Type-dependent prompt CycleQAG : Cycle consistency for Multi-hop Question Generation

Seungyeon Lee[1], **Minho Lee**[1,2,*]
[1]Department of Artificial Intelligence, Kyungpook National University
[2]ALI Co., Ltd.
statai3237@knu.ac.kr, mholee@knu.ac.kr

Abstract

Multi-hop question generation (QG) is the process of generating answer related questions, which requires aggregating multiple pieces of information and reasoning from different parts of the texts. This is opposed to single-hop QG which generates questions from sentences containing an answer in a given paragraph. Single-hop QG requires no reasoning or complexity, while multi-hop QG often requires logical reasoning to derive an answer related question, making it a dual task. Not enough research has been made on the multi-hop QG due to its complexity. Also, a question should be created using the question type and words related to the correct answer as a prompt so that multi-hop questions can get more information. In this view, we propose a new type-dependent prompt cycleQAG (cyclic question-answer-generation), with a cycle consistency loss in which QG and Question Answering (QA) are learnt in a cyclic manner. The novelty is that the cycle consistency loss uses the negative cross entropy to generate syntactically diverse questions that enable selecting different word representations. Empirical evaluation on the multi-hop dataset with automatic and human evaluation metrics outperforms the baseline model by about 10.38% based on ROUGE score.

1 Introduction

Question Generation (QG) problem that automatically generates a question from a given document with a correct answer is a challenging and an interesting task in the field of natural language processing (Chan and Fan, 2019; Pan et al., 2021; Yu et al., 2020; Dong et al., 2019). With the advent of deep learning, the pre-trained language models (Devlin et al., 2019; Radford et al., 2018; Liu et al., 2019; Raffel et al., 2020; Clark et al., 2020; Peng et al., 2021) were proposed, after which the study of natural language processing began to develop rapidly. These works not only use single-hop QA dataset such as SQuAD (Rajpurkar et al., 2016), which is a representative of research on Question Answering (QA), but also the multi-hop QA dataset such as HotpotQA (Yang et al., 2018). The QA dataset consists of (Context, Question, Answer) pairs along with a lot of QA data, that enables research on Automatic Question Generation (AQG). Most of the question generation methods evaluated questions using the single-hop QA datasets (Duan et al., 2017; Du et al., 2017; Sultan et al., 2020). However, in real-world situations, the questions can be very complex and sometimes require a complicated reasoning process (Gupta et al., 2020; Pan et al., 2021; Yu et al., 2020).

Multi-hop QG requires combining several pieces of information and reasoning over them to derive an answer related a question, making it a dual task. Multi-hop questions that can be encountered in the real world are largely divided into two types, bridges and comparisons. As shown in Fig. 1, the middle side is an example of a bridge-type question. When the question is "Who played Selby Wall in the film that Charlize Theron won an Academy Award for?", the first thing we need to know is what film Theron won the Academy Award. Second, we should be able to obtain information about the actors who played Selby Wall among the actors in the movie. Here, Monster, the movie that connects the two, serves as a bridge. On the other hand, the comparison type shown on the right side of Fig. 1 is to create a question that can be answered by comparing two objects.

Some of the methods for multi-hop QG transform the input text into an intermediate representation such as a parsing tree (Ji et al., 2021), and then convert the resulting form into a question by some well-designed templates or general rules. In (Gupta et al., 2020), they use multi-task learning with an auxiliary loss for sentence-level supporting fact prediction. Graph-based methods (Su et al.,

*Corresponding author

Proceedings of the 29th International Conference on Computational Linguistics, pages 6301–6314
October 12–17, 2022.

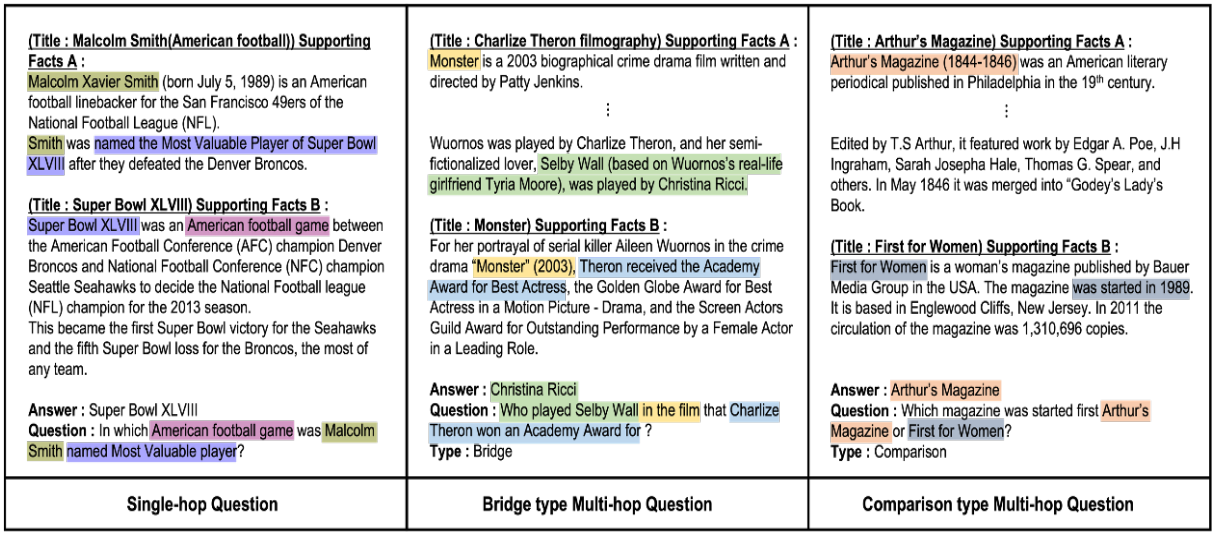

Figure 1: Examples of Single-hop QAG and Multi-hop QAG pair in HotpotQA (Yang et al., 2018) dataset. Multi-hop QG for reasoning multi-hop by finding the contact points between the given Answer and supporting facts A and B. The left is a bridge type, and the right is an example of a comparison type multi-hop question. Both question types are multi-hop, but generate questions with different characteristics.

