Synthesize, Prompt and Transfer: Zero-shot Conversational Question Generation with Pre-trained Language Model

Hongwei Zeng[1,2], **Bifan Wei**[2,3]*, **Jun Liu**[1,2], **Weiping Fu**[1,2]
[1]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering,
School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, China
[3]School of Continuing Education, Xi'an Jiaotong University, China
hongwei.zeng@foxmail.com, {weibifan@, liukeen@, fuweiping@stu.}xjtu.edu.cn

Abstract

Conversational question generation aims to generate questions that depend on both context and conversation history. Conventional works utilizing deep learning have shown promising results, but heavily rely on the availability of large-scale annotated conversations. In this paper, we introduce a more realistic and less explored setting, **Zero**-shot **C**onversational **Q**uestion **G**eneration (**ZeroCQG**), which requires no human-labeled conversations for training. To solve ZeroCQG, we propose a multi-stage knowledge transfer framework, **S**ynthesize, **P**rompt and tr**A**nsfer with p**Re**-**T**rained l**A**nguage model (**SPARTA**) to effectively leverage knowledge from single-turn question generation instances. To validate the zero-shot performance of SPARTA, we conduct extensive experiments on three conversational datasets: CoQA, QuAC, and DoQA by transferring knowledge from three single-turn datasets: MS MARCO, NewsQA, and SQuAD. The experimental results demonstrate the superior performance of our method. Specifically, SPARTA has achieved 14.81 BLEU-4 (88.2% absolute improvement compared to T5) in CoQA with knowledge transferred from SQuAD.

Conversational Question Generation
Context: Friedrich, pausing at Gross-Nossen, and perhaps a little surprised to find no Loudon meddling with him, pushes out, first one party and then another, Dalwig, Bulow, towards Landshut Hill-Country, to threaten Loudon's Bohemian roads;–who, singular to say, do not hear the least word of Loudon thereabouts. **Conversation History**: Q1: Who paused at Gross-Nossen? A1: Friedrich Q2: What was he caught off guard about? A2: No Loudon meddling with him **Answer**: Dalwig and Bulow
Question: What parties did *he* push out?

Table 1: A conversational QG instance in CoQA dataset (Reddy et al., 2019).

1 Introduction

Question Generation (QG) aims to automatically generate questions from the given context and answer. It plays a vital role in knowledge testing (Heilman and Smith, 2010; Lindberg et al., 2013; Ghanem et al., 2022) and information seeking (Shum et al., 2018; Rosset et al., 2020; Zamani et al., 2020) by creating quiz questions and spanning question suggestions, respectively. Most existing QG research has usually focused on generating single-turn questions, which are formalized as independent interactions (Zhou et al., 2017; Zhao et al., 2018; Tuan et al., 2020). However, it is a more natural way to achieve complex information need through conversations involving a series of interconnected questions (Reddy et al., 2019). Different from single-turn QG, the task of conversational QG (Gao et al., 2019) aims to generate questions which depend on both context and conversation history.

Recent conversational QG models (Gao et al., 2019; Pan et al., 2019; Wang et al., 2022b) which utilize a separate neural encoder to handle the conversation history, have achieved great performance on CoQA dataset (Reddy et al., 2019). However, these deep models rely heavily on large-scale annotated conversations which provides the dependency between conversation history and the follow-up question. As shown in Table 1, we cannot infer who the *he* in the question is referring to without taking into account the conversation history. Therefore, it is also impossible to generate a conversational question with a referential phenomenon to the history, e.g., *Friedrich* and *he*.

In this paper, we propose a more realistic and less explored setting, **Zero**-shot **C**onversational **Q**uestion **G**eneration (**ZeroCQG**), which requires no human-labeled conversational datasets for training. To solve ZeroCQG, we propose to transfer knowledge from single-turn QG instances and the

[*] Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8989–9010
July 9-14, 2023 ©2023 Association for Computational Linguistics

pre-trained Language Model (LM). The relation of question to context and answer plays an important role in both single-turn and conversational QG, while single-turn QG instances are often abundant and easier to obtain. However, there is still a significant domain gap between the two QG tasks due to the lack of conversation history in single-turn QG. More recently, pre-trained LMs brings remarkable performance improvement on the task of conversational QG (Do et al., 2022; Fei et al., 2022) due to their massive amounts of linguistic knowledge and powerful contextual representation capabilities (Li et al., 2021). However, the different input and output paradigms will also increase the domain gap between the objective of pre-trained LM and the conversational QG.

To address these issues, we propose a multi-stage knowledge transfer framework, **S**ythesize, **P**rompt and tr**A**nsfer with p**R**e-**T**rained l**A**nguage model (**SPARTA**) to effectively leverage knowledge from single-turn QG instances. **(1) Synthesize.** We synthesize conversation for each single-turn QG instance to alleviate the domain gap between single-turn and conversational QG tasks. Specifically, we first retrieve question-answer pairs with similar contextual contents and sequential dependencies from the whole single-turn QG dataset to stimulate history for each single-turn QG instance. Then, we incorporate anaphora characteristics into the single-turn question by replacing entity co-occurring in both the question and the simulated history with co-referenced pronouns. **(2) Prompt.** We propose conversation prompting to alleviate the domain gap between the objective of pre-trained LM and conversational QG. Specifically, this prompting method reformulates the conversational QG as a masked question-filling task similar to T5 (Raffel et al., 2020) where the input and output are organized by prompt templates with semantic prefixes to better steer the expressive power of pre-trained LM. **(3) Transfer.** We fine-tune pre-trained LM on the synthetic dataset with conversation prompting. Then, the fine-tuned pre-trained LM with the same conversation prompting are directly applied for inference of conversational QG without using any annotated conversations for training.

To validate the zero-shot performance of our proposed SPARTA, we conduct extensive experiments on three conversational datasets: CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018) and DoQA (Campos et al., 2020) by transferring knowledge from three single-turn datasets: MS MARCO (Nguyen et al., 2016), NewsQA (Trischler et al., 2017) and SQuAD (Rajpurkar et al., 2016) based on different pre-trained LMs: T5 (Raffel et al., 2020), BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020). The experimental results demonstrate that our proposed SPARTA significantly improves the performance of ZeroCQG on most transfer settings. For example, SPARTA (T5) achieves 14.81 BLEU-4 (88.2% absolute improvement compared to T5) in CoQA with knowledge transferred from SQuAD, We further conduct extensive ablation studies and discussions to explore the effectiveness of each component of the proposed SPARTA.

We summarize our main contributions as follows:

- We introduce a novel task setting, ZeroCQG, which requires no human-labeled conversations for training.
- We propose a multi-stage knowledge transfer framework, SPARTA, which effectively leverages knowledge from single-turn QG instances and pre-trained LM for ZeroCQG.
- We have conducted extensive experiments to demonstrate the superior performance of SPARTA in most transfer settings.

2 Problem Definition

In this section, we first introduce the definition of conversational QG task. Given a context c^u, a conversation history $h_t^u = \{(q_1^u, a_1^u), \ldots, (q_{t-1}^u, a_{t-1}^u)\}$, a answer a_t^u, the conversational QG task aims to generate a followup question q_t^u at t-th turn:

$$q_t^u = \arg\max_q P(q|c^u, h_t^u, a_t^u) \qquad (1)$$

in which the generated question should be coherent with the conversation history and be conversational.

Furthermore, the task of ZeroCQG is defined to generate conversational questions without using any human-labeled conversations.

3 Methodology

In this section, we introduce the proposed SPARTA which mainly contains three stages: synthesize, prompt, and transfer. The overall framework is illustrated in Figure1.

