshotqa: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6081–6090.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2019. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. *arXiv preprint arXiv:1908.00059*.

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

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. *arXiv preprint arXiv:2205.09073*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 582–587.

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. *Advances in Neural Information Processing Systems*, 32.

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

Alexander Richard Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513.

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 (ACL)*, pages 4853–4862.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

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*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

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.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

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 (TACL)*, 7:452–466.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Huihan Li, Tianyu Gao, Manan Goenka, and Danqi Chen. 2021. Ditch the gold standard: Re-evaluating conversational question answering. *arXiv preprint arXiv:2112.08812*.

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

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.

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

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick SH Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *NAACL-HLT*.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826.

Peng Qi, Yuhao Zhang, and Christopher D Manning. 2020. 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.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1133–1136.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1391–1400.

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.

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.

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. *arXiv preprint arXiv:1809.01494*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv e-prints*, pages arXiv–2201.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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. *arXiv preprint arXiv:2109.04780*.

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

A Experimental Details

A.1 Datasets

QuAC QuAC (Choi et al., 2018) consists of 100k QA pairs in information-asymmetric dialogues, where a questioner asks questions based on a topic with background information, and an answerer returns the answers in the form of text spans in Wikipedia document. Restricting the questioners from accessing the answer-containing document, the authors encourage them to seek new information on a topic via conversation. Following Choi et al. (2018), we evaluate models with the F1 score for QuAC. Since the test set is only available in the QuAC leaderboard, we evaluate models on the development set[10]. HEQ measures whether a CQA model finds more accurate answers than humans in each granularity (HEQ-Q for question, HEQ-D for dialogue)[11].

OR-QuAC Qu et al. (2020) extend the original QuAC dataset to open-domain setup[12]. It assumes that a ground-truth document is not given in advance, which means the answerers do not know what to be asked before a conversation begins. Instead, they first need to search relevant passages from web-scale documents (about 11M chunked passages) based on the given conversational history and current question. After reading the retrieved passage, they predict an answer to the question. Following the original setup in Qu et al. (2020), we only regard previous questions $\{q_1, q_2, ..., q_{t-1}\}$ as history without answers. Since OR-QuAC is similarly partitioned with our QuAC splits (see details in Table 8), we use same synthetic conversations that are used for CQA task. For evaluation, mean reciprocal rank (MRR), Recall@5 (R@5), and Recall@20 (R@20) are used to evaluate first stage conversational retrieval.

A.2 SimSeek for Semi-supervised Setup

For semi-supervised CQA setup, we set QuAC$_{seen}$ as $\mathcal{D}$ and documents in QuAC$_{unseen}$ as $\mathcal{C}$. The number of turns T is set to 6 in our generations. For OR-QuAC experiment, we follow the semi-supervised setup since OR-QuAC shares the same document split with the semi-supervised setup. All CQG models are based on T5-large (Raffel et al., 2020) model of 770M parameters, and we use 5 for beam size of beam search and 0.98 for top-p value of nucleus sampling (Holtzman et al., 2020) with 1.2 temperature. We employ the same backbone for the CAF with corresponding CQA student models, RoBERTa-base (125M), RoBERTa-Large (355M), and Longformer-large (435M) (Liu et al., 2019; Beltagy et al., 2020).

A.3 Baselines for Synthetic CQA Generation

We introduce strong baselines for synthesizing CQA datasets and compare them with our methods. For a fair comparison, we train all components of approaches on the same labeled dataset $\mathcal{D}$, and generate the synthetic dataset $\hat{\mathcal{D}}$ on the unlabeled corpus $\mathcal{C}$.

PAQ-CANARD For the single-turn QA generation, Lewis et al. (2021) propose PAQ, the pipeline strategy composed of three phases, answer extraction, question generation, and round-trip filtration. Even though it is not designed to generate context-dependent questions, we generate decontextualized conversations like CANARD (Elgohary et al., 2019). Thus, we fine-tune every component of the PAQ on CANARD$_{train}$. Then, we include it as one of the baselines leveraging single-turn QA.

