Encoding Paraphrase Information Generated by ChatGPT and Introducing Interrogative Word Control Mechanism for Question Generation

Fangfang Wang
Dongguan University of Technology
Dongguan, China
2112115044@dgut.edu.cn

Jian Zhang*
Dongguan University of Technology
Dongguan, China
zjian03@gmail.com

Huaqiang Yuan
Dongguan University of Technology
Dongguan, China
yuanhq@dgut.edu.cn

***Abstract*—Answer-aware question generation means generating a relevant question based on a passage of text and an answer, where the answer is a subsequence within the text. In fact, different expressions of the same thing can affect human understanding, and the same holds true for models. Therefore, in this paper, we use paraphrase information to mimic different perspectives on the same fact and propose a new question generation model. Specially, we utilize ChatGPT to obtain paraphrase of the context. What sets our approach apart from previous work is that we directly model both the context and its paraphrase information using a shared encoder, enabling the model to capture semantic information from different perspectives and fully understand the context. Furthermore, we propose a new self-guided conditional layer normalization module (SCLN) that introduces interrogative words as conditional information into the encoder for the first time. This module adaptively maps interrogative word information to context information, enhancing the model's understanding of the context and controlling the interrogative word that generate question, thereby enhancing the answerability of the question. We conduct experiments on the SQuAD dataset and the NewsQA dataset. The experimental results show that our proposed model achieves significant performance improvements compared to several baseline models. Further human evaluation demonstrates that the questions generated by our model can comprehensively capture contextual information to improve question answerability.**

***Index Terms*—Question generation, text paraphrase, interrogative words**

I. Introduction

Question generation is an important task in NLP that has received widespread attention in recent years. Its goal is to generate related questions based on input context. Question generation can be applied in various scenarios, such as constructing question-answering datasets to provide training data for question-answering or reading comprehension. It can also be used in medical question-answering systems to assist doctors with their diagnoses, or in educational systems to help teachers evaluate their students' grasp of knowledge. This work focuses on sentence-level text question generation.

Significant progress has been made in question generation. Previous work has used recurrent neural networks to obtain sequential contextual information, graph neural networks and other structures to obtain rich structural information, or knowledge graphs to provide external knowledge for contexts, with the aim of obtaining sufficient contextual information. However, these methods all ignore the sensitivity of the model to different expressions of the same meaning, and thus different modeling results may arise for contexts that have the same meaning but are described differently. To address this issue, Jia et al. [1] proposed utilizing the diversity of expressions obtained through paraphrases to aid question generation, enabling the model to learn more question patterns. However, the aforementioned methods indirectly utilize paraphrase information through multi-task learning, thereby failing to fully utilize it. In real life, humans always re-describe the same fact based on their own understanding. Therefore, we believe that different representations of the same context are parallel relations, and the context and paraphrase information should be modeled at the same time, so that the model can obtain the deep semantics of the context from different perspectives.

In addition, previous work has shown that interrogative word can guide question generation. Huang et al. [2] heuristically extracted interrogative word from the answer. Zhou et al. [3] used a softmax on the hidden state at the position of the answer to predict the interrogative word. However, all these works introduce the interrogative word at the decoding stage to initialize the decoder to guide question generation. After analyzing the experimental results, we found that the above methods cannot fully control the way the model asks questions. Therefore, we propose to use interrogative word information as conditional information to control question generation and facilitate question generation.

Inspired by the above, this paper proposes a question generation model that utilizes paraphrase information. Specifically, we first acquire the paraphrase information of the context and expand the dataset accordingly. Then, we employ a shared encoder to explicitly model both the context and its paraphrase information. This allows the model to capture different representations of the same fact, thereby obtaining context-level semantic information after feature fusion. Finally, the encoded information is passed to the decoding stage for

This work was supported by the Natural Science Foundation of China (Grant No.61972090). *Corresponding author: Jian Zhang.

question generation. Additionally, we introduce the SCLN module, which incorporates interrogative word information into the encoding stage. This enables the model to capture the implicit relationship between interrogative word and the context, mapping the interrogative word features into the context information to control the question generation process. Furthermore, we also introduce interrogative word information in the decoding stage to further guide question generation and improve the quality of generated questions.