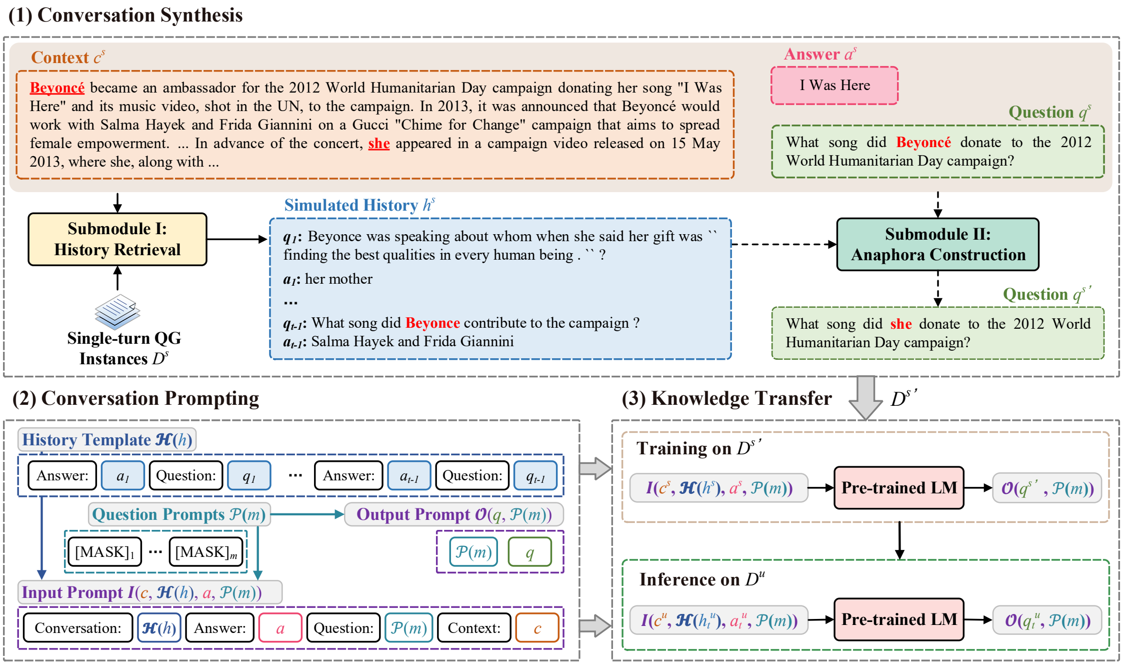

(1) Conversation Synthesis
Context c^s
Beyoncé became an ambassador for the 2012 World Humanitarian Day campaign donating her song "I Was Here" and its music video, shot in the UN, to the campaign. In 2013, it was announced that Beyoncé would work with Salma Hayek and Frida Giannini on a Gucci "Chime for Change" campaign that aims to spread female empowerment. ... In advance of the concert, she appeared in a campaign video released on 15 May 2013, where she, along with ...
Answer a^s
I Was Here
Question q^s
What song did Beyoncé donate to the 2012 World Humanitarian Day campaign?
Submodule I: History Retrieval
Single-turn QG Instances D^s
Simulated History h^s
q_1: Beyonce was speaking about whom when she said her gift was `` finding the best qualities in every human being . `` ?
a_1: her mother
...
q_{t-1}: What song did Beyonce contribute to the campaign ?
a_{t-1}: Salma Hayek and Frida Giannini
Submodule II: Anaphora Construction
Question $q^{s'}$
What song did she donate to the 2012 World Humanitarian Day campaign?
(2) Conversation Prompting
History Template $\mathcal{H}(h)$
Answer: a_1 Question: q_1 ... Answer: a_{t-1} Question: q_{t-1}
Question Prompts $\mathcal{P}(m)$
[MASK]$_1$ ··· [MASK]$_m$
Output Prompt $\mathcal{O}(q, \mathcal{P}(m))$
$\mathcal{P}(m)$ q
Input Prompt $I(c, \mathcal{H}(h), a, \mathcal{P}(m))$
Conversation: $\mathcal{H}(h)$ Answer: a Question: $\mathcal{P}(m)$ Context: c
(3) Knowledge Transfer
$D^{s'}$
Training on $D^{s'}$
$I(c^s, \mathcal{H}(h^s), a^s, \mathcal{P}(m))$
Pre-trained LM
$\mathcal{O}(q^{s'}, \mathcal{P}(m))$
Inference on D^u
$I(c^u, \mathcal{H}(h_i^u), a_i^u, \mathcal{P}(m))$
Pre-trained LM
$\mathcal{O}(q_i^u, \mathcal{P}(m))$

Figure 1: Illustration of the proposed SPARTA for ZeroCQG. **(1) Synthesize.** For each single-turn QG instance $(c^s, a^s, q^s) \in D^s$, the conversation synthesis module will retrieve $t-1$ question-answer pairs from D^s as simulated history h^s, and transform the q^s into $q^{s'}$ if there exist co-reference with pronoun, e.g. *Beyoncé* and *she*. We term the dataset with synthesized conversation as $D^{s'} = \{(c^s, h^s, a^s, q^{s'})\}$. **(2) Prompt.** We propose conversation prompting to reformulate the input and output of conversational QG. **(3) Transfer.** We fine-tune pre-trained LM on the prompted $D^{s'}$. Then, the fine-tuned pre-trained LM with the same conversation prompting is directly applied for inference on D^u to generate conversational questions.

3.1 Conversation Synthesis

To alleviate the domain gap between single-turn and conversational QG, we synthesize conversation for each single-turn QG instance.

3.1.1 History Retrieval

History is the most differentiating aspect between single-turn and conversational QG. We retrieve question-answer pairs from the whole dataset to simulate the history for each single-turn QG instance. Specifically, we first retrieve question-answer pairs with similar contextual content to the context c^s of the single-turn question q^s as candidates. The similarity score of question-answer pairs is calculated through the dot product between the TF-IDF weighted bag-of-word vectors of the corresponding contextual content. Therefore, the retrieved questions are likely to be answerable given the context c^s of the query question q^s. Notably, for examples in datasets like SQuAD and NewsQA, we can directly adopt the multiple question-answer pairs corresponding to the same context as simulated history candidates of each other.

Then, we rank the question-answer pairs in the candidate set according to their relevance to q^s. Specifically, we leverage the Next Sentence Prediction (NSP) based on the pre-trained BERT (Devlin et al., 2019) to capture the intrinsic sequential dependencies between question pairs. We take the concatenation of the candidate history question q and the query question q^s, like "[CLS] q [SEP] q^s", as input to NSP, and obtain the probability that q^s can be semantically inferred from q, with label *isNext*. Then, we select the highest $t-1$ question-answer pairs to stimulate conversation history according to this probability. The question-answer pair with higher probability is closer to q^s in the synthesized conversation. Therefore, for each single-turn QG instance $(c^s, a^s, q^s) \in \mathcal{D}^s$, we can obtain a ranked list of question-answer pairs as the simulated history $h^s = \{(q_i, a_i)\}_{i=1}^{t-1}$.

3.1.2 Anaphora Construction

Anaphora is the most common characteristic in conversation systems (Reddy et al., 2019). To incorporate anaphora into the single-turn question and the simulated history, we replace co-occuring entities

Domain	Dataset	Train	Dev	Test	History Turns	L_C	L_Q	L_A
Single-turn	**MS MARCO**	73,794	9,030	-	-	83.00	6.05	17.05
	NewsQA	92,549	5,166	-	-	446.52	7.63	4.94
	SQuAD	89,644	10,570	-	-	138.32	11.30	3.36
Conversational	**CoQA**	-	-	5,945	7.01	312.81	6.35	3.21
	QuAC	-	-	4,869	3.44	521.81	7.58	17.17
	DoQA	-	-	2,714	2.03	143.80	15.34	18.67

Table 2: Dataset statistics for ZeroCQG. L_C, L_Q, L_A refer to the average length of context, question and answer respectively. The average length is calculated after word tokenization using NLTK (Wagner, 2010).

with co-referenced pronouns. Specifically, we first concatenate the context c^s, the simulated history h^s and the question q^s into one long text. Then, we employ a pre-trained document-level co-reference resolution model, SpanBERT (Joshi et al., 2020), to cluster mentions in the long text which refer to the same real-world entities. Finally, we transform question q^s into $q^{s'}$ by replacing the co-occurring entities appearing in both q^s and h^s with pronouns in the same mention cluster, e.g. *Beyoncé* and *she*.

Overall, we can synthesize a conversational QG dataset $D^{s'} = \{(c^s, h^s, a^s, q^{s'})\}$ with simulated history h^s and transformed question $q^{s'}$.

3.2 Conversation Prompting

To alleviate the domain gap between the objective of pre-training LM and the conversational QG task, we reformulate the objective of conversational QG as a masked question-filling task. Specifically, the input and output prompt are detailed as follows:

Input Prompt For the conversation history, the template $\mathcal{H}$ concatenates $h = \{(q_i, a_i)\}_{i=1}^{t-1}$ into a text sequence where the components are identified by semantic prefixes "question:" and "answer:" respectively, rather than newly introduced tokens. For the masked question, multiple consecutive prompt tokens $\mathcal{P}(m) = [\text{[MASK]}_1, \cdots, \text{[MASK]}_m]$ are replaced in the corresponding position of the conversation after the history, where m is the length of question prompt tokens and each prompt token [MASK]_i has trainable parameters equal to the size of embedding vector. Finally, the input prompt is composed of context c, history template $\mathcal{H}(h)$, answer a, and question prompts $\mathcal{P}(m)$, formalized as $\mathcal{I}(c, \mathcal{H}(h), a, \mathcal{P}(m))$ with additional semantic prefixes "conversation:" and "context:".

Output Prompt The same question prompt tokens $\mathcal{P}(m)$ used in the input are prepended before the target question q as the model output, formalized as $\mathcal{O}(q, \mathcal{P}(m))$. A longer sequence of question prompt tokens means more trainable parameters, and therefore more expressive power to steer pre-trained LMs to capture the semantic representation of the question prompt in the corresponding position of the input and provide direct guidance for the generation of output question.

3.3 Knowledge Transfer

SPARTA transfers knowledge from single-turn QG to conversational QG based on pre-trained LM. The training and inference is detailed as follows:

Training Our model is continuously trained based on the pre-trained LM as an intermediate task (Pruksachatkun et al., 2020) on the synthesized dataset $D^{s'}$. Specifically, we leverage conversation prompting to transform each instance $(c^s, h^s, a^s, q^{s'}) \in D^{s'}$ as instantiated input $X^s = \mathcal{I}(c^s, \mathcal{H}(h^s), a^s, \mathcal{P}(m))$ and output $Y^s = \mathcal{O}(q^{s'}, \mathcal{P}(m))$ for training. The model with parameter θ is optimized with negative entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{l_{Y^s}} \log P_\theta(Y_i^s | X^s, Y_{<i}^s) \quad (2)$$

where $l_{Y^s} = m + l_q$ and l_q refer to the length of output Y^s and question $q^{s'}$ respectively.