PAQ-QuAC We construct a baseline by using a straightforward way to extend the single-turn QG framework, e.g., PAQ (Lewis et al., 2021), to a conversational setup. We replace the question generator in PAQ with CQG$_{answer}$ model that also takes the conversation history as input. Different from our SimSeek-sym, the baseline utilizes the original answer extractor model of Lewis et al. (2021), which extracts answer candidates regardless of conversational history, i.e. $p_a(a \mid c)$. From a given answer-containing passage c, top-k answer candidates are extracted by the model in advance. Then, we randomly take out an answer from the candidates to feed it to the CQG$_{answer}$ at every turns.

A.4 Baslines in Downstream tasks

CQA Models After building synthetic CQA datasets upon the unlabeled corpus $\mathcal{C}$, the baseline CQA models are trained on the datasets. By comparing the resulting CQA performances, we evaluate the effectiveness of the generated dataset $\hat{\mathcal{D}}$. We test three backbone architectures for CQA, base and large size of RoBERTa (Liu et al., 2019), and Longformer-large (Beltagy et al., 2020). By contrasting various sizes of pre-trained models, we

[10] quac.ai
[11] Evaluation scripts are provided by quac.ai
[12] github.com/prdwb/orconvqa-release

show the different effects of data augmentation. In addition, we involve Longformer architecture that has been shown to be effective for encoding much longer history (Zhao et al., 2021), which achieves competitive performance with the state-of-the-art approach.

Conversational Search We employ dual-encoder based dense retriever, DPR, for our baseline (Karpukhin et al., 2020). Especially, we initialize the encoders with pre-trained DPR model on Natural Questions (Kwiatkowski et al., 2019). To represent query input, we concatenate questions $\{q_1, q_2, ..., q_t\}$ with `[SEP]` token. We truncate the input length when longer than 128 but retain first question q_1 at the same time (Qu et al., 2020). The context input is concatenation c and its title with `[SEP]`. The maximum length for the context input is 384. We train the model for 10 epochs with 128 for batch size, 3e-5 for lr, 0.1 for lr warming up, and 0.01 for weight decay. All DPR models are trained by using in-batch negative without any usage of hard negatives (Karpukhin et al., 2020).

B Additional Experiments

B.1 Intrinsic Evaluation of CQG models

Table 7 presents intrinsic evaluation results of our two kinds of CQG models. Scores represent the lexical similarity of the generated questions with the ground-truth questions when ground-truth conversational history is given. The sub-component of SIMSEEK-SYM, CQG$_{answer}$ significantly outperforms CQG$_{prior}$ in BLEU scores of all n-gram levels. The contrasting results to our experiments (Section 4.2) imply that accurate generation grounded on the answer is not enough to generate realistic conversation. Instead, we presume other vital factors, such as question based on information asymmetry, proper answer selection in natural conversational flow, and their chained interactions, contribute to a better synthetic CQA generation.

B.2 Ablation Study on the Filtration

We perform an ablation study on the filtering process and compare SIMSEEK-SYM to the strong baseline, PAQ-QuAC. The two frameworks show similar performance in Table 1. Despite the small gap, we observe the end-CQA performances are often saturated by the filtering procedure. Without the filtration ($\hat{\mathcal{D}}_{\text{unfilt}}$), SIMSEEK-SYM consistently outperforms PAQ-QuAC over all backbone baselines. It also shows better filtration efficiency by a gap of approximately 20%$_{\text{p}}$ in the success rate. The results indicate that SIMSEEK-SYM greatly advances the generation frameworks for simulating CQA datasets.

C Implementation Details

All our implementations are based on huggingface's transformers library (Wolf et al., 2019).

C.1 Training models in SimSeek

We train overall four models, CAE, CQG$_{answer}$, CQG$_{prior}$, and CAF on QuAC$_{seen}$ split for SimSeek. We optimize all models using AdamW optimizer with linear learning rate scheduling (Kingma and Ba, 2017). The best-performing checkpoint is selected according to validation score.

We employ 2D span extraction model proposed in PAQ with `bert-base-uncased` backbone for CAE (Devlin et al., 2019; Lewis et al., 2021). We observe that using the previous question and answer pair, (q_{t-1}, a_{t-1}), instead of the whole history $\mathcal{H}_t$ is enough to get reasonable performance. The qa pair is appended to c with `[SEP]` token for input representation. We set overall maximum sequence length to 512 and the maximum history length 32. We train it for 3 epochs with 8 for batch size, 3e-5 for lr on 1 32GB V100 GPU. To evaluate the model, we check whether the ground-truth answer span a_t is in the predicted top-10 answer spans $\hat{A}_t$, i.e. Recall@10.

