Controllable Question Generation with Semantic Graphs

Zhifei Xu

***Abstract*—Generating questions from answers and articles is an interesting and difficult task. Recent works have mostly focused on quality of generating a single question from a given article. However, question generation is not only a one-to-one problem that maps an article to one single question, because everyone may focus on different parts of the article and ask different questions accordingly. This makes the diversity of generated questions equally important with the quality.**

In this paper, we discuss the quality and diversity of question generation, and propose a controllable question generation model to improve these two aspects. Specifically, we associate the article with the dependencies parse tree and merge it with the article by constructing different triples. Secondly, because triples are different, we can generate different questions based on different triples and articles. By selecting different triples, we can control the content of generated questions and improve both the quality and the diversity. Experiment results on the SQuAD dataset show that our proposed method can significantly improve the diversity of generated questions, especially from the perspective of using different question types. Compared with the existing methods, our model achieves a better trade-off between the quality and diversity of generated questions, and we can generate diverse questions in a more controlled way.

***Index Terms*—Automatic question generation, semantic graphs, pretraining language models, BART.**

I. Introduction

As the reverse task of question answering (QA), question generation (QG) is the task of automatically generating a question as output that from a given passage or context as input data. In fact, question generation can be applied to a variety of fields, one use case is in the realm of dialogue. Firstly, it can serve as a cold starter to start a topic or ask questions to get feedback. Secondly, it can be applied to automatic consultation system. Another use case is the computer-assisted/guided learning. In online education, the question generation system can generate different type of questions for exam or quiz in courses to test the knowledge acquisition of learners.

In real-world, questions often come in various types and forms, e.g., short answer, open-ended, multiple-choice, and gap questions. In question generation, an article can generate a board range of questions, which may include any aspect, such as what, how or comparison with two keywords, etc. Moreover, different real-world applications also demand different questions. In a conversational system, we need to ask customer for name, address, gender and so on; but in quiz generation we need factoid questions to test learner's knowledge, such as *Is the Mississippi River the largest river in the world?* Therefore, to build a QG system that can adapt to different application scenarios, it is crucial to endow the system with the ability to *control* the types or contents of the questions being generated.

Building a controllable question generation system is not a trivial task. In traditional question generation models, most of them use the encoder-decoder model to generate questions, such as Seq2Seq [1] model, BART [2] model, and T5 [1] model, as shown in Fig. 1. The generated questions only have a one-to-one mapping relationship with article which means if we use one context as input, its output can only generate one question. Therefore, questions generated from the encoder-decoder model are monotonous, not diverse. Due to the limitations of the traditional encoder-decoder model, an article can only produce one question, and the same article can not produce different types of questions, so it is hard to produce diversity of questions in one context.

To address the above challenges, in this paper, we propose a novel question generation method, which achieves the controllable question generation by first selecting what contents should be included in the question and then realizing the question in natural language. Through this model, we can easily generate different types of questions based on a given context, and the model can be refined to accurately produce controllable questions.

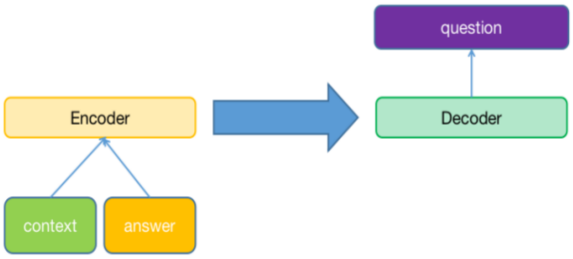

Fig. 1. General framework for encoder-decoder model.