Inference We directly use the fine-tuned model with parameter θ for inference on D^u. Specifically, we use the same conversation prompting to transform the input of conversational QG as $X^u = \mathcal{I}(c^u, \mathcal{H}(h_t^u), a_t^u, \mathcal{P}(m))$. Then, the output is generated as:

$$Y^u = \arg\max_Y P_\theta(Y | X^u) \quad (3)$$

By removing prompt tokens $\mathcal{P}(m)$ from Y^u, we can obtain the generated conversational question.

4 Experiment

4.1 Datasets

We use three single-turn datasets as the source datasets: MS MARCO (Nguyen et al., 2016),

Source Dataset	Model	CoQA			QuAC			DoQA		
		B-4	MR	R-L	B-4	MR	R-L	B-4	MR	R-L
MS MARCO	PEGASUS	2.75	11.08	19.23	1.23	7.41	13.43	1.25	6.65	12.23
	BART	3.26	11.51	20.15	1.33	7.98	14.96	0.87	6.41	11.74
	T5	3.97	12.35	20.93	1.46	7.98	14.84	1.31	7.06	12.70
	SPARTA (PEGASUS)	2.67	9.16	16.76	3.00	**9.12**	18.04	1.74	**7.77**	13.89
	SPARTA (BART)	3.54	11.26	19.91	2.89	9.07	18.10	1.33	7.09	13.09
	SPARTA (T5)	**5.33**	**12.60**	**26.23**	**3.04**	9.09	**19.01**	**1.97**	7.74	**13.94**
NewsQA	PEGASUS	7.17	16.38	37.44	2.42	9.97	27.56	0.99	6.76	20.82
	BART	9.08	18.52	40.69	3.56	10.73	28.92	1.27	6.94	20.87
	T5	9.74	18.84	40.06	3.54	10.58	27.51	1.52	7.29	19.95
	SPARTA (PEGASUS)	9.35	16.86	40.63	5.17	11.66	32.89	1.80	7.63	21.62
	SPARTA (BART)	11.54	17.93	43.40	5.58	11.66	33.44	1.46	7.46	**22.10**
	SPARTA (T5)	**13.34**	**18.86**	**45.02**	**6.21**	**12.07**	**33.47**	**2.28**	**8.09**	21.51
SQuAD	PEGASUS	7.19	18.22	35.67	2.51	10.21	23.51	1.97	8.01	19.70
	BART	7.40	18.61	35.66	2.51	10.24	23.07	1.85	7.90	18.70
	T5	7.87	18.58	35.11	2.81	10.01	22.10	2.10	8.15	19.02
	SPARTA (PEGASUS)	11.12	19.69	42.36	4.61	11.70	28.60	2.47	8.57	20.95
	SPARTA (BART)	12.61	20.75	44.33	5.14	11.47	29.61	2.35	8.38	20.92
	SPARTA (T5)	**14.81**	**21.56**	**45.86**	**5.85**	**11.96**	**30.43**	**2.66**	**8.70**	**21.44**

Table 3: Zero-shot performance comparisons on three conversational benchmarks: CoQA, QuAC and DoQA with knowledge transferred from three single-turn datasets: MS MARCO, NewsQA and SQuAD respectively. B-4, MR and R-L refer to BLEU-4, METEOR and ROUGE-L respectively. The optimal values in one and all three source datasets are marked in **bold** and underline respectively.

NewsQA (Trischler et al., 2017), and SQuAD (Rajpurkar et al., 2016) and three conversational datasets as the target datasets: CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), and DoQA (Campos et al., 2020). The processed dataset statistics are displayed in Table 2. More details of datasets are in Appendix A.1.

4.2 Baselines

As this novel task setting of ZeroCQG has not been explored in previous work, there is no existing method to compare with. Therefore, we used three commonly used encoder-decoder style pre-trained LMs: T5 (Raffel et al., 2020), BART (Lewis et al., 2020), and PEGASUS (Zhang et al., 2020), as baselines. More details are in Appendix A.2.

4.3 Main Results

Table 3 presents the zero-shot performance comparison on three conversational datasets. From that, we have the following findings:

(1). SPARTA significantly outperforms baseline models across most transfer settings in terms of various metrics. For example, SPARTA (T5) outperforms T5 by a large margin on the transfer from SQuAD to CoQA obtaining 88.2% absolute improvement in BLEU-4, 16.2% absolute improvement in METEOR, and 30.6% absolute improvement in ROUGE-L. When transferring from NewsQA to QuAC, SPARTA achieves an absolute improvement of 2.75, 2.02, and 2.67 in BLEU-4 compared to vanilla PEGASUS, BART, and T5, respectively.

(2). T5 has better zero-shot generalization performance on conversational QG task. We have observed that T5 achieves better results than BART and PEGASUS in most transfer settings. Similarly, SPARTA (T5) also outperforms SPARTA (BART) and SPARTA (PEGASUS). This may be because the span corruption object in T5 is more generable compared to the gap-sentence generation object designed for abstractive summarization in PEGASUS and the corrupted text reconstruction object using denoising auto-encoder in BART.

(3). Short answers are easier to understand and thus lead to better transfer results. As shown in Table 2, the average answer lengths L_A of NewsQA, SQuAD, and CoQA are shorter than 5, while those of MS MARCO, QuAC, and DoQA are longer than 17. We can observe that the performance of transferring from NewsQA or SQuAD to CoQA is significantly higher than other transfer settings. Among them, the knowledge transferred from MS MARCO has the worst generalization ability. This is probably because the answers in MS MARCO are human-generated, lengthy, and difficult for ma-

Model	MS MARCO			NewsQA			SQuAD		
	CoQA	QuAC	DoQA	CoQA	QuAC	DoQA	CoQA	QuAC	DoQA
SPARTA (T5)	**5.33**	**3.04**	1.97	**13.34**	**6.21**	**2.28**	**14.81**	**5.85**	2.66
- w/o CS	4.64	1.59	1.40	11.69	3.81	1.35	9.79	4.12	2.04
- w/o AC	4.82	2.43	1.92	12.56	5.69	2.27	13.60	5.47	2.70
- w/o CP	2.82	2.65	**2.04**	11.51	5.69	**2.28**	11.55	4.56	**2.83**

Table 4: Ablation results in terms of BLEU-4 score. CS, AC and CP refer to conversation synthesis, anaphora construction and conversation prompting respectively. Notably, AC module depends on simulated history.

chines to understand, and the instance number is much fewer. Besides, models learned from different single-turn data all perform poorly on DoQA. This may be due to the fact that the question-answer pairs in DoQA are all domain-specific, and the length distribution of the question-answer is quite different from the single-turn datasets.

(4). The closer the average question lengths L_Q of the single-turn and conversational datasets are, the better the zero-shot generalization performance will be. As shown in Table 2, compared to SQuAD, L_Q in NewsQA is closer to L_Q in CoQA and QuAC, but farther away from L_Q in DoQA. Similarly, as shown in Table 3, we observe that the baselines trained on NewsQA achieve better performance on CoQA and QuAC than that trained on SQuAD, but worse on DoQA.

4.4 Ablation Studies

We conduct ablation experiments over different variants of the best-performing model SPARTA (T5) to better understand the relative importance of the proposed SPARTA framework. As shown in Table 4, most variants lead to worse performance and yet still outperform the baseline model T5.

Conversation Synthesis. When we transfer knowledge from the single-turn dataset without using CS, the performance of our model drops significantly. For example, when transferring knowledge from SQuAD, the BLEU-4 score drops from 14.81, 5.85, and 2.66 to 9.79, 4.12, and 2.04 in CoQA, QuAC and DoQA respectively. This confirms that the dependency on conversation history is important to the conversational QG task. This module alleviates the domain gap between single-turn and conversational QG with simulated history and constructed anaphora, thus improving the transfer result.

Anaphora Construction. By turning off the AC module, the BLEU-4 score drops to 2.43, 5.69, and 5.47 in QuAC with knowledge transferred from MS MARCO, NewsQA, and SQuAD respectively. The same performance decrease phenomenon can also be seen in the other transfer settings. This demonstrates that there is a difference between single-turn and conversational questions. Training on synthetic datasets with constructed anaphora characteristics is able to generate more conversational questions. While the AC module is mainly based on the co-reference resolution model, SpanBERT. The ablation of AC also verifies the effectiveness of the co-reference resolution model in understanding anaphora phenomena in conversation.

Conversation Prompting The variant without CP formalizes the input similar to that commonly used in conversational question answering systems (Reddy et al., 2019), i.e. appending the conversation history and target answer before the context as $\langle a \rangle\ a_1\ \langle q \rangle\ q_1 \cdots \langle a \rangle\ a_{t-1}\ \langle q \rangle\ q_{t-1}\ \langle a \rangle\ a_t\ \langle sep \rangle\ c$. $\langle a \rangle$ and $\langle q \rangle$ are special tokens used to identify answers and questions, respectively. c is the context. And the question q_t is taken as output without using any prompts to guide the decoding process. We can see that this variant leads to a large decrease in BLEU-4 scores on CoQA and QuAC, but a slight increase on DoQA. This may be because DoQA is a domain-specific FAQ dataset with longer questions and answers, which has a larger domain gap with the source datasets than CoQA and QuAC. This result shows that CP can enhance the zero-shot generalization ability of the pre-trained LM when the domains are relevant but has limitations when the domain gap becomes large.