For both CQG models, we employ same t5-large backbone but different input representations. First, c, `<sep>`, q_1, `<sep>`, a_1, ..., a_{t-1}, `<mask>`, a_t, `<sep>` are concatenated to represent input for CQG$_{answer}$, where the `<sep>` and `<mask>` are special seperator and masking token, respectively. And the output representation of it is concatenation of `<bos>`, q_t, and `<eos>`. As mentioned in Section 3.1, the c is highlighted by `<hl>` tokens to emphasize rationale for a_t (Gu et al., 2021). Second, $\mathcal{B}$, `<sep>`, q_1, `<sep>`, a_1, ..., a_{t-1}, `<mask>` are concatenated to represent input for CQG$_{prior}$ and the output representation is the same with that of CQG_{answer}. Actually, the $\mathcal{B}$ is composed of three textual inputs, title, section title, and background (abstractive description) [13]. They are also concatenated with the `<sep>` token to represent the $\mathcal{B}$. The masked question prediction scheme is inspired by Chada and Natarajan (2021) and we find the scheme is more sample efficient in our preliminary experiment. We

[13] Please see Table 9

Synthetic CQA Generation	Filtration		RoBERTa-base		RoBERTa-large		Longformer-large	
	#($\hat{\mathcal{D}}$)	%(Success)	$\hat{\mathcal{D}}_{\text{unfilt}}$	$\hat{\mathcal{D}}$	$\hat{\mathcal{D}}_{\text{unfilt}}$	$\hat{\mathcal{D}}$	$\hat{\mathcal{D}}_{\text{unfilt}}$	$\hat{\mathcal{D}}$
Human ($\hat{\mathcal{D}}$ = QuAC$_{\text{unseen}}$)	37,753	-	-	65.3	-	65.0	-	72.3
PAQ-QuAC	11,794	28.5 %	44.9	**55.9**	47.1	51.5	42.7	**61.7**
SIMSEEK-SYM	19,550	46.6 %	**51.8**	55.5	**53.3**	**54.3**	**53.6**	60.8

Table 6: Comparison over two baselines with detailed statistics for the filtering process. $\hat{\mathcal{D}}_{\text{unfilt}}$ represents the CQA dataset that is not filtered by our filtering process. Although the two frameworks achieve similar performances in Table 1, SIMSEEK-SYM largely outperforms PAQ-QuAC in the no-filtering setup.

Trained on	Model	B-1	B-2	B-3	B-4
QuAC$_{seen}$	CQG$_{answer}$	28.2	18.0	11.9	9.1
	CQG$_{prior}$	23.9	14.4	9.0	6.4
QuAC$_{full}$	CQG$_{answer}$	29.6	19.3	12.8	9.7
	CQG$_{prior}$	24.5	15.2	9.8	7.4

Table 7: Automatic evaluation over two different CQG models of our frameworks on QuAC development set. The B-* indicate BLEU scores.

QuAC	Train			Dev
QuAC$_{split}$	Seen	Unseen	Valid	Dev
# Passages	4,383	6,694	490	1,000
# Questions	31,527	37,753	3,430	7,354
OR-QuAC	Seen	Unseen	Dev	Test
# Passages	4,383	6,694	490	771
# Questions	31,527	-	3,430	5,571

Table 8: Data statistics of QuAC dataset used in our experiments. Note that we use questions and answers in QuAC$_{unseen}$ to represent human upper bound. OR-QuAC also contains 11M of chunked passages collection for the retrieval. We split datasets following CANARD (Elgohary et al., 2019), which is smilar with OR-QuAC (Qu et al., 2020)

train both CQG models for 10 epochs with 16 for batch size, 3e-5 for lr, 0.1 for lr warming up, and 0.01 for weight decay on 2 32GB V100 GPUs. We set maximum sequence length for the input representations to 512 and maximum context length to 384. The context means c and $\mathcal{B}$ for CQG$_{answer}$ and CQG$_{prior}$, respectively.