It is common for QG systems to use a combination of semantic pattern matching, syntactic features, and template methods to create questions. Typically, these systems look for patterns of syntax, keywords, or semantic roles that appear in input document. This paper also adopts a similar idea. Our model proposes to use a dependency parse tree to convert all important elements of the article into different triples and each triple can represent an aspect to ask about ("questioning point") for the article. Then we randomly select triples and merge the selected triples with the source article. We train the model with the original text and the selected triplet as the new input data. Finally, by using the predicted data and completing the training model, we put these data in our model and the predicted questions are obtained. Following the steps described above, if we had chosen different triples at the beginning, we would have obtained different questions from the model. This is how we control the article so that the question is manageable and therefore

Manuscript received September 16, 2021; revised May 5, 2022.

Zhifei Xu is with the Cornell College, China (email: ZXu22@cornellcollege.edu).

doi: 10.18178/ijmlc.2022.12.6.1122

can generate different types of questions.

To verify the effectiveness of our proposed model, we evaluate our model on the SQuAD dataset [2] and make comparison with other models. We compare our model with three state-of-the-art question generation models: Seq2Seq [1], BART [2], and Google-T5 [3]. The experimental results demonstrate the feasibility of our model compared to other models on the SQuAD dataset.

Compared with the traditional Seq2Seq, BART, and T5 models, our proposed model not only generates more accurate questions but also generates more diverse questions. It shows that our model can generate higher quality questions, and the generated questions meet the needs of real life, it is satisfied the needs of people in different scenarios, and present forward questions that are conform with people's needs.

Overall, the main contributions of this paper are as follows these three ways:

1) This paper presents a novel model for controllable question generation based on the dependencies parse tree method to control question generation
2) The model controls the type of question based on the triples generated by the article, and after merging with the article, randomly selects the required triples to generate the question
3) Our model can generate higher quality questions, and we can generate more diversity of questions through a controlled method.

II. Related Work

In recent years, automatic question generation has attracted increasing attention from the natural language generation community, and many datasets reflect this, such as SQuAD, HotpotQA [3], and so on. Traditional question generation methods are mainly based on rules, which first convert the source information in the data set into syntax or semantic representation, and then use the model to generate relevant questions. However, these methods rely on strict heuristic rules to a large extent, which is not easy to be generalized.

Compared with the rule-based method, they [6], [7] present that the neural network method has a sequence to sequence framework and attention, and shows promising results when combining rich answer perception and attention mechanisms. These models are trained in an end-to-end way and are more flexible. It [8] put forward and integrate answer centered information to improve the correlation between answers and questions.

Some work noticed the inherent diversity of QG and came up with ways to consider this characteristic. It [9] uses language features and additional sentences as variables to embed them for modeling. However, these characteristics cannot be closely related to diversity. Recently, [10] proposes a method based on knowledge graph encoder, which aims to generate triples for entity domain, description information and relationship hierarchy information. Different from their methods, we use the method of dependencies parse tree to generate triples and use potential variable triples for modeling, which may capture more changes inherent in content selection, so as to achieve a controllable method generation problem.

III. Method

We propose a structural framework similar to the encoder-decoder model, which has two novel features for question generation (QG):

1) The key information in the article is filtered and extracted by using of dependencies parse tree, so that triples can be generated by the article, and the controllable effect of the question is better than realized through the triples.
2) Through the selective task of triples, the context and the triples which you selected are recombined into a new training task to decode, thus increasing the validity of the text and the precise reasoning of the question to constitute the diversity of the question.

Fig. 2 shows the main idea of the framework in the model, which consists of three modules: the generation of triples, which builds triples based on dependencies parse tree for known input context; random selective the representation of triples in each sentence, in order to simulate the process of choosing triples artificially; then let the first and second parts are combined to build novel training data and generate controllable questions through decoding steps. In the following sections, we describe the details that make up each module.

A. Graph Parser

As mentioned in the introduction, the relationship between triples is an important clue to determine the content of the context and the type of reasoning question. In order to extract effective information from documents, we explore a method that is based on dependencies parse tree [11] to generate and construct triples according to the context.