In summary, the contributions of this paper are:

- We propose to simultaneously model both the context and its paraphrase information, enabling the model to capture the semantic information of the text from multiple perspectives and thus enhancing the contextual semantic information.
- We propose SCLN module, which introduces interrogative word features in the vector space, uses interrogative word as conditional information to control question generation, and improves the accuracy of question types and the answerability of generated questions.
- Experimental results on the SQuAD dataset and NewsQA dataset demonstrate that our model outperforms the baselines, validating the effectiveness of the proposed methods. Human evaluation indicates that the questions generated by our model are more answerable and exhibit better grammaticality.

II. Related Work

A. Question Generation

In recent years, the attention of researchers on question generation has grown significantly due to the rapid development of question-answering systems and machine reading comprehension. Previous approaches to question generation can be classified into two categories: rule-based and neural network-based.

Rule-based question generation relies on the utilization of manually designed templates to transform declarative sentences into interrogative form. However, this rule-based method presents challenges in terms of requiring a significant amount of human annotation and limitations in its extensibility. Therefore, while the rule-based method is somewhat effective, the creation of templates presents a challenging task.

Question generation has made significant progress thanks to the advancements in neural networks and machine translation. The task of question generation was first defined as a sequence-to-sequence task by [4], who also introduced neural networks to this domain for the first time. The majority of existing research methods in question generation are built on the sequence-to-sequence approach. To enhance the model's ability to grasp contextual information, [5] supplemented the encoder with diverse language lexical features and answer information, enabling the model to acquire more comprehensive feature information for generating questions. However, previous works have ignored the rich structural information implicit in the text. Therefore, [6] introduced the Graph2Seq architecture to capture structural information in the text. Additionally, they proposed a deep alignment network that incorporates answer information at various granularities. Huang et al. [2] utilized GCN and LSTM to capture both textual structural information and sequential information, and use heuristic rules to predict interrogative word, and use interrogative word to guide question generation.

The advancement of large-scale pre-trained language models has led to the impressive performance of certain models in question generation task, including UniLM [7] and ProphetNet [8]. Fei et al. [9] used BERT to obtain contextual semantics and proposed a decoding method based on an iterative graph network to model the structural information of previously generated text, capturing information from previously generated words to assist with subsequent text generation. Duong et al. [10] used pre-trained model to jointly learn question generation and sentence selection tasks. In this paper, we utilize the pre-trained model T5 [11] and BART [12] as the base model, leveraging its powerful modeling ability to accomplish the question generation task.

B. Paraphrase generation

Paraphrases refer to texts that convey the same meaning using different words or sentence structures. Paraphrase generation is one of the widely studied tasks in natural language processing. Paraphrase knowledge has been used to enhance various NLP tasks. For example, [13] used paraphrasing techniques to expand the training dataset for the maritime field. Thompson et al. [14] employed paraphrase information to aid in assessing the quality of machine translation, bringing the model's output closer to human references.

Previous studies have integrated paraphrase information into question generation, such as [15], who extracted triplets from text paraphrases using OpenIE. They used the subject or object as an answer to generate questions and acquire QA pairs. However, this work primarily focused on reducing lexical overlap through paraphrases to improve question quality. Additionally, [1] adopted a multi-task learning approach to simultaneously generate paraphrases and questions, aiming to enhance expression diversity by paraphrase information. In contrast, we directly model paraphrase information, enhancing the model's contextual understanding by integrating human understanding from different perspectives on the same fact, thus facilitating question generation.

III. Method

A. problem definition

The question generation task is defined as follows: Given a sentence $S = \{s_1, s_2, \ldots, s_m\}$ and an answer $A = \{a_1, a_2, \ldots, a_n\}$, where the answer A is a subsequence of the sentence S, the objective of question generation is to generate a question Q related to S, with A as its answer.