We adopt three CQA backbone architectures, RoBERTa-base, RoBERTa-large (Liu et al., 2019), and Longformer-large (Beltagy et al., 2020), which are shown to be effective in CQA task. Please note that any CQA models can be used for CAF model as a teacher. For all CQA models, we concatenate the title, sub-title, and previous history to question text, separating with the special token `[SEP]`. We train all models for 2 epochs without weight decay on all datasets and set maximum answer length 64. CQA models return `CANNOTANSWER` when all scores of answer logits do not exceed a pre-defined threshold. RoBERTa backbones are trained for batch size 12 per each GPU without weight decay. We set the maximum length for query and input sequences as 128 and 512, respectively. Due to their limitation of the input sequence length, a single question-answer pair at previous turn $(t-1)$ is included to the input, shown to be most effective in prior works (Qu et al., 2019a). When Longformer architecture is adopted, we find the optimal setup of the maximum length for query and sequence as 768 and 2048, respectively. It encodes all previous history and titles when providing answers. They are trained with batch size 1 per GPU. For the large-size models, we train them with a learning rate 1.5e-5 on 8 32GB V100 GPUs.

C.2 Training on Wiki-SimSeek

When trained on WIKI-SIMSEEK, we validate models every 20000 steps with the validation split of QuAC. We evaluate best performing models on the development set

D Dataset Statistics

Table 8 shows data statistics used in our experiments.

D.1 Case Study

We explore how SimSeek-sym fails to simulate realistic conversations, but SimSeek-asym successfully mimics information-seeking behaviors.

The first case in Table 9 shows the synthetic conversation simulated by SimSeek-sym. In the example, consecutive and disjoint spans are selected for answers from the evidence document as the conversation progresses. Moreover, all questions contain a common phrase "What happened $\cdots$ " while mentioning keywords that have appeared in

SimSeek-sym
Title : Native Americans in the United States **Section Title** : Self-determination

Document c :
. . .
Upset with tribal government and the failures of the federal government to enforce treaty rights, about 300 Oglala Lakota and AIM activists took control of Wounded Knee on February 27, 1973. Indian activists from around the country joined them at Pine Ridge, and the occupation became a symbol of rising American Indian identity and power. Federal law enforcement officials and the national guard cordoned off the town, and the two sides had a standoff for 71 days.
. . .

. . .
q_3 : What happened at Wounded Knee?
a_3 : Indian activists from around the country joined them at Pine Ridge, and the occupation became a symbol of rising American Indian identity and power.
q_4 : What happened after they took control of Pine Ridge?
a_4 : Federal law enforcement officials and the national guard cordoned off the town, and the two sides had a standoff for 71 days.
q_5 : What happened during the standoff?
a_5 : During much gunfire, one United States Marshal was wounded and paralyzed.
. . .

SimSeek-asym
Title : Thor Heyerdahl **Section Title** : Kon-Tiki expedition

Background $\mathcal{B}$
Thor Heyerdahl (October 6, 1914 - April 18, 2002) was a Norwegian adventurer and ethnographer with a background in zoology, botany, and geography. He became notable for his Kon-Tiki expedition in 1947, · · ·

. . .
q_4 : What were some of the things he found on the Kon-Tiki expedition?
a_4 : The raft proved to be highly manoeuvrable, and fish congregated between the nine balsa logs in such numbers that ancient sailors could have possibly relied on fish for hydration in the absence of other sources of fresh water.
q_5 : Are there any other interesting aspects about this article?
a_5 : The documentary film of the expedition entitled Kon-Tiki won an Academy Award in 1951.
q_6 : Why did the film win an Academy Award?
a_6 : CANNOTANSWER

Table 9: Examples of the resulting datasets simulated by SimSeek-sym and SimSeek-asym. In the first case (above), SimSeek-sym asks unspecific questions repeatedly, which can effortlessly achieve the goals, answer relevance and coherence with the conversation; but it leads to the shallow conversation. On the contrary, SimSeek-asym successfully mimics diverse information-seeking behaviors that are commonly occurred in human dialogue.

previous answers. Asking these ambiguous questions repeatedly would be the best option for the answer-grounded CQG to achieve answer relevance and coherence with the conversation easily.