1) Dependencies parse tree graph of triples

Based on the dependencies Parse Tree framework, we complete the data preprocessing work through three tasks. (1) Find the only head *H* node that is not contained to aimed paragraphs. Its function is to confirm the corresponding position in the context based on the question and answer from data. (2) Find the head *H* node that relies on and then get the words. In this step, *H* is taken as a child node to retrieve the superior node of H, and the relevant text is obtained to ensure that all information about this question and answer in the context paragraph is successfully obtained. (3) Find the words that are linked to *H* nodes but not aimed paragraphs. Its goal is to get the body of the other word outside of the paragraph. Step 2 and 3 are sibling steps. In this way, the triples are obtained, and the articles are associated with the triples.

B. Controllable Question Generation Model

In our model, the data is need preprocessed at first, as mentioned in the method of Section I, using dependencies Parse Tree, which have genertated and constructed valid triples from the context is based on the question and answer. We use dependencies parser tree to analysize and generate triplets, then we according to the answer from context to filter which triplet is we want. After analyzing and judging the superior nodes of the target root node triplet, the triplet composed of effective information is combined with the context into a new data set, which is the process of datapreprocessing. After finishing the data preprocessing, we need to put the processed data into the Encoder-Decoder model. To simulate the uncertain factors by random selection, all the data of the SQuAD dataset are processed again in this model, and training the model. To confirm that the randomly selected data belongs to the valid part of the content, can form the expected effect of the question, to form a novel and the model of controllable question generation.

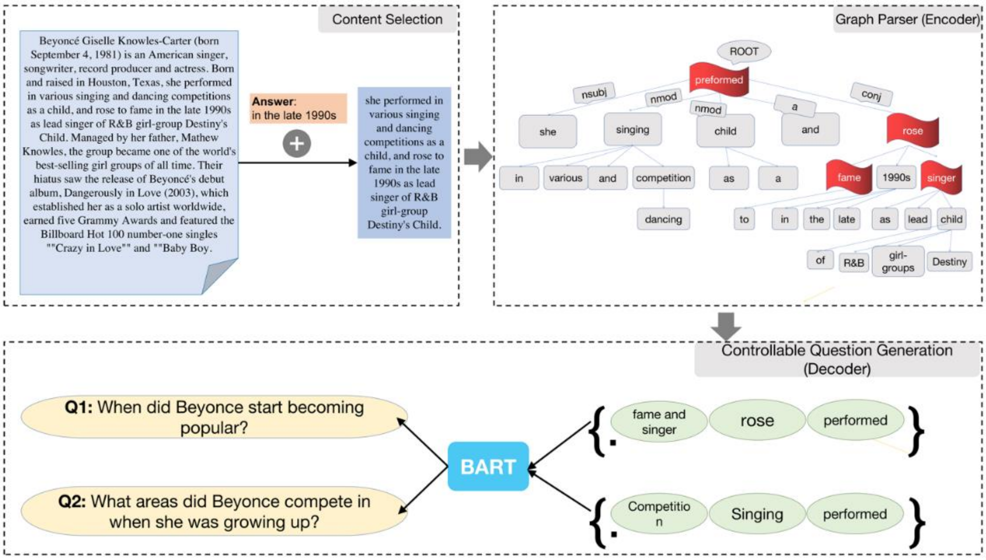
Content Selection
Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles ""Crazy in Love"" and ""Baby Boy.
Answer: in the late 1990s
she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child.
Graph Parser (Encoder)
ROOT
preformed
nsubj
nmod
nmod
a
conj
she
singing
child
and
rose
in
various
and
competition
as
a
fame
1990s
singer
dancing
to
in
the
late
as
lead
child
of
R&B
girl-groups
Destiny
Controllable Question Generation (Decoder)
Q1: When did Beyonce start becoming popular?
Q2: What areas did Beyonce compete in when she was growing up?
BART
fame and singer
rose
performed
Competition
Singing
performed

Fig. 2. Framework of our proposed method method.