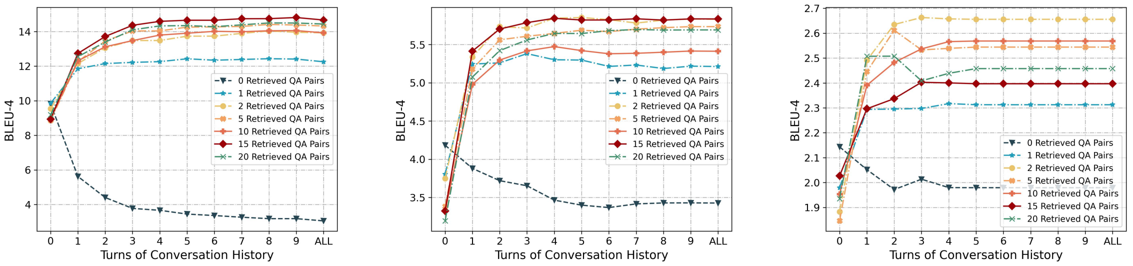
4.5 Analysis of Question Ranking Method

History selection is an important module in conversational systems (Zaib et al., 2022). As shown in Table 5, we have explored different question ranking algorithms to investigate the effectiveness of retrieved question-answer pairs for conversation synthesis. The observations are as followings:

(1). All of these question ranking methods lead to significant performance gains. We observe that

Question Ranking	MS MARCO			NewsQA			SQuAD		
	CoQA	QuAC	DoQA	CoQA	QuAC	DoQA	CoQA	QuAC	DoQA
SPARTA (T5) (-w NSP)	5.33	3.04	**1.97**	13.34	**6.21**	**2.28**	**14.81**	**5.85**	**2.66**
-w TF-IDF	4.95	2.99	1.71	13.10	5.58	2.06	13.59	5.69	2.60
-w Levenshtein	4.80	**4.16**	1.74	12.23	5.75	2.27	13.88	5.74	2.64
-w Dense Retrieval	**6.16**	2.53	1.59	**14.16**	6.08	1.93	13.96	5.47	2.28

Table 5: Performance comparison of the BLEU-4 score with different question ranking methods for conversation synthesis. The Dense Retrieval method encodes the query and candidate questions into low-dimensional embedding using the pre-trained BERT and performs retrieval in the embedding space with dot product. Bold and underlined values indicate the best and worst score, respectively.



(a) CoQA

(b) QuAC

(c) DoQA

Figure 2: Performance comparison of SPARTA (T5) trained with different maximum number of single-turn question-answer pairs retrieved from SQuAD. The x-axis refers to maximum number of conversational history turns used during inference on conversational datasets.

all these variants achieve higher BLEU-4 scores compared to the variant SPARTA (T5) (-w/o CS) shown in Table 4. This also demonstrates the importance and robustness of the conversation synthesis module in ZeroCQG.

(2). NSP is best suited for retrieving question-answer pairs to simulate conversation history. We observe that NSP achieves the best or second-best performance in all settings. The pre-training objective of NSP (Devlin et al., 2019) is to predict whether two sentences appear consecutively in a document. Thus, NSP is able to capture the intrinsic sequential dependencies between question pairs.

(3). Explicit word overlap facilitates retrieval of questions that are more likely to appear in the conversation history. We observe that TF-IDF and Levenshtein distance had fewer worst scores compared to Dense Retrieval. This may be because explicit word match relates to the paraphrased nature of a question.

4.6 Analysis of History Turns

We have explored how the different number of history turns affect knowledge transfer. From Figure 2, we obtain the following observations:

(1). When training on single-turn datasets without conversation synthesis (retrieved QA pairs = 0), inference with ground-truth conversation history leads to significant performance degradation. And as the turns of ground-truth history increases, the performance drops more severely.

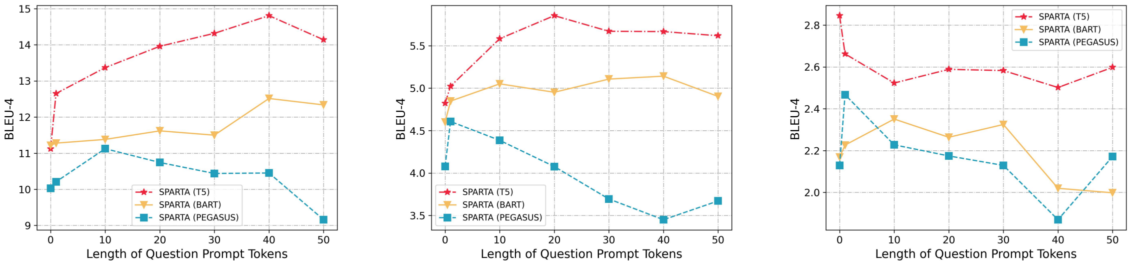
(2). When training on single-turn datasets with conversation synthesis, inference without ground-truth conversation history will also result in a significant performance drop.

(3). The BLEU-4 score increases up to a threshold (15 for CoQA, 2 for QuAC, and 2 for DoQA) as the number of retrieved question-answer pairs increases in single-turn training, and then a slight performance drop occurs. Larger question-answer pairs mean more relevant evidence, while potentially introducing more noise.

(4). The performance increases up to a threshold (9 for CoQA, 5 for QuAC, and 3 for DoQA) as the turns of ground-truth conversation history increase, followed by a very slight fluctuation. The difference here may be reflected in the average history turns in the CoQA, QuAC, and DoQA datasets shown in Table 2, respectively.

Prompts	MS MARCO			NewsQA			SQuAD		
	CoQA	QuAC	DoQA	CoQA	QuAC	DoQA	CoQA	QuAC	DoQA
SPARTA (T5) (-w QP$_{same}$)	5.33	3.04	**1.97**	**13.34**	**6.21**	**2.28**	**14.81**	**5.85**	2.66
- w QP$_{diff}$	4.82	2.92	1.76	12.99	5.92	2.24	14.10	5.41	2.81
- w/o SP	4.83	3.24	1.74	13.13	5.70	2.13	13.47	5.46	2.73
- w/o QP$_{input}$	**5.64**	**3.40**	1.79	13.01	5.79	2.24	13.93	5.36	2.69
- w/o QP$_{output}$	3.86	2.45	1.79	12.05	5.89	2.27	12.36	4.57	**2.85**

Table 6: Performance comparison of the BLEU-4 score with different prompt designs. SP refers to semantic prefixes. QP refers to Question Prompt. QP$_{input}$ and QP$_{output}$ refer to the prompt tokens used in the input and output, respectively. QP$_{same}$ means QP$_{input}$ and QP$_{output}$ are the same. QP$_{diff}$ means QP$_{input}$ and QP$_{output}$ are different.



(a) CoQA

(b) QuAC

(c) DoQA

Figure 3: Performance comparison of different models with different length of question prompt tokens with knowledge transferred from SQuAD.

4.7 Analysis of Prompt Design

To evaluate the relative importance of the conversational prompt, we explore several variants as shown in Table 6. The observations are as followings:

(1). We have observed that semantic prefixes are more beneficial than introducing new special tokens. Removing semantics prefixes leads to a performance drop in most cases.

(2). Both QP$_{input}$ and QP$_{output}$ contribute to the overall prompt architecture in most cases, with QP$_{output}$ contributing more to the CP than QP$_{input}$. Table 6 shows that removing QP$_{output}$ leads to a larger and consistent performance drop, while removing QP$_{input}$ even improves the transfer from MS MARCO to CoQA and QuAC. This may be because the trainable question prompts used in the output are closer to the target question and thus can be better optimized to guide the generation process.

(3). It is better for QP$_{input}$ and QP$_{output}$ to be the same than different. We can observe the prompt variant QP$_{same}$ achieves higher score compared to QP$_{diff}$. This result suggests that using the same question prompts in both input and output will further improve the semantic connections and thus enhance QP$_{output}$ guidance on question generation.

4.8 Analysis of Question Prompt Length

To study the effects of question prompt length on knowledge transfer, we train the models with prompt length varying in {0, 1, 10, 20, 30, 40, 50}. Figure 3 shows the BLEU-4 score of the different models plotted as a function of the question prompt length. We can observe the optimal prompt length varies across models and datasets. Especially it shows large fluctuations on the DoQA dataset.

In CoQA and QuAC, the BLEU-4 score of SPARTA (T5) increases as the prompt length increases to a threshold (40 for both CoQA, and 20 for QuAC), and then decreases. Similar trends can also be seen on SPARTA (BART) and SPARTA (PEGASUS). Among them, the optimal prompt length of SPARTA (PEGASUS) is shorter than other models. Longer prompts mean more trainable parameters and therefore improve expressiveness. But it also increases the computational and time overhead of both training and inference.

5 Related Work

Conversational Question Generation. Question Generation (QG) aims to generate natural questions with targeted answers from textual inputs. Early works were mainly rule-based systems (Heilman, 2011), using linguistic rules and hand-crafted templates to transform declarative sentences into interrogative sentences. With the popularity of neural networks, many research works (Du et al., 2017; Zhou et al., 2017; Zhao et al., 2018) adopt the encoder-decoder framework which combines attention (Bahdanau et al., 2015) and pointer (See et al., 2017) mechanisms to deal with the question generation problem in an end-to-end fashion.