On the other hand, we observe various information-seeking behaviors in the second case from our SimSeek-asym. The lack of information impels questioners to ask open-ended questions using uncertain words such as "some of the things" in q_4. When they cannot find adequate follow-up questions on the conversation, they ask an additional information as in q_5. They sometimes fail to acquire new knowledge when the question cannot be answered by the evidence document (see q_6, a_6).

E Human Evaluation Details

Data Preparation We ask five workers to compare and rate each candidate follow-up QA generated by one of the models or sampled from the original dataset, given the dialogue history and document context (total 296 samples). The workers were asked to assess the five criteria ranging from the informativeness of the question and the answer accuracy.

As reported by Li et al. (2021), we also observe that the annotators exhibit a bias for questions that do not have an answer (i.e., CANNOTANSWER). In addition, we find that the annotators tend to score favorably "Anything else?" questions in most criterion since they often seem relevant to their answer and conversational context. Thus, we filtered out those types of QA pairs when reporting the scores.

Title : Esports **Section Title** : History Early history (1972–1989)

Document c

The earliest known video game competition took place on 19 October 1972 at Stanford Universit for the game "Spacewar". Stanford students were invited to an "Intergalactic spacewar olympics" whose grand prize was a yearś subscription for "Rolling Stone", with Bruce Baumgart winning the five-man-free-for-all tournament and Tovar and Robert E. Maas winning the Team Competition. The Space Invaders Championship held by Atari in 1980 was the earliest large scale video game competition, attracting more than 10,000 participants across the United States , establishing competitive gaming as a mainstream hobby. · · ·

Background $\mathcal{B}$

Esports (also known as electronic sports, e-sports, or eSports) is a form of competition using video games. Most commonly, esports takes the form of organized, multiplayer video game competitions, particularly between professional players, individually or as teams. Although organized online and offline competitions have long been a part of video game culture, these were largely between amateurs until the late 2000s, when participation by professional gamers and spectatorship in these events through live streaming saw a large surge in popularity. By the 2010s, esports was a significant factor in the video game industry, with many game developers actively designing toward a professional esports subculture.

Conversation in WIKI-SIMSEEK

q_1 : **What is the history of esports?**

a_1 : The earliest known video game competition took place on 19 October 1972 at Stanford University for the game "Spacewar".

q_2 : **What was the result of this competition?**

a_2 : Bruce Baumgart winning the five-man-free-for-all tournament and Tovar and Robert E. Maas winning the Team Competition.

q_3 : **Did esports grow from there?**

a_3 : The Space Invaders Championship held by Atari in 1980 was he earliest large scale video game competition, attracting more than 10,000 participants across the United States,