C. Training and Texting Time

1) Training

Based on the input representation of the context, we train the generation question through a model framework similar to the coder-decoder. We use the Dependencies Parse tree to approach the preprocess the data. As the preprocessing of the data, we conduct a directional search based on the questions and answers in the data to lock the original part of the question. Use dependencies Parse tree to form the desired part of the triple. The next step is to integrate the article with the obtained triples to complete the data preprocessing. On the other hand, because the result of generating a triple is not exactly the same or completely irrelevant information. Therefore, in order to simulate the process of a random selection of triples, we chose to predict all the information of triples during training to ensure that the accuracy of question generation meets the expected requirements. Finally, the preprocessed data and the corresponding question of each data are matched. The input port is the pretraining data, and the output port is the question Q. All the preprocessed data and questions are put into the code-decoder Model for training, to obtain the Controllable Question Generation Model (CQGM) [12].

2) Testing

After completing the operation of model training, we also need to preprocess the data for the test data set. However, in the test dataset, we removed the answer reference information from the question-answer section and kept only the triple section and the context as the test data. To prevent real life, the unknown answer of the data can not be effectively predicted. Finally, the test data set is put into the trained Controllable Question Generation Model for prediction, and bleu is used to evaluate and analyze the prediction questions. Meanwhile, to achieve test the diversity of generated questions, we also use self-bleu to evaluate and analyze the diversity of questions. As stated in the introduction, the model can be able to generate higher quality questions, and we can generate more variety of questions with a controlled approach.

IV. Experiment

A. Dataset

To assess the model's ability to generate the types of questions, we experimented with the SQuAD dataset, which consisted of 84773/11872 articles, each with a question attached. It contains questions and required triples inferred from the text and answers. In the QG task, we will support text and answers as input to generate questions. However, the current encoder-decoder model is difficult to generate accurate questions from the original text. At the same time, encoder-decoder model only supports one-to-one generation mode, which cannot generate multiple questions and controllable question types in one article. Therefore, the original data set was preprocessed to select the question type, that is, according to the triples and answers as evidence, the article was combined as the input data set. We set batch size to 32 and epoch to 5. Following the previous work, we used BLEU1-4 [13], METEOR [14], Rouge-Land [15], Self-BLEU [16] as automated evaluation indicators. BlEU measures the overlap times of words in N-gram in reference by using weighted set average. BLEU's n-gram overlap idea is applied specifically to machine translation and text summary evaluation by METEOR and ROUGE-L respectively. Based on Bleu, Self-Bleu changes the model to generate the three questions with the highest probability and calculates the types of question and the probability of generating the diversity of questions. In addition, we performed a human assessment, in which the annotator assessed the quality of deep-seated question generation in three important respects: fluency, relevance, and complexity.

TABLE I: PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS ON THE SQUAD DATASET

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	SELF-BLEU
Seq2Seq	0.2428	0.0937	0.0470	0.0260	0.0743	0.2583	0.9993
Bart	0.2839	0.1373	0.0810	0.0516	0.1197	0.2854	0.9167
Google T5	0.3010	0.1453	0.0866	0.0557	0.1186	0.2928	0.9146
Our	0.4064	0.2607	0.1887	0.1428	0.1906	0.3965	0.4897

B. Baseline

We compare our proposed model with several baselines for question generation.