2020; Kumar et al., 2019) used graph convolution networks(GCNs) to capture dependencies among different pieces of information for reasoning. However, those approaches should predefine graphs only from the question and candidate answers, lacking much key information for multi-hop reasoning.

To alleviate the above problems, we introduce the automated question generation to handle the lack of information for multi-hop reasoning and solve the predefined graph issue in an end-to-end manner. In this view, we propose a type-dependent prompt CycleQAG, which provides additional information and loss of cycle consistency for multi-hop QG. At first, an intermediate task of cyclically learning QG and QA is performed before the fine-tuning stage. In this process, we use cycle consistency loss, and in particular, we introduce the negative cross entropy (NCE), which is used to increase the lexical diversity of multi-hop questions. And in the final step, we use a prompt-based fine-tuning method that maximizes the information obtained from the intermediate task by giving information that can be provided according to the types of questions (eg, type and answer related words). Using the proposed model, we can generate complex questions by an end-to-end manner with semantically similar but diverse vocabulary.

We use the HotpotQA distractor setting and perform experiments with a multi-hop QA dataset. The proposed model outperforms the baseline models in automatic evaluation results such as ROUGE (Lin, 2004) for quantitative evaluation. However, qualitative part of the multi-hop question is evaluated using fluency, relevance, answerability, complexity, and diversity for human evaluation. We evaluate the diversity of vocabulary through qualitative evaluation, and show examples to prove this.

2 Related Works

Question Answering. Machine reading comprehension (MRC) is originally inspired by language proficiency tests, and the machine aims to answer a question by reading and understanding a given context (Zhu et al., 2021). (Seo et al., 2016) introduced the Bi-Directional Attention Flow (BIDAF) network, and proposed a model structure to represent contexts at various levels using a multi-level hierarchical structure. QANet (Yu et al., 2018) models an architecture that does not require a recurrent network and only consists of convolution model and self-attention. Recently, research on new multi-hop QA datasets that require more complex and diverse information, such as HotpotQA (Yang et al., 2018), HybridQA (Chen et al., 2020), MultiModalQA (Talmor et al., 2021), is being actively conducted. For example, (Xiong et al., 2021) used a simple recursive framework to solve open domain multi-hop QA, and configured the model to use dense search for multi-hop setups. In this work, we propose a method to solve multi-hop QA reasoning by a top-down approach to find a specific answer in a whole context.

Question Generation. The ultimate goal of the QG task is to automatically generate questions from texts or knowledge data. With the advent of machine reading comprehension datasets such as SQuAD and pre-trained language models, QG research is conducting the multi-hop reasoning research that deals with more complex and inference-demanding thorny questions, to mimic humans (Pan et al., 2020; Yu et al., 2020; Pan et al., 2021). (Yu et al., 2020) proposed a whole generator evaluator network for generating questions by creating an entity graph to integrate various entities scattered in the texts. (Pan et al., 2021) proposed a multi-hop QG method that used predefined basic operators to search, generate, and aggregate information of each input according to the types of inputs. They also defined and used six inference types of reasoning graphs. In particular, an off-the-shelf template was used for generating a comparison type question that compares two subjects. Although such pre-defined templates or structured models can generate accurate questions for given data, they can be fatal in both quantitative and qualitative aspects when new complex data are given. To overcome some of these issues, we propose a new end-to-end approach to generate multi-hop questions.

Dual task of QA and QG. QA and QG are separate but closely related tasks. In (Tang et al., 2017), they jointly train the two tasks by exploiting the probabilistic correlation between QA and QG. In particular, the parameterized model was jointly trained to minimize the loss function according to the constraints. (Duan et al., 2017) used question generation as an auxiliary task to improve the text-based QA task. They calculated the relevance score between the input question and the answer candidates, and chose the highest relevance score as a correct answer. (Sun et al., 2020) generated additional training instances to further improve the QA model in (Tang et al., 2017), each consists of a question, an answer, and a label for a category. In addition, the question was created by clamping the answer part and providing the answer to the QG model. Many efforts have been made to improve each module by using QA and QG together. In this view, we not only propose a method of using cycle consistency to increase the robustness of QA and QG but also introduce the NCE that increases the diversity of questions.

3 Proposed model

The proposed model for question generation includes an intermediate task execution phase before the fine-tuning step. In the intermediate task, QA and QG are trained to have cycle consistency, where question paraphrasing and similarity are additionally used to increase the robustness of the question generation. We focus on using the multi-hop QG and QA together as a supplement to increase the performance of QG. We define QA as the top-down approach to find a right answer, and QG as the bottom-up approach to use abundant information from entities or sentences. The overall framework of our proposed model is explained in Fig. 2.

3.1 Intermediate Task Training

The intermediate task is to fine-tune the pre-trained model for a task of interest before fine-tuning. We fine-tune the models used for the intermediate task based on the Google-T5 model (Raffel et al., 2020). In the intermediate task stage, QG and QA learn the "cycle consistency". This property was first introduced in the back-translation by Brislin (Brislin, 1970). This translates English to French, and translates the translated French back to English so that the original sentence can be reconstructed. Mathematically, this can be represented as a translator $G : X \rightarrow Y$, $F : Y \rightarrow X$, where G and F are inverse of each other, and are connected like a bijection. Inspired by those properties, when generating questions and answers in the intermediate stage, the proposed QG and QA model uses a cycle-consistency loss that exchanges inputs and outputs in the reverse direction, respectively.