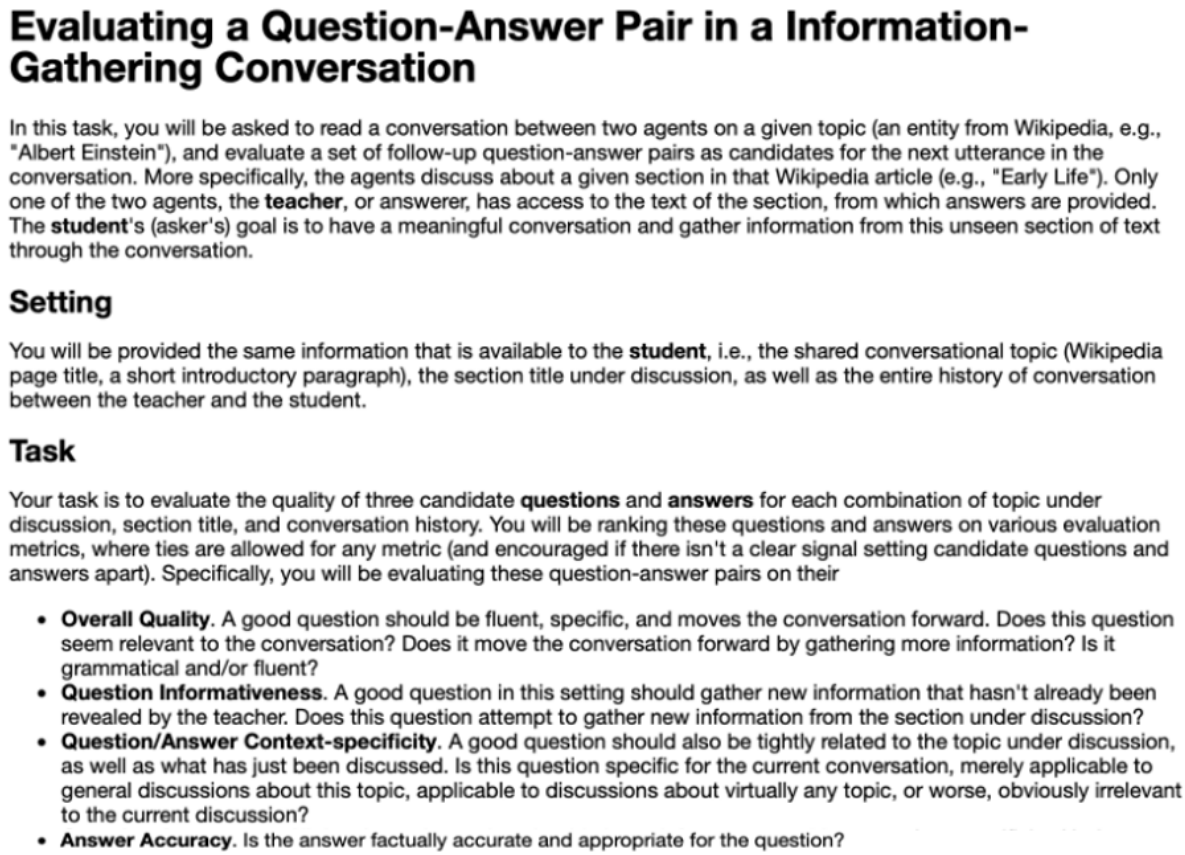
q_4 : **What happened after the Space Invaders Championship?**

· · ·

Table 10: Another qualitative example in WIKI-SIMSEEK. Especially, it shows that our framework works well even for topics of out-of-domain, i.e., “esports”, which is not person-related categories as in original QuAC.

F Other Details

Computational Cost We conduct training and inference once for all experiments since it takes huge computational cost. Training Longformer-large on WIKI-SIMSEEK takes 4 days for machines with 8 V100 GPUs.



Evaluating a Question-Answer Pair in a Information-Gathering Conversation

In this task, you will be asked to read a conversation between two agents on a given topic (an entity from Wikipedia, e.g., "Albert Einstein"), and evaluate a set of follow-up question-answer pairs as candidates for the next utterance in the conversation. More specifically, the agents discuss about a given section in that Wikipedia article (e.g., "Early Life"). Only one of the two agents, the **teacher**, or answerer, has access to the text of the section, from which answers are provided. The **student**'s (asker's) goal is to have a meaningful conversation and gather information from this unseen section of text through the conversation.

Setting

You will be provided the same information that is available to the **student**, i.e., the shared conversational topic (Wikipedia page title, a short introductory paragraph), the section title under discussion, as well as the entire history of conversation between the teacher and the student.

Task

Your task is to evaluate the quality of three candidate **questions** and **answers** for each combination of topic under discussion, section title, and conversation history. You will be ranking these questions and answers on various evaluation metrics, where ties are allowed for any metric (and encouraged if there isn't a clear signal setting candidate questions and answers apart). Specifically, you will be evaluating these question-answer pairs on their

- **Overall Quality**. A good question should be fluent, specific, and moves the conversation forward. Does this question seem relevant to the conversation? Does it move the conversation forward by gathering more information? Is it grammatical and/or fluent?
- **Question Informativeness**. A good question in this setting should gather new information that hasn't already been revealed by the teacher. Does this question attempt to gather new information from the section under discussion?
- **Question/Answer Context-specificity**. A good question should also be tightly related to the topic under discussion, as well as what has just been discussed. Is this question specific for the current conversation, merely applicable to general discussions about this topic, applicable to discussions about virtually any topic, or worse, obviously irrelevant to the current discussion?
- **Answer Accuracy**. Is the answer factually accurate and appropriate for the question?

Figure 4: The detailed instructions given to the crowdworkers during human evaluation on Amazon Mechanical Turk. The task description and the settings were based on prior work (Qi et al., 2020)