B. Paraphrase expansion

Paraphrase generation refers to generate an output sentence that has the same meaning as the input sentence but with

different sentence structure and keywords. By leveraging this characteristic of paraphrase, we can simulate different perspectives on a fact and provide additional explanations for the source context. Note that any advanced method can be used to generate paraphrases. In this paper, we leverage the powerful ability of ChatGPT to obtain paraphrases of the source text and then use them to expand the context for question generation. The prompt for the paraphrase generation task are constructed, as illustrated at the top of Fig. 1. We build prompt to get contextual paraphrase. The term "**S**" corresponds to the context. We fill it in using the formulated prompt and then use it as input. The original dataset consists of $< S, A, Q >$ triplets, and we expand it to include the paraphrased context, resulting in $< S, A, Q, S_P >$ quadruplets. Here, S represents the context, A represents the answer, Q represents the corresponding question, and S_P represents the paraphrased text of the context.

C. Question Generation Model

The architecture of our model follows an encoder-decoder structure, illustrated in Fig. 1. We first obtain the paraphrase corresponding to the given context, expanding the dataset. Then, we concatenate the source sentence and its corresponding paraphrase with the answer using special tokens (<context>, <answer>), forming separate inputs. In order to acquire more profound semantic information from the context, we combine the semantic features of X and X_P to generate an amplified semantic information feature denoted as H.

$$X = (< context >, S, < answer >, A) \tag{1}$$

$$X_P = (< context >, S_P, < answer >, A) \tag{2}$$

$$H_X = Encoder(Embed(X)) \tag{3}$$

$$H_{XP} = Encoder(Embed(X_P)) \tag{4}$$

$$H = H_X + \lambda H_{XP} \tag{5}$$

Where λ is a hyperparameter, $Embed(X), Embed(X_P)$ respectively represents embeddings of X and X_P. In this paper, we assume that the paraphrase information is reliable and set $\lambda = 1$.

Once the hidden states H of the encoder are acquired, the decoder takes the fused feature representation generated by the encoder as input and performs autoregressive generation to generate the corresponding question Q.

SCLN: AdaIN is a widely employed technique in style transfer to incorporate diverse image styles. Karras et al. [16] transformed latent variables into intermediate variables through a mapping network and utilized AdaIN to introduce image features. Building upon this concept, we introduce SCLN module (depicted on the right side of Fig. 2), which integrates interrogative word as conditional information into the model for controlling question generation.

Initially, to gather interrogative words, we categorize them into seven common types, labeling any types beyond these as "other". Subsequently, we derive interrogative words by obtaining Named Entity Recognition (NER) features from the answer through heuristic rules. These interrogative words are then incorporated into the question generation model. More precisely, we substituted the layer normalization layer within the transformer block and performed a mapping between the interrogative word feature and the context representation. This allows the model to dynamically learn the connection between the interrogative word and the context, thereby exerting control over question generation.

$$SCLN = \gamma * LN(Embed(X)) + \beta \tag{6}$$

$$\gamma, \beta = MLP(E_{style}) \tag{7}$$

LN represents Layer Normalization operation. E_{style} represents embedding of the interrogative words. γ and β are derived by applying an affine transformation to the E_{style}.

IV. Experiments

A. Datasets

SQuAD [17]: The SQuAD dataset is the most widely used question-answering dataset and is also a widely used and authoritative dataset for question generation. Consequently, we conducted experiments on this dataset and employed Du's [4] method to split it, focusing on generating questions at the sentence level. Ultimately, we obtained 76,718 training examples, 7,821 validation examples and 7,837 test examples.

NewsQA [18]: It was extracted from news articles of the American Cable News Network. We processed the dataset by segmenting the paragraphs to make the data suitable for sentence-level question generation. Ultimately, the dataset consisted of 73,960 training examples, 4,228 validation examples, and 4,131 test examples.

B. Baselines

UniLM [7]: A pre-trained language model based on BERT.

BART [12]: A sequence-to-sequence transformer model used for pre-training by leveraging a denoising autoencoder.