More recently, conversational QG which involves multi-turn interactions has attracted increasing attention. (Gao et al., 2019) utilized the multi-source encoder-decoder model with coreference alignment to refer back and conversation flow to maintain coherent dialogue transition. (Pan et al., 2019) proposed a reinforced dynamic reasoning network to better understand what has been asked and what to ask next with the reward defined by the quality of answer predicted by a question-answering model. (Gu et al., 2021) designed a two-stage architecture that learns question-answer representations across multiple dialogue turns using flow propagation-based training strategy. (Wang et al., 2022b) proposed to distill knowledge from larger pre-trained LM into a tiny answer-guided network for efficient conversational question generation with fewer parameters and faster inference latency. (Do et al., 2022) utilized the top-p strategy to dynamically select the most relevant sentences and question-answer pairs from context and history respectively. (Ling et al., 2022) proposed a review and transit mechanism to identify question-worthy content for informative question generation in open-domain conversations. However, these models rely heavily on large-scale annotated conversations. As far as we know, this is the first research work to explore conversational question generation in the zero-shot learning setting.

Transfer Learning. Transfer learning focuses on adapting knowledge gained while solving one task to a different but related task (Pan and Yang, 2010). Fine-tuning is a commonly used approach in transfer learning, where a pre-trained model is adapted to a new task. The pre-trained models are typically trained on large-scale datasets, which can be either labeled images, such as ImageNet (Deng et al., 2009), or unlabeled text, such as BooksCorpus (Zhu et al., 2015) and Wikipedia. It has been successfully applied to many domains, such as computer vision and Natural Language Processing (NLP). In NLP, the well-known utilization of static word embedding (Mikolov et al., 2013; Pennington et al., 2014) and contextualized word embedding (Peters et al., 2018; Devlin et al., 2019), also called pre-trained LMs, in downstream task can also be referred as applications of transfer learning. In addition, prompt learning (Liu et al., 2021) is a new paradigm that can enhance the knowledge transfer capability by refactoring downstream tasks into the forms that are close to the pre-training objectives and thus alleviate the domain gap problem. More related works about zero-shot learning are detailed in Appendix C.

6 Conclusion

In this paper, we introduce a novel task setting, named ZeroCQG, which requires no human-labeled conversations for training. To solve ZeroCQG, we propose a multi-stage knowledge transfer framework SPARTA. Specifically, SPARTA synthesizes conversations for each single-turn QG instance to alleviate the domain gap between the two QG tasks. Besides, SPARTA leverage conversation prompting to reformulate conversational QG into a masked question-filling task similar to T5 to alleviate the domain gap between the objective of pre-trained LM and conversational QG. Extensive experiments conducted on the knowledge transfer from three single-turn QG datasets: MS MARCO, NewsQA, and SQuAD to three conversational QG datasets: CoQA, QuAC, and DoQA demonstrate the superior performance of our method.

7 Limitations

Although our proposed method achieves promising performance in the novel direction of ZeroCQG, it still has the following limitations: (1) retrieval-based conversation synthesis is limited to predefined question-answer pairs and may introduce repeated question-answer pairs with small differences (discussed in Appendix B.1). Future work may include exploring generative-based approaches to generate new and diverse question-answer pairs for better conversation synthesis. (2) Existing question transformation only explore one of the most common conversational characteristics, anaphora. However, other different characteristics, such as ellipsis, should also be considered in the future. (3) The conversation prompting has limitations when the domain gap becomes large (discussed in Sec. 4.4). More robust prompt learning should be explored in the future.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2022YFC3303600), National Natural Science Foundation of China (62137002, 62293553, 62293554, 61937001, and 62176209), Innovative Research Group of the National Natural Science Foundation of China (61721002), "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project, Natural Science Basic Research Program of Shaanxi (2023-JC-YB-293), the Youth Innovation Team of Shaanxi Universities.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa - accessing domain-specific faqs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7302–7314. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Xuan Long Do, Bowei Zou, Liangming Pan, Nancy F. Chen, Shafiq R. Joty, and Ai Ti Aw. 2022. Cohs-cqg: Context and history selection for conversational question generation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 580–591. International Committee on 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1342–1352. Association for Computational Linguistics.

Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuanjing Huang. 2022. CQG: A simple and effective controlled generation framework for multi-hop question generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6896–6906. 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4853–4862. Association for Computational Linguistics.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer McIntosh von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2131–2146. 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, EACL 2021, Online, April 19 - 23, 2021*, pages 2061–2070. Association for Computational Linguistics.

Michael Heilman. 2011. Automatic factual question generation from text. *Language Technologies Institute School of Computer Science Carnegie Mellon University*, 195.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 609–617. The Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert:

Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

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

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2105.10311*.

David Lindberg, Fred Popowich, John C. Nesbit, and Philip H. Winne. 2013. Generating natural language questions to support learning on-line. In *ENLG 2013 - Proceedings of the 14th European Workshop on Natural Language Generation, August 8-9, 2013, Sofia, Bulgaria*, pages 105–114. The Association for Computer Linguistics.

Yanxiang Ling, Fei Cai, Jun Liu, Honghui Chen, and Maarten de Rijke. 2022. Generating relevant and informative questions for open-domain conversations. *ACM Transactions on Information Systems (TOIS)*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems*, volume 1773 of *CEUR Workshop Proceedings*.

Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019. Reinforced dynamic reasoning for conversational question generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2114–2124. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *CoRR*, abs/2005.00628.

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

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul N. Bennett. 2020. Leading conversational search by suggesting useful questions. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1160–1170. ACM / IW3C2.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *CoRR*, abs/2209.07511.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers Inf. Technol. Electron. Eng.*, 19(1):10–26.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 191–200. Association for Computational Linguistics.

Luu Anh Tuan, Darsh J. Shah, and Regina Barzilay. 2020. Capturing greater context for question generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9065–9072. AAAI Press.

Wiebke Wagner. 2010. Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit - o'reilly media, beijing, 2009, ISBN 978-0-596-51649-9. *Lang. Resour. Evaluation*, 44(4):421–424.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022a. What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 22964–22984. PMLR.

Zekun Wang, Haichao Zhu, Ming Liu, and Bing Qin. 2022b. Tagnet: a tiny answer-guided network for conversational question generation. *International Journal of Machine Learning and Cybernetics*, pages 1–12.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: a survey. *Knowl. Inf. Syst.*, 64(12):3151–3195.

Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 418–428. ACM / IW3C2.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3901–3910. Association for Computational Linguistics.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Prompt consistency for zero-shot task generalization. *CoRR*, abs/2205.00049.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, volume 10619 of *Lecture Notes in Computer Science*, pages 662–671. Springer.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

A Details of Experiment Setup

A.1 Datasets

Single-turn Datasets: (1) **MS MARCO v1.1** (Nguyen et al., 2016) contains 100K question-answer pairs where questions are sampled from Bing's search query, answers are generated by human and passages are retrieved from Bing search engine. (2) **NewsQA** (Trischler et al., 2017) contains 120K question-answer pairs based on over 10K CNN news articles. (3) **SQuAD v1.1** (Rajpurkar et al., 2016) contains more than 100K question-answer pairs based on 536 Wikipedia articles. We can observe that each passages in the MS MARCO corresponds to only one question-answer pair, while each article in the NewsQA or SQuAD corresponds to multiple question-answer pairs.

Conversational Datasets: (1) **CoQA** (Reddy et al., 2019) contains 8k conversations about text passages from seven diverse domain. It consists of 127K question-answer pairs almost half of which refer back to conversation history using anaphors. (2) **QuAC** (Choi et al., 2018) contains 14K simulated information seeking dialogues where each interaction is between a student and a teacher on a hidden wikipedia passage. (3) **DoQA** (Campos et al., 2020) contains 2,437 real domain-specific information seeking dialogues collected from FAQ sites, such as Stack Exchange. This makes DoQA a challenging dataset with more coherent conversations and less factoid questions.

Our experiments are conducted with the accessible part. In particular, we use the validation set of CoQA and QuAC as the test set. For conversational datasets, we remove the examples with too little information in the answer, such as unknown, yes or no, to avoid the generation of questions full of randomness.

A.2 Baselines

BART (Lewis et al., 2020) is pre-trained using a denoising auto-encoding task which aims to recover corrupted documents to the original ones. BART can adapt well on both discriminative and generative tasks.

T5 (Raffel et al., 2020) is pre-trained to fill in randomly corrupted text spans. T5 is applicable to varies natural language processing problems by formulating tasks in text-to-text format.

PEGASUS (Zhang et al., 2020) is pre-trained to generate the most important sentences extracted from an unlabeled document. PEGASUS is mainly designed for abstractive summarization.

Model	Parameters	Hidden Size
BART	139M	768
T5	220M	768
PEGASUS	272M	768

Table 7: The parameter and hidden size of baseline models.

The baseline pre-trained LMs in this paper are all base sizes. When we employ question prompts on the baseline models, the number of extra parameters is equal to the corresponding hidden size multiplied by the length of question prompt tokens.