1) **Seq2Seq:** The basic Seq2Seq [1] model in AllenNLP, which uses the context as input data to decode and generate the question.
2) **BART:** The Bidirectional and Auto-Regressive Transformer or BART [2] is a Transformer that combines the Bidirectional Encoder (BERT) [17] with an Autoregressive decoder [18] into one Seq2Seq model. It uses a standard transformer-based neural machine translation architecture BART is a denoising autoencoder for pretraining Seq2Seq Models. (Data Framework ["input text", "target text"] input text is the pretraining data and source document. Target text is predict question from model)
3) **Google T5**: In "Exploring the Limits of Transfer Learning with a Unified text-to-text Transformer" [19], it present a large-scale empirical survey to determine which transfer learning techniques work best and apply these insights at scale to create a new model that call the Text-To-Text Transfer Transformer (T5). Where input and output are always text strings, bert-style models can output only class labels or input spans. The T5 text-to-text framework allows the same models, missing functions, and hyperparameters to be used in any NLP task, including machine translation, document summarization, question solving, and classification tasks. (Data Framework ["prefix", "input text", "target text"] prefix is index for source data, input and target text are same with BART model)

1) Implementation details part

To ensure the unique variables and fairness of the experiment, we use the three models described above with our proposed model applied to the SQuAD dataset. All baselines are made using a 1-layer GPU document encoder and question decoder. For the batch size of the model, we set it as 32, and the training epochs were 5 times.

C. Comparison with Baseline Models

Experimental results compared with all baseline methods are shown in Table I. We have two main observations:

1) Our model consistently outperforms all other baselines in BLEU. Specifically, our model based on the combination of triples and articles has advantages in bleU1-4, METEOR and rouge-L scores. We use the same encoder-decoder model as Seq2Seq, BART and T5 models. This shows that our model has a significant effect on the accuracy of question generation.
2) Compared with other methods, the results in Self-Bleu also show the diversity of questions generated by our model. We set the number of questions returned per article to three. In other words, all four models, including ours, return the top three questions with the highest probability. Then the absolute improvement of over 0.43 was found by comparing the Self-Bleu value, so the similarity degree of the three generated questions was judged to be diverse and capable of generating different types of questions, while the similarity degree of the questions generated by other models was between 0.91-0.99. This means the questions are basically the same.

D. Impact of Graph Parser

Our model uses the form of dependencies parse tree to generate triples. Specifically, the answer is used to lock the useful information in the text and the header node related to the answer. For example, the head node of 1990s is rose, and the others are 1990s. Therefore, record the head node rose. The second is to find the parent node of the node according to the node obtained in the first step. For example, due to rose dependent on performed node, then record the father head node is performed. The last step is to find the part related to the problem in the peer node of the first step. Both fame and singer are depend on rose but not in the range of answer. The triples performed, rose, singer and fame are obtained through the above steps.

In the scene graph parser [20], it is used as a python toolkit similar with Stanford scene graph based on parsing sense. Its parser is written entirely in Python. This project is inspired by the Stanford Scene Graph Parser [21]. At the same time, it has an easy-to-use user interface and configuration design. It parses sentences into graphs, where nodes are nouns (with modifiers, such as determinants or adjectives), and edges are the relationships between nouns. As in our example, she, variable fame singing and dancing competition, destination's child in 1990s are used as triples.

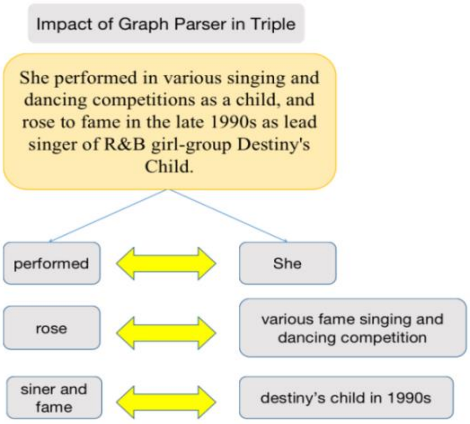
Impact of Graph Parser in Triple
She performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child.
performed
She
rose
various fame singing and dancing competition
siner and fame
destiny's child in 1990s

Fig. 3. Comparison of different parsers.

By contrast, our approach is based on grammatical information. The resulting triples are also more grammatical logic, which is a shallow analysis method. The method of scene graph parser depends on semantic knowledge. However, the accuracy of the triples obtained based on semantics is not as good as our method. We show examples of the results returned by different parsers in Fig. 3. We can also see that there is too much useless information in the generated triples, such as variable competition, destination's child. Therefore, using our method, we can get the key information about the components of triples more accurately.