3.1.1 Question Generation

We attempt to handle the multi-hop question generation using the answers and the context. First, we use Google-T5 (Raffel et al., 2020) as the baseline model to automatically generate the output question with a given input answer. While generating a question, as introduced in (Chan and Fan, 2019), there may be more than one instance of the same tokens as the correct answer in the context, then it may be confusing for the model to focus on question generation, we surround the annotated answer span tokens in the context with two tokens. Therefore, the format of the input can be represented as <sep>

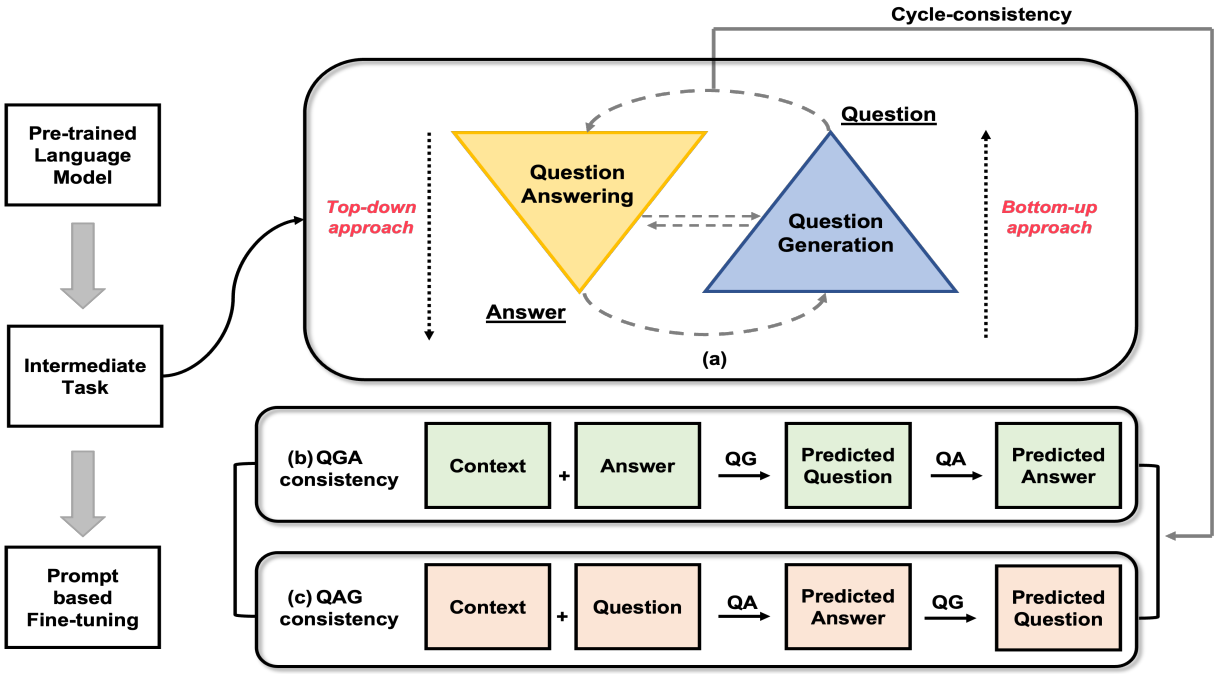

Figure 2: The framework of our proposed model. The proposed model configures QA and QG to learn cycle consistency (a). In the intermediate task, it is possible to create richer questions and more accurate answers through QA and QG interaction. In (b) and (c), the process of QGA-consistency and QAG-consistency to which cycle consistency is applied is shown in detail.

c_1,c_2,c_3 <hl> answer <hl> c_4,c_5, ..., c_m <sep>, where c_i is token of the context. The question generation module of the proposed model not only generates single-hop questions but also produces complex multi-hop questions that require multiple pieces of information. We train the model in such a way that the proposed QG module generates semantically similar and syntactically diverse questions. In particular, we introduce the NCE and the cross-entropy(CE) loss to train a QG model that generates more lexically diverse questions. It controls the probability of adopting information so that it can be semantically similar but syntactically diverse. When training QG, unlike in previous studies, we use the NCE. The probability of occurrence of a word can be lowered, but the essential meaning of a word does not change, thereby enriching the diversity of meaning.

Negative Cross Entropy loss (NCE loss). We use the NCE loss to generate questions more diverse. In general, most studies use the cross entropy (CE) to better train the model to maximize the probability of the correct class (Marek et al., 2021). However, as shown in Fig. 2 (a), in order to generate questions with a more diverse vocabulary for the bottom-up QG, we use the NCE to flatten the word occurrence probability distribution and increase the diversity of vocabulary. In this work, we use the NCE loss to reduce the distance between the predicted value and the actual value such that the generated question has similar meaning with increased lexical diversity.

Question Paraphrasing. In addition, to increase the robustness of the QG module, several questions are generated through a question paraphrasing process. This enables QG with the same meaning but different expressions. We use a paraphrasing model fine-tuned in advance using a Google-T5 model which was Quora Question Pair (QQP) * as the question paraphrasing dataset.

Similarity for generated paraphrasing question. Since it is important to ensure that the meaning of the generated question is the same even if a question with various expressions is generated, the similarity of the generated question should be measured. To find the similarity among paraphrased questions, Sentence-BERT (SBERT) is used, and the overall method is the same as that introduced in (Reimers and Gurevych, 2019), but uses T5 instead

*https://www.kaggle.com/c/quora-question-pairs.

of BERT [†]. We set the similarity value between 0 and 1. The similarity value obtained during the learning process is converted to $1 - similarity$ to train the model in such a way that the loss value decreases as the similarity increases.