A.3 Implementation Details

We run our experiments on a GTX 3090 GPU. The model is trained with Adam optimizer (Kingma and Ba, 2015). The learning rate is initialized with 1e-4 and decays linearly. The batch size for each update is set as 8 with accumulation steps as 4. The maximum training epochs is 4. Early stopping is performed based on the BLEU score on the development set evaluated every 1000 training steps. We conduct beam search with the beam width 5. Inputs that exceed the maximum input length 768 will be truncated from right. The maximum and minimal decoding steps are equal to the length of question prompt tokens m plus 15 and 1 respectively. Decoding stops when the maximum decoding step is reached or the special end token is generated. After the generation finish, we will remove all the question prompt and special tokens. This paper conduct extensive experiments and ablation studies over different transfer setting varies at single-turn datasets, conversational datasets and pre-trained language models. Performing multiple runs for each experiment was not feasible due to computational and time constraints. Therefore, we use the same seed (42) for all experiments. The metrics, such as BLEU, ROUGE and METEOR, are calculated with the package released by (Sharma et al., 2017).

A.4 Hyperparameters

Table 8 shows the optimal maximum number of retrieved question-answer pairs across different transfer settings. These numbers are selected from {0, 1, 2, 5, 10, 15, 20} This corresponds to the main results shown in Table 3.

Source Dataset	CoQA	QuAC	DoQA
MS MARCO	10	10	2
NewsQA	5	2	2
SQuAD	15	2	2

Table 8: The optimal maximum number of retrieved question-answer pairs for different transfer settings.

B Supplementary of Experiment Results

B.1 Case Studies

Synthesized Conversation We present two examples each from MS MARCO, NewsQA and SQuAD, as shown in Table 9, Table 10 and Table 11 respectively. We can observe that the retrieved question-answer pairs are correlated and helpful for answering the original single-turn question. By replacing co-occurring entities with pronouns identified by the co-reference tool, we introduce anaphora between the transformed questions and the simulated conversation history. However, we also find the following limitations.

1. Retrieval-based methods may introduce repeated question-answer pairs with small differences. As shown in the first example in Table 9, many retrieved questions are asking for the same thing "how long to cook chicken legs in oven". For the first example in Table 10, the same answer "European travel guidebooks" are corresponding to three different but highly correlated question q_2, q_3 an q_4. These highly repetitive question-answer pairs rarely appear in a normal conversation.

2. Existing question transformation methods are not applicable when the context and retrieved question-answer pairs provides no co-reference information. For example, the retrieved question-answer pairs in Table 9 mentioned "chicken legs" and "kidney" multiple times respectively, but the method cannot transform such examples.

Generated Question We present ZeroCQG instances generated from different models with knowledge transferred from MS MARCO, NewsQA and SQuAD respectively. As shown in Table 12, there is a large discrepancy between ground truth questions and the questions generated by different models trained on MS MARCO, but relatively small differences between the questions generated by models trained on NewsQA or SQuAD. As shown in Table 13, our proposed SPARTA allows different pre-trained LMs to correctly generate *she* to refer to *Bouvier*, while the baseline model without SPARTA cannot. As shown in Table 14, although the questions generated by different models all mentioned some relevant content, such as "drunk driving", "traffic", "new years eve", etc., there are still some semantic differences compared to the ground-truth question. This may be because the target answer is a long sentence, rather than a short span, thus requiring more complex comprehension.

C Related Work of Zero-shot Learning

Zero-shot learning aims to learn a model that can generalize to new classes or tasks for which no training data is available. Large language models pre-trained with self-supervised objectives on unstructured text data have shown promising zero-shot generalization in a wide variety of downstream tasks with properly designed prompts (Radford et al., 2019; Liu et al., 2021; Wang et al., 2022a). (Zhou et al., 2022) proposed prompt consistency regularization on multiple prompts for a single task with unlabeled data to improve the zero-shot performance. (Shu et al., 2022) proposed test-time prompt tuning to learn adaptive prompts on the fly through a single test sample without requiring task-specific annotations. (Wei et al., 2022) proposed the instruction tuning to improve zero-shot performance of large language model on unseen tasks through natural language instructions. In this paper, we focus on the zero-shot conversational QG with knowledge transferred from single-turn QG and pre-trained LMs with synthesis and prompt strategies respectively.

Context
Directions . 1 Arrange chicken thighs/leg quarters skin side up in a shallow baking dish . 2 Sprinkle with garlic powder . 3 Drizzle about 1/2 teaspoon soy sauce on each piece . 4 Bake at 350 degrees Fahrenheit for 45 minutes to an hour , until the skin is crisp and brown and the meat is ready to fall off the bones .
Synthesized Conversation
Retrieved QA Pairs $\mathbf{q}_1$: how long do you cook chicken thighs on the stove $\mathbf{a}_1$: 18 mins $\mathbf{q}_2$: how long to cook whole chicken legs in oven $\mathbf{a}_2$: 45 to 50 minutes . $\mathbf{q}_3$: how long do i need to cook chicken thighs in the oven $\mathbf{a}_3$: 30 minutes . $\mathbf{q}_4$: how long to bake frozen chicken thighs in oven $\mathbf{a}_4$: 375 degrees . $\mathbf{q}_5$: how long to boil chicken legs $\mathbf{a}_5$: 15 minutes $\mathbf{q}_6$: how long to cook chicken legs in oven $\mathbf{a}_6$: 35 to 40 minutes $\mathbf{q}_7$: how long to cook chicken thighs in oven $\mathbf{a}_7$: 45 mins to an hour . $\mathbf{q}_8$: how long to cook a chicken thigh in the oven $\mathbf{a}_8$: 1 hour (or up to a day) . Preheat oven to 375 degrees . $\mathbf{q}_9$: how do you prepare chicken legs to bake and how long at what temp . $\mathbf{a}_9$: Preheat the oven to 400°F . Bake the chicken , uncovered , for 35 to 40 minutes or until the chicken is no longer pink inside . You can bake the chicken legs in a 375°F . oven , if desired . Increase the baking time to 45 to 50 minutes. 35 to 40 minutes . 400°F . $\mathbf{Q}_{10}$: how long to cook chicken legs in the oven $\mathbf{A}_{10}$: 35 to 40 minutes or until the chicken is no longer pink inside . **Question**: how long cooking chicken legs in the big easy **Answer**: 45 minutes to an hour

Context
Diabetic kidney disease , or diabetic nephropathy , is a complication of type 1 or type 2 diabetes caused by damage to the kidneys ' delicate filtering system . Your kidneys contain millions of tiny blood vessel clusters (glomeruli) that filter waste from your blood .
Synthesized Conversation
Retrieved QA Pairs $\mathbf{q}_1$: what is diabetic peripheral neuropathy $\mathbf{a}_1$: Peripheral neuropathy is nerve damage caused by chronically high blood sugar and diabetes , it leads to numbness , loss of sensation , and sometimes pain in your feet , legs , or hands , it is the most common complication of diabetes . $\mathbf{q}_2$: what is the primary function of nephrons in the kidney $\mathbf{a}_2$: Filtering the blood is the primary function of the kidney . $\mathbf{q}_3$: what is diabetic retinopathy $\mathbf{a}_3$: Diabetic retinopathy is the result of damage caused by diabetes to the small blood vessels located in the retina . $\mathbf{q}_4$: what hormones do the kidneys produce and what are their function $\mathbf{a}_4$: The kidneys remove waste products and excess water from the body and so help to regulate blood pressure . $\mathbf{q}_5$: what is involved in a kidney scan $\mathbf{a}_5$: 1 . Assessment of the blood flow through the kidneys . 2 See how a transplanted kidney is working . 3 Check the extent of kidney damage . 4 Find an obstruction in the kidney or ureter 5 Find growths in the kidneys $\mathbf{q}_6$: what is polycystic kidney disease $\mathbf{a}_6$: Polycystic kidney disease (PKD) is an inherited disorder in which clusters of cysts develop primarily within your kidneys . Cysts are noncancerous round sacs containing water-like fluid . $\mathbf{q}_7$: what causes kidney and liver failure in dogs $\mathbf{a}_7$: Bacteria associated with advanced dental disease enter the blood stream and invades multiple organs , causing irreversible damage to the heart , liver and kidneys . $\mathbf{q}_8$: what is polycystic kidney disease symptoms $\mathbf{a}_8$: 1 High blood pressure . 2 Back or side pain . 3 Headache . 4 Increase in the size of your abdomen.5 Blood in your urine . 6 Frequent urination . 7 Kidney stones . 8 Kidney failure.9 Urinary tract or kidney infections . $\mathbf{q}_9$: what are kidneys made of $\mathbf{a}_9$: The kidney is made of a majority of cells called nephrons . $\mathbf{q}_{10}$: what is a kidneys function $\mathbf{a}_{10}$: To remove waste products and excess fluid from the body . **Question**: what is a diabetic kidney **Answer**: A complication of type 1 or type 2 diabetes caused by damage to the kidneys ' delicate filtering system .

Table 9: Our synthesized conversation examples from MS MARCO dataset. The original single-turn question-answer pair can be treated as turn 11 of the synthesized conversation.