E. Result and Analysis

1) Qualitative analysis

The experimental results on the SQuAD dataset are shown in Table I. The table shows that the scores of Bleu1-4, meteor and rouge-l in the quality of the questions generated by our method have a score advantage of more than 0.1 in each average score of our model compared with seq2seq, Bart and T5, which shows that our method is more accurate in the quality of the questions and our questions are closer to the target questions from dataset. In addition, we highlight each generation of content selected from the model in Fig. 4, which shows the effectiveness of our content selection module.

2) Diversifying question types

In addition, through the measurement of question types, we can find that our model has significant improvements in coverage and diversity, which is caused by the triples in the article. The diversity of question types can be explicitly controlled by selecting different triples we can observe from Self-Bleu that the higher the score, the more similar the three questions. The smaller the score, the greater the difference with the continuous growth of its value, the diversity is decreasing. Through calculation, it can be seen that our model has an absolute score difference of 0.43 and is ahead of other models in performance. From the perspective of diversity, the performance of our method is significantly better than other models, so as to achieve the best trade-off between diversity and quality, as shown in each index in the table. In addition, we show examples of the generated questions for each baseline method in Fig. 4. The examples show that our model can generate more diverse questions than baselines such as BART and Seq2Seq.

F. Human Evaluation

We conducted a human assessment of 100 random test samples, including 70 basic questions and 30 diversity questions. We asked eight staff members to rate the 100 questions generated and the basic authenticity between 1 (poor) and 5 (good) according to three criteria:(1) Fluency, indicating whether the question is grammatical and logical; (2) Relative, indicating whether the question is answerable or relevant to the article; (3) Answerability indicates whether questions can be answered according to the article. We average the rater's score on each question and report on the performance of the four models in Table II. The rater does not know the identity of the model beforehand. Let's explain two observations in detail:

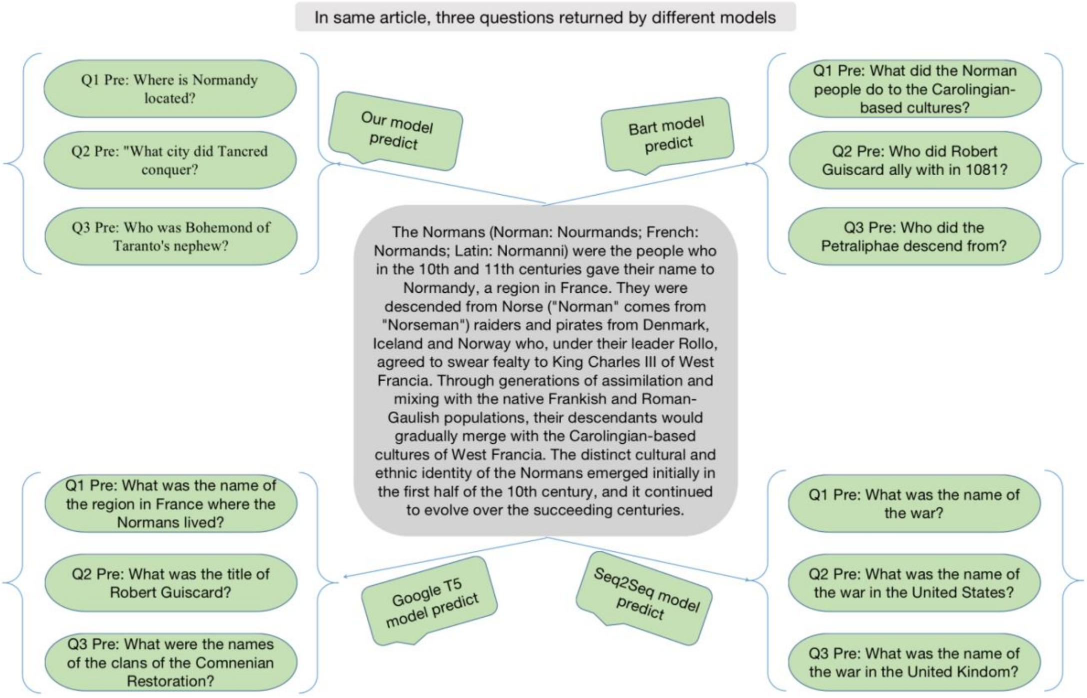
In same article, three questions returned by different models
Q1 Pre: Where is Normandy located?
Q2 Pre: "What city did Tancred conquer?
Q3 Pre: Who was Bohemond of Taranto's nephew?
Our model predict
Bart model predict
Q1 Pre: What did the Norman people do to the Carolingian-based cultures?
Q2 Pre: Who did Robert Guiscard ally with in 1081?
Q3 Pre: Who did the Petraliphae descend from?
The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.
Q1 Pre: What was the name of the region in France where the Normans lived?
Q2 Pre: What was the title of Robert Guiscard?
Q3 Pre: What were the names of the clans of the Comnenian Restoration?
Google T5 model predict
Seq2Seq model predict
Q1 Pre: What was the name of the war?
Q2 Pre: What was the name of the war in the United States?
Q3 Pre: What was the name of the war in the United Kindom?