Algorithm 1 Procedure of CycleQAG Framework

Input: Context = $(c_1, \ldots, c_n)$, Answer for QG
Context = $(c_1, \ldots, c_n)$, Question for QA
Output: Multi-hop Question

1: Initial QG $\leftarrow$ Generate question by QG Input
2: Paraphrasing the generated question
3: Calculate the cosine-similarity between the generated questions and the original question
4: Initial QA $\leftarrow$ Generate answer by QA Input
5: **for** $k \leftarrow 1$ to N **do**
6: $\quad C_k \leftarrow cycle(QA, QG)$
7: **end for**
8: **return** Multi-hop question, answer

3.1.2 Question Answering

To build a model that infers an answer using a given (question, passage) pair, the QA model is also trained using the Google-T5 model. In this paper, QA is used to improve the performance of QG, where the ultimate goal of the QA model is to approach the sentence related to the question in a given paragraph and access a correct answer.

3.1.3 Answer related words generation

Multi-hop QG requires more than two pieces of information when asking a question that can fit a correct answer. This requires gathering information in order to create a question. To this end, we use the title information of each supporting paragraph provided as in Fig. 1 to generate words related to a correct answer. Usually, the titles help by providing significant information in generating questions to arrive at a correct answer. We explain this in the appendix C with examples.

3.1.4 Cycle Consistency

We propose the cycle consistency which is widely used in the image field for QG and QA. By using this method, as shown in Fig. 2, QG and QA modules can help generate a robust model that can match the question and answer, respectively. There are not many, but existing multi-hop QG models use graphs or templates (Pan et al., 2021; Su et al., 2020; Kumar et al., 2019). However, we introduce the cycle consistency loss to train a text-based model that can learn by an end-to-end manner. We define the cycle consistency loss to reduce the difference between the predicted value and the actual value. The overall learning flow of the model with cycle consistency is given in Fig. 2 (a). Fig. 2 (b) refers to the QGA-consistency which predicts a question through QG using a given context and an answer, and then predicts an answer through the QA again. Conversely, Fig. 2 (c) refers to QAG-consistency for finding an answer with QA using a given context and a question and then predicting the question with QG. In here, the process flow shown in Fig. 2 (b) is to predict a correct answer, and it is necessary to predict question well through QG to predict correct answer through QA as shown in Fig. 2 (c). Our model uses the cycle consistency property so that answers and questions are learnt better. We describe the overall flow of the CycleQAG framework in Algorithm 1. In algorithm 1, N is the number of samples of the dataset and we use a common early stopping approach for cycle training.

3.2 Prompt-based fine-tuning

We use the multi-hop dataset for fine-tuning a target task after training an intermediate task. Unlike general fine-tuning, the prompt-based fine-tuning adds an element called a prompt. Prompt shows that GPT-3 (Brown et al., 2020) achieves remarkable performance in a few-shot setting, and has been used in many recent studies (Shin et al., 2020; Lester et al., 2021; Gao et al., 2021). In particular, prompt-based fine-tuning (PFT) aims to investigate the knowledge gained from pre-training by reducing the distribution gap between pre-training and fine-tuning stages. Considering these points, we use PFT instead of fine-tuning to make most of the information obtained from the intermediate task. The detailed process of the PFT process is described in Algorithm 2.

A multi-hop QG aims to generate a question using several pieces of information related to a correct answer. For this, we construct a prompt by extracting words related to a correct answer from an intermediate task. This not only uses the types of questions and the correct answers, but also words

[†]Since the T5 is essentially an encoder-decoder model, it is assumed that the decoder knows the meaning of the entire input sentence while generating the first token prediction. This means that the output embedding of the first decoder can grasp the meaning of the sentence like the $[CLS]$ token of the BERT.

Algorithm 2 Procedure of Type-dependent prompt fine-tuning for QG

Input: Prompt (P) ; Context with answer (X)
Output: Multi-hop Question (Y)

1: Configure required prompt token per context
2: Merge prompt token and context with answer
3: Maximize the likelihood of multi-hop question Y.
4: **return** $Pr_\theta(Y|[P;X])$, while keeping the model parameters, θ, fixed.

related to the correct answers obtained from section 3.1.3 as a prompt. This enables providing more information when generating multi-hop questions. We show the results of questions obtained through PFT in appendix C.

3.3 Model Training

In this section, we describe in detail how the model trains an intermediate task. The total loss of the intermediate task consists of QA loss, QG loss, cycle consistency loss, and similarity loss. Therefore the loss is given by Eq. (1).

$$\mathcal{L}_{All} = \mathcal{L}_{QA} + \mathcal{L}_{QG} + \mathcal{L}_{cycle} + \mathcal{L}_{sim} \quad (1)$$

Eq. (2) defines the cycle consistency loss which includes the loss for QA model and loss for QG. We learn QA and QG cyclically with the cycle consistency loss, allowing QA to narrow the range of correct answers, and QG to express more questions in an enriched expression. The similarity loss L_{sim} determines whether the paraphrased questions are semantically close while learning the QG. L_{QA} uses CE loss to find a right answer for a given question. L_{QG} uses the CE loss and the NCE loss to generate a variety of questions that are similar to the original question.

$$\mathcal{L}_{cycle} = \frac{1}{2}[\underbrace{\mathcal{L}_{CE}}_{\text{QGA}} + \underbrace{\{\lambda_1 \mathcal{L}_{NCE} + (1-\lambda_1)\mathcal{L}_{CE}\}}_{\text{QAG}}] \quad (2)$$

The first term of L_{cycle} for learning the cycle-consistency is the loss obtained using the QGA-consistency, which is explained in section 3.1.4 and Fig. 2 (b). Here, the QGA learns to get closer to an original answer by generating a question using a answer and context and then generates a correct answer through the QA again. The remaining terms of the L_{cycle} describe the process of learning the QAG-consistency in Fig. 2(c). While the previous methods use the CE alone, we propose to use the NCE to train the QAG consistency to improve diversity. The QAG learns to get closer to an original question by performing QG through QA. In this part, we adjust the NCE and the CE with λ_1 so that the semantic and the lexical are properly balanced.