Context

You ŕe all alone , surrounded by dank mist and the realization that it was these monks who kept literacy alive in Europe . To give you an idea of their importance , Charlemagne , who ruled much of Europe in the year 800 , imported Irish monks to be his scribes . Rounding Slea Head , the point in Europe closest to America , the rugged coastline offers smashing views of deadly black-rock cliffs and the distant Blasket Islands . The crashing surf races in like white horses , while longhaired sheep graze peacefully on the green hillside . Study the highest fields , untouched since the planting of 1845 , when the potatoes never matured and rotted in the ground . The great famine of that year , through starvation or emigration , nearly halved Ireland ś population . Because its endearing people have endured so much , Ireland is called " The Terrible Beauty . " Take your time at the Gallaras Oratory , circa A.D. 800 , the sightseeing highlight of your peninsula tour . One of Ireland ś best-preserved early Christian churches , its shape is reminiscent of an upturned boat . Its watertight dry-stone walls have sheltered travelers and pilgrims for 1,200 years . From the Oratory , continue up the rugged one-lane road to the crest of the hill and then coast back to Dingle Town – hungry , thirsty , and ready for a pint . **Rick Steves** writes European travel guidebooks and hosts travel shows on public television and public radio . E-mail him at rick@ricksteves.com, or write to him c/o P.O. Box 2009 , Edmonds , Wash. 98020 .

Synthesized Conversation

Retrieved QA Pairs
$\mathbf{q}_1$: What stations do **his** TV series air on ?
$\mathbf{a}_1$: public television
$\mathbf{q}_2$: What kind of books does Rick Steve write ?
$\mathbf{a}_2$: European travel guidebooks
$\mathbf{q}_3$: What types of books does Rick Steves write ?
$\mathbf{a}_3$: European travel guidebooks
$\mathbf{q}_4$: What does Rick Steves write ?
$\mathbf{a}_4$: European travel guidebooks
$\mathbf{q}_5$: What does **Rick Steves ’** company do ?
$\mathbf{a}_5$: writes European travel guidebooks and hosts travel shows on public television and public radio
Transformed Question: Where does **his** show air ?
Answer: public television

Context

-LRB- CNN -RRB- – Author **Arthur C. Clarke** , whose science fiction and non-fiction works ranged from the script for " 2001 : A Space Odyssey " to an early proposal for communications satellites , has died at age 90 , associates have said . Visionary author Arthur C. Clarke had fans around the world . Clarke had been wheelchair-bound for several years with complications stemming from a youthful bout with polio and had suffered from back trouble recently , said Scott Chase , the secretary of the nonprofit Arthur C. Clarke Foundation . **He** died early Wednesday – Tuesday afternoon ET – at a hospital in Colombo , Sri Lanka , where he had lived since the 1950s , Chase said . " He had been taken to hospital in what we had hoped was one of the slings and arrows of being 90 , but in this case it was his final visit , " he said . In a videotaped 90th birthday message to fans , Clarke said he still hoped to see some sign of intelligent life beyond Earth , more work on alternatives to fossil fuels – and " closer to home , " an end to the 25-year civil war in Sri Lanka between the government and ethnic Tamil separatists . " I dearly wish to see lasting peace established in Sri Lanka as soon as possible , " he said . " But I ḿ aware that peace can not just be wished – it requires a great deal of hard work , courage and persistence . " Clarke and director Stanley Kubrick shared an Academy Award nomination for best adapted screenplay for " 2001 . " The film grew out of Clarke ś 1951 short story , " The Sentinel , " about an alien transmitter left on the moon that ceases broadcasting when humans arrive . As a Royal Air Force officer during World War II , Clarke took part in the early development of radar . In a paper written for the radio journal " Wireless World " in 1945 , he suggested that artificial satellites hovering in a fixed spot above Earth could be used to relay telecommunications signals across the globe . He is widely credited with introducing the idea of the communications satellite , the first of which were launched in the early 1960s . But he never patented the idea , prompting a 1965 essay that he subtitled , " How I Lost a Billion Dollars in My Spare Time . " His best-known works , such as " 2001 " or the 1953 novel " Childhood ś End , " combined the hard science he learned studying physics and mathematics with insights into how future discoveries would change humanity . David Eicher , editor of Astronomy magazine , told CNN that Clarke ś writings were influential in shaping public interest in space exploration during the 1950s and 6́0s . Watch how Clarke stands among sci-fi giants "

Synthesized Conversation

Retrieved QA Pairs
$\mathbf{q}_1$: Arthur C. Clarke dies in Sri Lanka at age 90 , aide says
$\mathbf{a}_1$: whose science fiction and non-fiction works
$\mathbf{q}_2$: Who died in Sri Lanka ?
$\mathbf{a}_2$: Arthur C. Clarke
$\mathbf{q}_3$: What did he and Stanley Kubrick share ?
$\mathbf{a}_3$: Academy Award nomination for best adapted screenplay for “ 2001’
$\mathbf{q}_4$: Clarked lived in Sri Lanka since when ?
$\mathbf{a}_4$: the 1950s
$\mathbf{q}_5$: Where did **Arthur C. Clarke** die ?
$\mathbf{a}_5$: Colombo , Sri Lanka
Transformed Question: Where did **he** live ?
Answer: Colombo , Sri Lanka

Table 10: Our synthesized conversation examples from NewsQA dataset. The original single-turn question-answer pair can be treated as turn 6 of the synthesized conversation. The co-reference mentions are marked with **underline**.

Context

In December , Beyoncé along with a variety of other celebrities teamed up and produced a video campaign for “ Demand A Plan ” , a bipartisan effort by a group of 950 US mayors and others designed to influence the federal government into rethinking its gun control laws , following the Sandy Hook Elementary School shooting . **Beyoncé** became an ambassador for the 2012 World Humanitarian Day campaign donating her song “ I Was Here ” and its music video , shot in the UN , to the campaign . In 2013 , it was announced that **Beyoncé** would work with Salma Hayek and Frida Giannini on a Gucci “ Chime for Change ” campaign that aims to spread female empowerment . The campaign , which aired on February 28 , was set to her new music . A concert for the cause took place on June 1 , 2013 in London and included other acts like Ellie Goulding , Florence and the Machine , and Rita Ora . In advance of the concert , **she** appeared in a campaign video released on 15 May 2013 , where she , along with Cameron Diaz , John Legend and Kylie Minogue , described inspiration from their mothers , while a number of other artists celebrated personal inspiration from other women , leading to a call for submission of photos of women of viewers ’ inspiration from which a selection was shown at the concert . Beyoncé said about her mother Tina Knowles that her gift was “ finding the best qualities in every human being . ” With help of the crowdfunding platform Catapult , visitors of the concert could choose between several projects promoting education of women and girls . Beyoncé is also taking part in “ Miss a Meal ” , a food-donation campaign , and supporting Goodwill charity through online charity auctions at Charitybuzz that support job creation throughout Europe and the U.S .

Synthesized Conversation

Retrieved QA Pairs

q$_1$: Beyonce was speaking about whom when she said her gift was “ finding the best qualities in every human being . “ ?
a$_1$: her mother
q$_2$: Who did Beyoncé work with in 2013 on the Chime for Change campaign ?
a$_2$: Salma Hayek and Frida Giannini
q$_3$: What is the name of the campaign that Beyoncé and others are involved in that deals with gun control ?
a$_3$: Demand A Plan
q$_4$: Beyonce is contributing to which food-donation campaign ?
a$_4$: Miss a Meal
q$_5$: What song did **Beyonce** contribute to the campaign ?
a$_5$: I Was Here
Transformed Question: What song did **she** donate to the 2012 World Humanitarian Day campaign ?
Answer: I Was Here

Context

New Zealand has a strong hunting culture . The islands making up New Zealand originally had no land mammals apart from bats . However , once Europeans arrived , **game animals** were introduced by acclimatisation societies to provide New Zealanders with sport and a hunting resource . Deer , pigs , goats , rabbits , hare , tahr and chamois all adapted well to the New Zealand terrain , and with no natural predators , their population exploded . Government agencies view **the animals** as pests due to their effects on the natural environment and on agricultural production , but hunters view **them** as a resource .

Synthesized Conversation

Retrieved QA Pairs

q$_1$: What were the the only land mammal in New Zealand ?
a$_1$: bats
q$_2$: What was the only land mammal native to New Zealand ?
a$_2$: bats
q$_3$: Why did the population of pigs and rabbits explode in New Zealand ?
a$_3$: no natural predators
q$_4$: Game animals were introduced here by whom ?
a$_4$: acclimatisation societies
q$_5$: Why were **game animals** introduced by acclimatisation societies ?
a$_5$: to provide New Zealanders with sport and a hunting resource
Transformed Question: What resulted having no natural predators for **them** introduced ?
Answer: their population exploded

Table 11: Our synthesized conversation examples from SQuAD dataset. The original single-turn question-answer pair can be treated as turn 6 of the synthesized conversation. The co-reference mentions are marked with **underline**.