ProphetNet [8]: A transformer-based seq2seq architecture designed for various natural language generation tasks.

T5 [11]: A unified model framework utilizing a prefix format to adapt to diverse downstream tasks, converting all tasks into seq2seq tasks.

SP-T5: Our method is implemented based on the T5 model.

SP-BART: Our method is implemented on the BART model.

C. Automatic Evaluation Metrics

BLEU [19]: Originally employed for machine translation, the BLEU metric is also widely used in question generation tasks. It gauges the similarity between the generated question and a reference question by comparing their n-gram overlap. A higher score indicates better performance.

ROUGE-L [20]: ROUGE-L assesses similarity between the generated and target questions using the longest common subsequence. A higher score indicates better alignment.

METEOR [21]: METEOR considers various factors, including precision, recall, semantic matching, and word order accuracy. Higher scores denote better results.

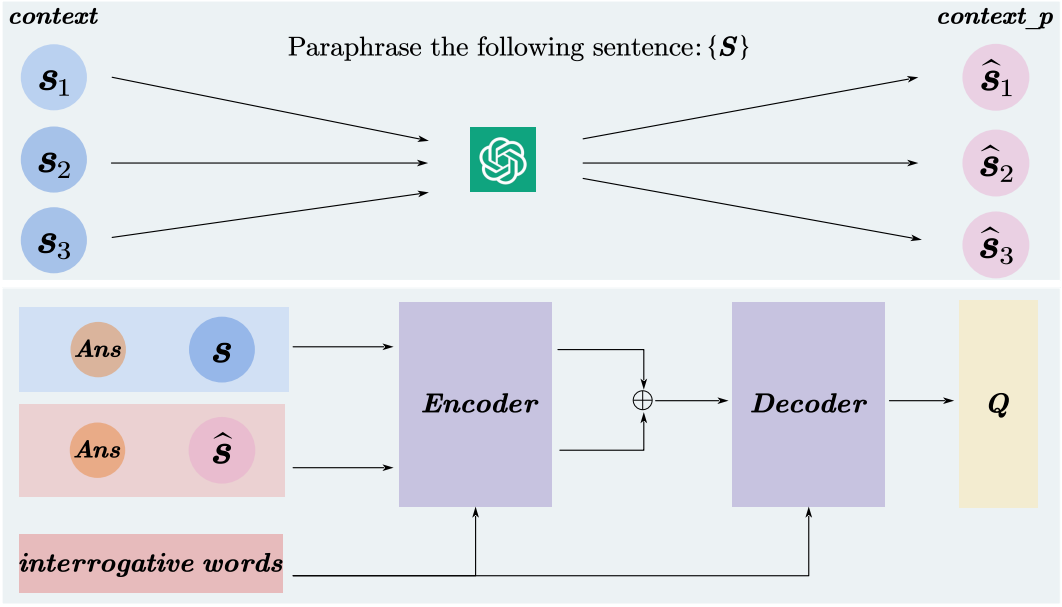

Fig. 1. Illustration of our proposed model.

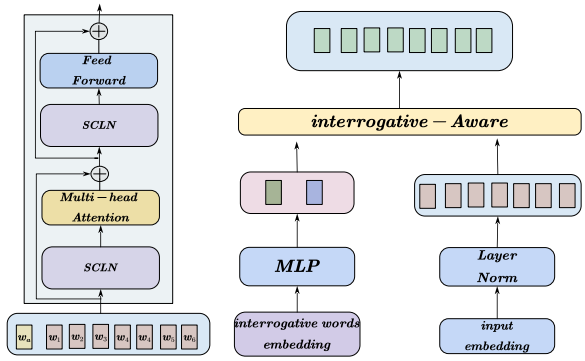

Fig. 2. On the left is the structure diagram of the transformer block; on the right is the structure diagram of the SCLN model.

D. Implementation Details

We conducted our experiments using PyTorch. All models were trained on the GeForce RTX 3090 card. During training, we employed the Adam optimizer. The initial learning rate was set to 3e-5. The batch size was configured as 8. For testing, we chose a beam size of 3. Due to equipment limitations, all models used in this paper are based on the base versions.