For generating questions with a similar meaning, to increase the diversity of questions and reduce the occurrence of most probability words, we use the NCE as shown in Eq. (3). In other words, it is intended to flatten the probability of occurrence of words, so that words can appear in various ways.

$$\mathcal{L}_{NCE} = \frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \quad (3)$$

However, if the model is trained only using Eq. (3), NCE may diverge (to -∞), so we adjust the Eq. (4) and hyperparameter λ_1 values such that the probability of the word appearing in the question is lowered only to a certain level. We heuristically adjust λ_1 as 0.2.

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \quad (4)$$

4 Experiments

In the following experiments, we evaluate multi-hop QG based on semantic similarity and lexical diversity. We also evaluate whether the intermediate task has an affect on QG module performance. The baseline model is initialized with a Google-T5 model from HuggingFace Transformer (Wolf et al., 2020), fine-tuned with 3 epochs, with batch size 8. The GPU used in the experiment is 4 Quadro RTX 8000.

4.1 Dataset

We evaluate our model with a focus on multi-hop QA, HotpotQA (Yang et al., 2018). HotpotQA is a multi-hop dataset that is more complex and requires reasoning than existing single-hop QA datasets.

As mentioned in RefNet (Nema et al., 2019), since the test set of HotpotQA is hidden, the validation set is used as the test set, and a part of the training set is used as a validation set. In the experiments, a dataset similar to the HotpotQA

	MODEL	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L
Baselines	B1. MQA-QG	36.01	25.79	21.88	17.83	26.89	39.95
	B2. BART	36.35	26.70	22.42	18.02	26.96	40.85
	B3. Google-T5	36.89	26.89	22.14	18.27	27.26	41.02
Proposed	P1. Type-dependent prompt CycleQAG	**38.28**	**29.77**	**24.32**	**20.51**	**29.01**	**44.10**
Ablation	A1. w/o type-dependent prompt	36.47	27.61	22.45	18.24	27.94	42.34
	A2. w/o Cycle (intermediate task)	36.96	27.88	22.91	18.98	27.70	41.90
	A3. w/o similarity, paraphrase	37.20	28.18	23.11	19.33	28.57	43.44

Table 1: Performance comparison with baseline and the ablation study. The best performance is bold.

dataset type is additionally used to improve the performance of multi-hop question generation. Representatively, SQuAD, a single-hop dataset, is used. Also, question paraphrasing is used to increase the robustness of the question.

Stanford Question Answering Dataset v1.1 (SQuAD) (Rajpurkar et al., 2016) is a machine reading comprehension dataset with over 100,000 questions created based on Wikipedia articles. Quora Questions Pairs (QQP) provides a label for detecting whether the intent is the same and whether the question text pairs correspond to semantically identical queries, with a focus on various issues related to Quora. We construct a QQP dataset with the same proportions as (Thakur et al., 2021). A detailed description of the dataset and data statistics are shown in appendix A.

4.2 Baselines

Since the multi-hop QG has not yet been explored much, there are few comparison models that can be compared with ours. We use a text-based multi-hop QG model and a model with excellent performance in QG research as our baseline models.

MQA-QG (Pan et al., 2021) generates a question according to a predefined reasoning graph according to the types of questions. In particular, they defined and used 11 templates for comparison type questions. We experiment with the same experiment settings as published in their paper. BART (Lewis et al., 2020) is a model that combines a Bidirectional Transformer and an Auto-Regressive Transformer, and is a pre-trained model using the denoising autoencoder method. In particular, it shows excellent performance in natural language generation. Google-T5 (Raffel et al., 2020) processes the NLP task using the text-to-text input and the output using C4 (Colossal Clean Crawled Corpus), a very large dataset and achieves the highest level in benchmarks such as SuperGLUE. We implement it using open code published by huggingface [‡] for BART and Google-T5.

MODEL	BERTSCORE
MQA-QG	91.88
BART-large	91.03
Google-T5	91.27
Type-dependent prompt CycleQAG (ours)	**93.87**
CycleQAG w/o type-dependent prompt	91.90
CycleQAG w/o Cycle	91.94
CycleQAG w/o similarity,paraphrase	92.93

Table 2: Performance of BERTSCORE. The best performance is bold.

4.3 Multi-hop QG Results and Analysis

Quantitative automatic evaluation and qualitative human evaluation are used to evaluate our proposed model. To this end, we describe in detail the automatic and human evaluation methods and discuss the results.

4.3.1 Automatic Evaluation Metrics

We perform automatic evaluation using n-gram and pre-trained language model based metrics.

N-gram based Metrics. BLEU (Papineni et al., 2002) score is a precision-based evaluation that computes the overlap of n-grams. METEOR (Lavie and Agarwal, 2007) is a relaxed F-measure-based evaluation method in which the unigrams of the hypothesis and the reference do not have an exact level of agreement, but they are synonymous. ROUGE-L (Lin, 2004) is a measure of the sequence of the longest common part between a pair of sentences (Sai et al., 2022). We use the nlg-eval [§] package released by (Sharma et al., 2017) to evaluate an n-gram-based metric.