Context
Kendra and Quinton travel to and from school every day . Kendra lives further from the bus stop than Quinton does , stops every morning at Quinton 's house to join him to walk to the bus stop . Every afternoon , after school , when walking home from the bus stop they go in for cookies and milk that Quinton 's mother has ready and waiting for them . Quinton ca n't eat cheese or cake so they had the same snack every day . They both work together on their homework and when they are done they play together . Kendra always makes sure to leave in time to get home for dinner . She does n't want to miss story time which was right before bedtime . One morning Kendra walked up to Quinton 's house , she thought something might be wrong because normally Quinton was waiting outside for her and on this morning he was not to be found . Kendra went up to the door and knocked . She waited and waited and yet no one answered . She saw that Quinton 's mother 's car was n't in their driveway which was weird . She waited for a few bit looking up and down the block and getting worried when Quinton was nowhere to be found . Kendra did n't want to miss the bus to school and hurried off to make it in time . The bus driver saw that she was upset and that Quinton was not with her that morning . She told him what happened and he said that he was sure that everything would be okay . Kendra got to school , ran to her teacher and told him what happened that morning . The teacher smiled and told her not to worry , Quinton 's mother had called and he was going to the dentist and would be at school after lunch and that she would see him at the bus stop like normal tomorrow .
Conversation History
$\mathbf{q}_1$: Where do Quinton and Kendra travel to and from every day ? $\mathbf{a}_1$: school $\mathbf{q}_2$: What do they do every afternoon after school ? $\mathbf{a}_2$: go to Quentin 's house $\mathbf{q}_3$: What does Kendra not want to miss ? $\mathbf{a}_3$: story time $\mathbf{q}_4$: When is that ? $\mathbf{a}_4$: right before bedtime $\mathbf{q}_5$: What happened when Kendra knocked on Quinton 's door ? $\mathbf{a}_5$: no one answered $\mathbf{q}_6$: What did the bus driver see ? $\mathbf{a}_6$: that she was upset
Answer $\mathbf{a}_7$: everything would be okay
Ground Truth Question $\mathbf{q}_7$: what did he say ?
Generated Question with Knowledge Transferred from MS MARCO
PEGASUS: what happens to quinton when he goes to school **BART**: what happened to kendra after she got home from school **T5**: what happened to kendra when quinton was not with her **SPARTA (PEGASUS)**: what happened to quinton **SPARTA (BART)**:what did the bus driver tell kendra that she was missing **SPARTA (T5)**: what did the bus driver see
Generated Question with Knowledge Transferred from NewsQA
PEGASUS: what did the bus driver say ? **BART**: what did the bus driver promise kendra ? **T5**: what did the bus driver say ? **SPARTA (PEGASUS)**: what did the bus driver say ? **SPARTA (BART)**: what did the bus driver say ? **SPARTA (T5)**: what did the bus driver say ?
Generated Question with Knowledge Transferred from SQuAD
PEGASUS: what did quinton 's teacher tell him ? **BART**: what did the bus driver say after kendra told him about **T5**: what did the bus driver say he was sure of ? **SPARTA (PEGASUS)**: what did the bus driver say ? **SPARTA (BART)**: what did the bus driver say ? **SPARTA (T5)**: what did he say ?

Table 12: An example of generated questions in the CoQA by different models with knowledge transferred from MS MARCO, NewsQA and SQuAD respectively.

Context
In the fall of 1947 , Bouvier entered Vassar College in Poughkeepsie , New York . She had wanted to attend Sarah Lawrence College , closer to New York City , but her parents insisted that she choose the more geographically isolated Vassar . Bouvier was an accomplished student who participated in the school 's art and drama clubs and wrote for its newspaper . Due to her dislike for the college , she did not take an active part in its social life and instead traveled back to Manhattan on the weekends . She had made her society debut in the summer before entering college and became a frequent presence in New York social functions . Hearst columnist Igor Cassini dubbed her the " debutante of the year " . Bouvier spent her junior year (1949-1950) in France – at the University of Grenoble in Grenoble , and at the Sorbonne in Paris – in a study-abroad program through Smith College . Upon returning home , she transferred to George Washington University in Washington , D.C. , graduating with a Bachelor of Arts degree in French literature in 1951 . During the early years of her marriage to John F. Kennedy , she took continuing education classes in American history at Georgetown University in Washington , D.C . While attending George Washington , Bouvier won a twelve-month junior editorship at Vogue magazine ; she had been selected over several hundred other women nationwide . The position entailed working for six months in the magazine 's New York City office and spending the remaining six months in Paris . Before beginning the job , Bouvier celebrated her college graduation and her sister Lee 's high school graduation by traveling with her to Europe for the summer . The trip was the subject of her only autobiography , One Special Summer , co-authored with Lee ; it is also the only one of her published works to feature Jacqueline 's drawings . On her first day at Vogue , the managing editor advised her to quit and go back to Washington . According to biographer Barbara Leaming , the editor was concerned about Bouvier 's marriage prospects ; she was 22 years of age and was considered too old to be single in her social circles . Bouvier followed the advice , left the job and returned to Washington after only one day of work . Bouvier moved back to Merrywood and was hired as a part-time receptionist at the Washington Times-Herald . A week later , she approached editor Frank Waldrop and requested more challenging work ; she was given the position of " Inquiring Camera Girl " , despite Waldrop 's initial concerns about her competence . The position required her to pose witty questions to individuals chosen at random on the street and take their pictures for publication in the newspaper alongside selected quotations from their responses . In addition to the random " man on the street " vignettes , she sometimes sought interviews with people of interest , such as six-year-old Tricia Nixon . Bouvier interviewed Tricia a few days after her father Richard Nixon was elected to the vice presidency in the 1952 election . During this time , Bouvier was also briefly engaged to a young stockbroker , John G. W. Husted , Jr. After only a month of dating , the couple published the announcement in The New York Times in January 1952 . She called off the engagement after three months , because she had found him " immature and boring " once she got to know him better .
Conversation History
$\mathbf{q}_1$: where did she go to College ? $\mathbf{a}_1$: Bouvier entered Vassar College in Poughkeepsie , New York .
Answer $\mathbf{a}_2$: 1947 ,
Ground Truth Question $\mathbf{q}_2$: what year did she go to college ?
Generated Question with Knowledge Transferred from MS MARCO
PEGASUS: when was bouvier born **BART**: when did jacqueline bouvier go to college **T5**: when did jacqueline bouvier enter vogue **SPARTA (PEGASUS)**: when did bouvier go to college **SPARTA (BART)**: when did she go to vogue **SPARTA (T5)**: when did she go to college
Generated Question with Knowledge Transferred from NewsQA
PEGASUS: what year did bouvier graduate from george washington university ? **BART**: when did she work for vogue ? **T5**: when did bouvier graduate ? **SPARTA (PEGASUS)**: when did she go to college ? **SPARTA (BART)**: when did she go to college ? **SPARTA (T5)**: when did she go to college ?
Generated Question with Knowledge Transferred from SQuAD
PEGASUS: in what year did bouvier enter vassar college ? **BART**: when did bouvier enter vassar college ? **T5**: when did bouvier enter vassar college ? **SPARTA (PEGASUS)**: when did she go to vassar college ? **SPARTA (BART)**: when did she enter vassar college ? **SPARTA (T5)**: when did bouvier enter vassar college ?

Table 13: An example of generated questions in the QuAC by different models with knowledge transferred from MS MARCO, NewsQA and SQuAD respectively.

Context
It should n’t be any worse than usual - it might even be a bit light ; Larchmont is a ways north of NYC proper , so I would n’t expect significant NYE related backups there . One thing that you should be wary about however , is drunk drivers ! There will probably be more of them on the road than usual that night , so be cautious and alert . (Similarly , there will probably be an above average number of police along the highway looking to catch said drunk drivers - and they wo n’t bee averse to writing you a citation for any other infraction which they might observe . Drive safely !)
Conversation History
$\mathbf{q}_1$: How bad is traffic from Boston to New York City on New Years Eve ? $\mathbf{a}_1$: It should n’t be any worse than usual - it might even be a bit light $\mathbf{q}_2$: How many hours would it take to go from Boston to Larchmont , NY ? $\mathbf{a}_2$: Larchmont is a ways north of NYC proper , so I would n’t expect significant NYE related backups there $\mathbf{q}_3$: would traffic be better before or after midnight ? $\mathbf{a}_3$: One thing that you should be wary about however , is drunk drivers
Answer $\mathbf{a}_4$: There will probably be more of them on the road than usual that night
Ground Truth Question $\mathbf{q}_4$: are there a lot of drunk drivers on new years eve ?
Generated Question with Knowledge Transferred from MS MARCO
PEGASUS: how many drunk drivers in nyc **BART**: drinking drivers in nyc **T5**: how many drunk drivers in larchmont nyc **SPARTA (PEGASUS)**: how bad is traffic in larchmont ny on new years eve **SPARTA (BART)**: would traffic be bad at larchmont nyc **SPARTA (T5)**: would traffic be worse in larchmont nyc on
Generated Question with Knowledge Transferred from NewsQA
PEGASUS: what should you be cautious about ? **BART**: what should you be careful about ? **T5**: what should you be cautious of ? **SPARTA (PEGASUS)**: will traffic be better before or after midnight ? **SPARTA (BART)**: what is the problem with drunk drivers ? **SPARTA (T5)**: would the traffic be better before or after midnight ?
Generated Question with Knowledge Transferred from SQuAD
PEGASUS: what is one thing that you should be wary of ? **BART**: are there more drunk drivers on the road ? **T5**: why should you be cautious and alert ? **SPARTA (PEGASUS)**: would there be more or less of them on the road ? **SPARTA (BART)**: are there more drunk drivers on the highways in new york **SPARTA (T5)**: would traffic be better or worse on new years eve ?

Table 14: An example of generated questions in the DoQA by different models with knowledge transferred from MS MARCO, NewsQA and SQuAD respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 4.7, Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 4, Appendix B

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.2, A.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.5, Section 4.7, Appendix A.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix A.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.1, Appendix A.3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.