V. Results

A. Main Results

Table I presents the experimental results of SP-T5 on SQuAD and NewsQA, respectively. From these comparative results, it's evident that our model outperforms the baseline model in both SQuAD and NewsQA datasets. For instance, when comparing our model with SP-T5 and T5 on SQuAD, we observe an improvement of 1.99% in the BLEU metric and 3.98% in the ROUGE-L metric. This improvement is due to two factors. On the one hand, the paraphrase information and context are modeled at the same time, so that the model can model the context from different perspectives and pay attention to more comprehensive context information; on the other hand, introducing interrogative words into the encoder can make the model adaptively map interrogative word features into the context. Additionally, the attention mechanism within the transformer architecture assists the model in capturing the underlying relationships between interrogative words, context, and answers. This allows the model to ask questions in a clear way and improving the answerability of the question.

We further compare our model with state-of-the-art pre-trained language models for the task of question generation. Despite the gap between SP-T5 and the SOTA models, our model still achieves competitive results compared to UniLM. In general, our model achieves comparable results at lower cost without requiring large amounts of external data and computing resources.

TABLE I
THE RESULTS OF THE SP-T5 MODEL ON THE SQUAD DATASET AND NEWSQA DATASET.(WITH * INDICATES THE RESULTS WE EXTRACTED DIRECTLY FROM THE OFFICIAL PAPER. "B" STANDS FOR "BLEU"; "RL" STANDS FOR "ROUGE-L"; "M" STANDS FOR "METEOR")

Models	SQuAD			NewsQA		
	B↑	RL↑	M↑	B↑	RL↑	M↑
UniLM	22.12*	51.07*	25.06*	-	-	-
prophetNet	23.91*	52.26*	26.60*	-	-	-
T5	19.95	45.16	23.29	13.30	36.77	18.68
SP-T5	**21.94**	**49.14**	**24.59**	**19.84**	**41.14**	**20.13**

Furthermore, we applied our method to the BART model for experimentation, and the results are presented in Table II, illustrating the efficacy of our approach. Specifically, when comparing SP-BART with BART, an increase of 2.59% in ROUGE-L is observed on the SQuAD dataset, along with a 4.58% ROUGE-L improvement on the NewsQA dataset.

TABLE II
THE RESULTS OF THE SP-BART MODEL ON THE SQUAD DATASET AND NEWSQA DATASET.("B" STANDS FOR "BLEU"; "RL" STANDS FOR "ROUGE-L"; "M" STANDS FOR "METEOR")

Models	SQuAD			NewsQA		
	B↑	RL↑	M↑	B↑	RL↑	M↑
BART	19.25	43.89	22.87	11.51	35.48	17.56
SP-BART	**19.99**	**46.48**	**23.43**	**13.66**	**40.06**	**18.87**

B. Human Evaluation results

Furthermore, we conducted a manual evaluation to assess the quality of the generated questions. For this purpose, we randomly selected 100 test samples and invited three volunteers. These volunteers are second-year graduate students with good English reading skills, and capable of effectively evaluating question quality. Clear instructions were provided to the volunteers, outlining three key criteria for assessment: grammaticality, relevance, and answerability. They were instructed to assign scores ranging from 1 to 5, with higher scores representing better quality.

As presented in Table III, our model achieves the highest score for answerability, highlighting the effective control our proposed SCLN lends to question generation, consequently enhancing answerability. Additionally, our model outperforms other baselines in terms of grammaticality. This observation suggests that the incorporation of paraphrase information aids the model in capturing more profound semantic nuances.

TABLE III
HUMAN EVALUATION RESULTS ON SQUAD(**GRAMMATICALITY:** WHICH ASSESSES WHETHER THE QUESTION ADHERES TO GRAMMAR RULES AND FLUENCY. **RELEVANCY:** WHICH DETERMINES THE EXTENT TO WHICH THE QUESTION IS RELATED TO THE INPUT CONTEXT. **ANSWERABILITY:** WHICH EVALUATES THE FEASIBILITY OF ANSWERING THE QUESTION USING THE GIVEN ANSWER.)