Pre-trained Language Model based Metrics. BERTSCORE (Zhang* et al., 2020) is a method of evaluating NLG and computes a similarity score of each token of the candidate correct answer and the ground truth. Whereas existing evaluation methods evaluate based on exact match, BERTSCORE is

[‡] https://huggingface.co/

[§] https://github.com/Maluuba/nlg-eval

MODEL	Fluency	Relevance	Answerability	Complexity	Diversity
MQA-QG	2.45	2.38	2.42	2.35	2.35
BART	2.28	2.14	2.30	2.29	2.34
Google-T5	2.39	2.42	2.45	2.40	2.47
Type-dependent prompt CycleQAG (ours)	**2.56**	**2.62**	**2.59**	**2.53**	**2.66**

Table 3: Human Evaluation Results.

effective for paraphrase detection because it uses contextual embedding (Devlin et al., 2019). We download the package for bert-score from (Zhang* et al., 2020)[¶] and use it.

Results and Analysis. We compare the QG performance of the proposed type-dependent prompt CycleQAG model with baseline models and show the automatic metric results in Table 1. Our type-dependent prompt CycleQAG model outperforms all automatic evaluation metrics including ROUGE-L when compared to other models using the same data. Tables 2 indicates whether contextual meaning can be reflected, where BERTSCORE shows excellent performance. Also, it can be seen that they are semantically similar to the original question.

Ablation study. In order to understand the influence of the components of our proposed model, we conduct an ablation study with experimental data for type-dependent prompt CycleQAG. When we do not use the fine-tuning of the type-dependent prompt format that we suggest, it can be observed that the performance is lowered. It can also be observed that the presence or absence of additional information determines the performance improvement when performing a fine-tuning. The additional information referred to here is the types of questions and words related to the answer. We confirm through experiments that their role helps to improve overall performance. Also, fine-tuning the data without the cycle-consistency loss performed in the intermediate task stage, overall performance is degraded. This confirms that the intermediate task is helpful in the QG module when comparing the performance with B3 as shown in Table 1. When we train to generate questions in an intermediate task, we use paraphrase and similarity methods together to increase the lexical diversity of questions. If these methods are removed and tested, the overall performance is slightly degraded.

[¶] https://github.com/Tiiiger/bert_score

4.3.2 Human Evaluation

In this section, we discuss the human evaluation metric. We employ fluency, relevance, answerability, complexity and diversity. Human evaluation is an additional support method for the reliability and robustness of automated evaluation. Here, we use fluency, relevance, and answerability to measure the quality of whether our proposed question is relevant to a given context and answer. Multi-hop QG has high complexity because it requires reasoning, and it is necessary to measure the complexity of the generated question. In addition, we use diversity to evaluate cases in which vocabulary expressions are expressed in various ways, although the meaning of the question is the same. We randomly select 50 question-and-answer pairs from the test set from 20 annotators to obtain evaluations of our model and other baseline models. In human evaluation, we perform the evaluations in a blind format. The range of scores used for evaluation is set to 1-3, and the higher the score, the better the evaluation. The results are shown in Table 3. Overall we consistently get better performance than the conventional models like the BART and Google-T5. We obtain significantly better results than other reference models, especially in terms of diversity and complexity.

5 Conclusion

In this work, we propose type-dependent prompt CycleQAG with cycle consistency. Since multi-hop QG needs to know more diverse information because it needs to gather more scattered pieces of information for generating a question, we introduce the NCE for the first time in the QG task. Also, we demonstrate that the intermediate task is effective in the QG task. Furthermore, we show a significant performance improvement by using prompt-style fine-tuning to make the most of the information obtained from the intermediate task. The experiments show that the proposed model outperforms in all automatic evaluations comparing with the existing text-based multi-hop model and several QG models. Although we use only multi-hop, single-hop-based

datasets, experiments can be performed without additional datasets later using the type-dependent prompt CycleQAG method. In other words, it is possible to learn QA and QG models using unsupervised learning. In the future, we would like to investigate a model that generates questions and answers by itself enough to imitate humans from knowledge through self-cyclic learning that is less influenced by data.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-02068, Artificial Intelligence Innovation Hub), and Center for Applied Research in Artificial Intelligence(CARAI) grant funded by Defense Acquisition Program Administration(DAPA) and Agency for Defense Development(ADD) (UD190031RD).

References

Richard W Brislin. 1970. Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, 1(3):185–216.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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.

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

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

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Deepak Gupta, Hardik Chauhan, Ravi Tej Akella, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Reinforced multi-task approach for multi-hop question generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2760–2775, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tianbo Ji, Chenyang Lyu, Zhichao Cao, and Peng Cheng. 2021. Multi-hop question generation using hierarchical encoding-decoding and context switch mechanism. *Entropy*, 23(11).

Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. 2019. Difficulty-controllable multi-hop question generation from knowledge graphs. In *International Semantic Web Conference*, pages 382–398. Springer.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Petr Marek, Vishal Ishwar Naik, Anuj Goyal, and Vincent Auvray. 2021. OodGAN: Generative adversarial network for out-of-domain data generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 238–245, Online. Association for Computational Linguistics.

Preksha Nema, Akash Kumar Mohankumar, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2019. Let's ask again: Refine network for automatic question generation. 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 3314–3323, Hong Kong, China. Association for Computational Linguistics.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Unsupervised multi-hop question answering by question generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5866–5880, Online. Association for Computational Linguistics.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

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

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. Mathbert: A pre-trained model for mathematical formula understanding. *arXiv preprint arXiv:2105.00377*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603.