Models	Grammaticality	Relevancy	Answerability
T5	4.37	4.19	3.98
SP_T5	4.52	4.48	4.23

C. Ablation results

We performed experiments to assess the effectiveness of each component in our model by separately employing paraphrase information and SCLN module. The results, presented in Table IV, clearly indicate that the removal of any component led to a substantial decrease in performance. Especially for the SCLN module, when the SCLN module are removed, the performance of BLEU drops by 1.43%, which fully demonstrates that the SCLN module can control the model to ask questions in the correct way and improve the quality of generated questions. Moreover, the removal of paraphrase information caused a decrease of 0.27% in the BLEU score, emphasizing the model's ability to acquire complementary contextual semantic information from paraphrase.

For the interrogative word, we conducted another experiment where we directly concatenated the interrogative word to the input sequence. This allowed the model to directly acquire the interrogative word and guide the question generation. The results showed that although this method also achieved good outcomes, it did not perform as well as the SCLN. This indicates that interrogative words are crucial for question generation, and our proposed SCLN module can better control the model's capture of relevant information, leading to the generation of high-quality questions.

TABLE IV
ABLATION STUDIES ON SQUAD. ("-P" DENOTES THE REMOVAL OF PARAPHRASE INFORMATION; "-SCLN" INDICATES THE EXCLUSION OF THE SCLN; "SP-T5(CONCAT)" SIGNIFIES DIRECTLY CONNECTING THE INTERROGATIVE WORD WITH THE CONTEXT AS INPUT)

Models	BLUE	R1	R2	RL	METEOR
SP-T5	21.94	51.83	29.13	49.14	24.59
-P	21.67	51.52	28.82	48.88	24.60
-SCLN	20.51	48.60	27.03	45.42	23.50
SP-T5(concat)	21.04	51.02	28.14	48.10	24.52

In addition, we conducted an experiment on the position of the interrogative word, and the results are shown in Table V. It was observed that the interrogative word played a role in both the Encoder and the Decoder. This confirms our conjecture that the introduction of interrogative word in the encoder can map interrogative word features to contextual information, capture the hidden relationship between interrogative word and context, and help capture key information from the context; introducing interrogative word in decoder, question types can be emphasized to guide question generation and improve question answerability.

TABLE V
ABLATION EXPERIMENT OF INTERROGATIVE WORD INTRODUCTION POSITION ("- ENCODER" MEANS TO REPLACE THE SCLN ON THE ENCODER SIDE WITH A REGULAR LAYERNORM;"- DECODER" MEANS TO REPLACE THE SCLN ON THE DECODER SIDE WITH A REGULAR LAYERNORM)

Models	BLUE	R1	R2	RL	METEOR
T5	19.95	48.33	26.56	45.16	23.29
SP-T5	21.94	51.83	29.13	49.14	24.59
- encoder	20.72	50.71	27.63	47.68	24.08
- decoder	21.68	51.62	28.89	48.91	24.49

D. Case Study

We present several examples of generated questions in Table VI. In Example 1, it can be observed that T5 model

demonstrates a lack of deep comprehension of the context, resulting in errors within the generated question such as inaccurate information (the text mentions "a delay of 80 seconds", while the generated question states "ten minutes"). In contrast, our model is able to capture the correct relevant contextual information, ask questions with correct interrogative word, and exhibit better grammar and answerability.

For Example 2, although the question generated by the T5 model exhibit grammatical correctness, they fail to incorporate specific background information (such as the mention of "as of 2012" in the text), thereby weakening the relevance of the generated questions. In contrast, our model is capable of fully utilizing contextual information to generate more specific and higher-quality questions.