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.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020. Multi-hop question generation with graph convolutional network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4636–4647, Online. Association for Computational Linguistics.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yibo Sun, Duyu Tang, Nan Duan, Tao Qin, Shujie Liu, Zhao Yan, Ming Zhou, Yuanhua Lv, Wenpeng Yin, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020. Joint learning of question answering and question generation. *IEEE Transactions on Knowledge and Data Engineering*, 32(5):971–982.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027*.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

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

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Jianxing Yu, Xiaojun Quan, Qinliang Su, and Jian Yin. 2020. Generating multi-hop reasoning questions to improve machine reading comprehension. In *WWW*, pages 281–291.

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

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

Appendix

A Data statistics

Dataset	Data type	Train set	Validation set	Test set
HotpotQA	All	90.4k	7.4k	7.4k
	Bridge type	58.5k	5.9k	(invisible)
	Comparison type	17.4k	1.5k	(invisible)
	Single-hop	14.5k	-	-
SQuAD	Single-hop	87.5k	10.5k	-
QQP	Paraphrase	254k	10k	10k

Table A1: Statistics of Datasets. Number of data instances in the train, validation and test set of HotpotQA, SQuAD and Quora Question Pairs (QQP).

Table. A1 is a statistic of the dataset used in the experiment. HotpotQA provides two dataset versions, a distractor setting and a full wiki setting. In this paper, we conduct all experiments on a distractor setting with 2 gold paragraphs and 8 distractor paragraphs. We use training and validation sets provided by HotpotQA to train and evaluate the model. The SQuAD dataset is similar to HoptotQA dataset, the answer to each question can be found in the form of a text span in the paragraph, and it consists of data that can answer diverse types of questions. The QQP dataset detects whether the intent of two given pairs of sentences is the same, and provides a label on whether the two sentences are semantically similar. The QQP dataset consists of more than 400,000 lines of potential question duplicate pairs, with a binary value indicating whether that row contains duplicate pairs. We obtain HotpotQA, SQuAD data through the datasets package provided by Huggingface and use it for experiments, which can be downloaded from https://huggingface.co/datasets. QQP dataset can be downloaded from https://www.kaggle.com/c/quora-question-pairs.

B Error Analysis

We perform error analysis to analyze the experimental results. In Table A2, Case1 has the same meaning, but only the expression of the lexicon is different, which is a well-predicted case. However, in Case 2, an important entity is omitted, so the meaning of the question is completely changed, and the gist of the question can not be grasped. In this question, an important "traditional sport considered a national sport" is predicted with the word "traditional", so it is a erroneously generated question because an entity related to the sport is omitted. This is not a well-formed question because we cannot conversely answer the question properly. Case 3 is the case of making a semantic error. The ground truth of Case 3 is to ask whether Mozart wrote Pomone or Idomeneo, but the meaning of the predicted question is a semantically different question because it asks which of Pomone or Idomendo premiered first. Case 4 is the case of generating a completely wrong question, where the intent of the question is completely changed by incorrectly predicting the meaning of "and" as "unlike".

Types	Example
Case1	(GT.) Who was the director of the James Bond movie in which Anatole Taubman appeared as a henchman? (Pred.) Who directed the James Bond movie in which Anatole Taubman appeared as a henchman?
Case2	(GT.) Which old, traditional sport is considered to be the national sport by some Swiss? (Pred.) What is an old tradition from the rural central cantons?
Case3	(GT.) Of Pomone and Idomeneo which one was written by Amadeus Mozart? (Pred.) Which opera premiered first, Pomone or Idomeneo?
Case4	(GT.) Unlike Xuzhou, where is Rugao under the adminstration of? (Pred.) Rugao and Xuzhou, Jiangsu are both county-level cities under the administration of who?

Table A2: Error Analysis. GT is ground truth, and Pred is prediction example.

C Experimental result examples

In this section, an example of generating a multi-hop question using the type-dependent prompt CycleQAG method presented by us is shown in detail. In Table A3 and A4, we mark the correct answer we want to obtain in red text and the words related to the correct answer obtained through section 3.1.3 in blue text. If the correct answer and the word related to the correct answer overlap, it is indicated in cyan text. In particular, we can confirm that the meaning of the original question and the generated question did not change, but a vocabulary with a similar meaning was used, making the question richer. In addition, it can be seen through the example of the generated question that it has a considerable influence when generating a question by using the question type, answer, and answer-related words as a prompt. More specifically, in prompt based fine-tuning, we set the input as question type: *type of question*, answer-related words: *combination of words related to the correct answer*, context: *context with answer* and set the output to *multi-hop question*.