TABLE VI
EXAMPLES OF GENERATED QUESTIONS. THE TEXT IN RED IS THE ANSWER, AND "GOLDQ" REPRESENTS THE REFERENCE QUESTION.

context	it has been claimed that the transmission of the first episode was delayed by ten minutes due to extended news coverage of the assassination of us president john f. kennedy the previous day; whereas in fact it went out after a delay of eighty seconds.
GoldQ	what major event u.s. occurred that made the bbc delay the broadcast?
T5	why was the first episode of american idol delayed ten minutes?
SP-T5	what event caused the delay in the first episode of american idol?
context	as of 2012, quality private schools in the united states charged substantial tuition, close to $40,000 annually for day schools in new york city, and nearly $50,000 for boarding schools.
GoldQ	about how much did a new york city day school cost annually in 2012?
T5	how much did day schools in new york city charge per year?
SP-T5	as of 2012, what was the annual tuition for day schools in nyc?

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a QG model that leverages paraphrase information to enhance the model's understanding of context. We leverage paraphrase information to mimic human understanding of different perspectives on the same fact and obtain paraphrases via ChatGPT. We then explicitly model the context and its paraphrase, enabling the model to capture deep semantic information from the context. Moreover, in order to improve the answerability of questions, we propose the SCLN module to control the interrogative words of the questions. Experimental results on the SQuAD dataset and the NewsQA dataset also demonstrated the effectiveness of our approach. Given the diverse nature of paraphrase information, in the future, we aim to further leverage paraphrase information to enhance the diversity of generated questions.

REFERENCES

[1] X. Jia, W. Zhou, X. Sun, and Y. Wu. How to ask good questions? try to leverage paraphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6130–6140. Association for Computational Linguistics, 2020.

[2] Q. Huang, M. Fu, L. Mo, Y. Cai, J. Xu, P. Li, Q. Li, and Ho-fung Leung. Entity guided question generation with contextual structure and sequence information capturing. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13064–13072, 2021.

[3] W. Zhou, M. Zhang, and Y. Wu. Question-type driven 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)*, pp. 6032–6037, 2019.

[4] X. Du, J. Shao, and C. Cardie. 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)*, pp. 1342–1352, 2017.

[5] X. Sun, J. Liu, Y. Lyu, W. He, Y. Ma, and S. Wang. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3930–3939, Brussels, Belgium, 2018.

[6] Y. Chen, L. Wu, and Mohammed J Zaki. Reinforcement learning based graph-to-sequence model for natural question generation. *arXiv preprint arXiv:1908.04942*, 2019.

[7] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, H.W.Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in neural information processing systems*, vol. 32

[8] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2401–2410, 2020.

[9] Z. Fei, Q. Zhang, and Y. Zhou. Iterative GNN-based decoder for question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2573–2582, 2021.

[10] Do Hoang Thai Duong, Nguyen Hong Son, Hung Le, and Minh-Tien Nguyen. Learning to generate questions by enhancing text generation with sentence selection, 2022.

[11] M. Lewis, Y . Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V . Stoyanov, and L. Zettlemoyer. Exploring the limits of transfer learning with a unified text-to-text transformer. In *The Journal of Machine Learning Research*, vol. 21, pp. 5485–5551, 2020.

[12] M. Lewis, Y . Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V . Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *arXiv preprint arXiv:1910.13461*, 2019.

[13] F. Shiri, T.Y. Zhuo, Zhuang Li, S. Pan, W. Wang, R. Haffari, Y.F. Li, and V. Nguyen. Paraphrasing techniques for maritime qa system. In *2022 25th International Conference on Information Fusion (FUSION)*, 2022.

[14] Brian Thompson and Matt Post. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 90–121, 2020.

[15] Dinesh Nagumothu, Bahadorreza Ofoghi, Guangyan Huang, and Peter W Eklund. Pie-qg: Paraphrased information extraction for unsupervised question generation from small corpora. *arXiv preprint arXiv:2301.01064*, 2023.

[16] Karras, Tero and Laine, Samuli and Aila, Timo. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019

[17] Pranav Rajpurkar, J. Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[18] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni,P. Bachman, and K. Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, 2017.

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

[20] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

[21] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, 2014.