Data fields	Example
Answer	Jacksonville station
Generated answer related words	Silver Meteor, Jacksonville station
Context	The Silver Meteor is a passenger train operated by Amtrak between New York City and Miami, Florida. The first diesel-powered streamliner between New York and Florida, since being introduced by the Seaboard Air Line Railroad (SAL) in 1939, it remains in operation now. The train is part of Amtrakś "Silver Service" along with the "Silver Star", another former SAL streamliner. Jacksonville station is an Amtrak train station in Jacksonville, Florida, United States. It serves the "Silver Meteor" and "Silver Star" trains as well as the Thruway Motorcoach to Lakeland. The station lies next door to a freight facility with its own platform and is also just east of Norfolk Southern's Simpson Yard.
Original question	Where does the train that runs from NYC and Miami station at Florida?
Generated question	Which Amtrak station serves the passenger train operated by Amtrak between New York City and Miami, Florida?
Answer	Julianne Moore
Generated answer related words	Emanuelle Goes to Dinosaur Land, Julianne Moore
Context	"Emanuelle Goes to Dinosaur Land" is the of the fourth season of the American television comedy series "30 Rock", and the 79th overall episode of the series. It was written by supervising producer Matt Hubbard and directed by Beth McCarthy-Miller. The episode originally aired on the National Broadcasting Company (NBC) network in the United States on May 13, 2010. Guest stars in this episode include John Anderson, Elizabeth Banks, Jon Hamm, Kristin McGee, Julianne Moore, Michael Sheen, Jason Sudeikis, and Dean Winters. Julianne Moore (born Julie Anne Smith; December 3, 1960) is an American actress, prolific in films since the early 1990s. She is particularly known for her portrayals of emotionally troubled women in both independent and Hollywood films, and has received many accolades, including the 2014 Academy Award for Best Actress.
Original question	What 2014 Academy Award winner guest starred in "Emanulle Goes to Dinosaur Land?
Generated question	Which guest star in "Emanuelle Goes to Dinosaur Land" won the 2014 Academy Award for Best Actress?
Answer	Lantern Waste
Generated answer related words	Lantern Waste, Tumnus
Context	Lantern Waste is a fictional place in "The Chronicles of Narnia" series by C. S. Lewis. It is a wood and is notable as the place where Lucy Pevensie and Mr. Tumnus meet, which is the first scene of Narnia described in the books. The lamppost in the wood is an iconic image of Narnia, and the question of its origin is what convinced Lewis to write more than one book on Narnia. One of King Edmund's titles is "Duke of Lantern Waste". Tumnus is a fictional character in C. S. Lewis' series "The Chronicles of Narnia". He is featured prominently in "The Lion, the Witch and the Wardrobe" and also appears in "The Horse and His Boy" and "The Last Battle". He is close friends with Lucy Pevensie and is the first creature she meets in Narnia, as well as the first Narnian to be introduced in the series. Lewis said that the first Narnia story, "The Lion, the Witch and the Wardrobe", all came to him from a single picture he had in his head of a faun carrying an umbrella and parcels through a snowy wood. In that way, Tumnus was the initial inspiration for the entire Narnia series.
Original question	What is the name of the place in The Chronicles of Narnia where Lucy Pevensie and Mr. Tumnus meet?
Generated question	What is the name of the fictional place where Lucy Pevensie and Mr. Tumnus meet?

Table A3: Example of generated bridge type multi-hop question.

Data fields	Example
Answer	Emory University
Generated answer related words	Emory University, Vanderbilt University
Context	Emory University is a private research university in metropolitan Atlanta, located in the Druid Hills section of DeKalb County, Georgia, United States. The university was founded as Emory College in 1836 in Oxford, Georgia by the Methodist Episcopal Church and was named in honor of Methodist bishop John Emory. In 1915, the college relocated to metropolitan Atlanta and was rechartered as Emory University. The university is the second-oldest private institution of higher education in Georgia and among the fifty oldest private universities in the United States. Emory is frequently cited as one of the world's leading research universities and one of the top institutions in the United States. Vanderbilt University (also known informally as Vandy) is a private research university located in Nashville, Tennessee. Founded in 1873, it was named in honor of shipping and rail magnate Cornelius Vanderbilt, who provided the school its initial $1 million endowment despite having never been to the South. Vanderbilt hoped that his gift and the greater work of the university would help to heal the sectional wounds inflicted by the Civil War.
Original question	Was Vanderbilt University or Emory University founded first?
Generated question	Which university is older, Vanderbilt University or Emory University?
Answer	Battle of Guam
Generated answer related words	Battle of Manila , Battle of Guam
Context	The Battle of Manila (February 3, 1945 – March 3, 1945) was a major battle of the Philippine campaign of 1944-45, during the Second World War. It was fought by American and Filipino forces against Japanese troops in Manila, the capital city of the Philippines. The month-long battle, which resulted in the death of over 100,000 civilians and the complete devastation of the city, was the scene of the worst urban fighting in the Pacific theater. Japanese forces committed mass murder against Filipino civilians during the battle. Along with massive loss of life, the battle also destroyed architectural and cultural heritage dating back to the city's foundation. The battle ended the almost three years of Japanese military occupation in the Philippines (1942–1945). The city's capture was marked as General Douglas MacArthur's key to victory in the campaign of reconquest.The Second Battle of Guam (21 July – 10 August 1944) was the American recapture of the Japanese-held island of Guam, a U.S. territory in the Mariana Islands captured by the Japanese from the U.S. in the 1941 First Battle of Guam during the Pacific campaign of World War II.
Original question	Which battle occurred first, the Battle of Manila or the Battle of Guam ?
Generated question	Which battle took place first, Battle of Guam or Battle of Manila?
Answer	Dracula
Generated answer related words	Dracula, Pistacia
Context	The orchid genus Dracula, abbreviated as Drac in horticultural trade, consists of 118 species native to Mexico, Central America, Colombia, Ecuador and Peru. The name "Dracula" literally means "little dragon", an allusion to the mythical Count Dracula, a lead character in numerous vampire novels and films. The name was applied to the orchid because of the blood-red color of several of the species, the strange aspect of the long spurs of the sepals. Pistacia is a genus of flowering plants in the cashew family, Anacardiaceae. It contains 10 to 20 species that are native to Africa and Eurasia from the Canary Islands, all of Africa, and southern Europe, warm and semidesert areas across Asia, and North America from Mexico to warm and semidesert United States, such as Texas or California.
Original question	Whiich genus has more species, Dracula or Pistacia?
Generated question	Which genus contains more species, Pistacia or Dracula?

Table A4: Example of generated comparison type multi-hop question.