Fig. 4. Example of generated questions for different models.

1) Compared with other models, improvements in answerability (+0.7) and correlation (+0.6) were more significant than in fluency (-0.01). The reason is that the baseline produces thinner questions (affecting correlation) or questions with grammatical errors (affecting fluency) Moreover, this is why our model can only be as fluent as other models. These indicate that our model, by combining grammar diagrams, produces fewer grammatical errors and uses more text.
2) When more than one question is generated, all metrics generally decrease, with the "correlation" decreasing most noticeably. When an article is entered, it is difficult for the model to capture the point of the question and make correct inferences, resulting in many similar problems. As the semantics and input noise increase, the quality of the question decreases.

TABLE II. HUMAN EVALUATION RESULTS

Model	Fluency	Relative	Answerability
Bart	3.69	2.3	3.59
T5	4.62	2.3	4.76
Seq2Seq	4.88	1.53	3.7
Our	4.6161	2.91	4.41

V. CONCLUSION AND FUTURE WORKS

In this paper, we explicitly diversify question generation from the perspective of dependencies Parse Tree generating triples and how triples are expressed. We introduce dependencies Parse tree for triplet generation and model concerns with random triples to allow control over the type of question generation. To ensure that random selection meets our needs, we consider various expressions of triples. On a common data set, our approach achieves an optimal trade-off between the quality of the question generated and the diversity of question types. Further analysis also proves the validity of our proposed model components.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

All the work are contributed by myself.

REFERENCES

[1] I. Sutskever, O. Vinyals, and V. Quoc Le, “Sequence to sequence learning with neural networks,” in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 3104-3112.

[2] M. Lewis, Y. H. 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,” arXiv preprint arXiv:1910.13461, 2019.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Q. Zhou, W. Li, and J. Peter Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” arXiv preprint arXiv:1910.10683, 2019.

[4] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” arXiv preprint arXiv:1606.05250, 2016.

[5] Z. L. Yang, Q. Peng, S. Z. Zhang, Y. S. Bengio, W. W. Cohen, R. Salakhutdinov, and D. Christopher Manning, “Hotpotqa: A dataset for diverse, explainable multi-hop question answering,” arXiv preprint arXiv:1809.09600, 2018.

[6] Q. Y. Zhou, N. Yang, F. R. Wei, C. Q. Tan, H. B. Bao, and M. Zhou, “Neural question generation from text: A preliminary study,” in *Proc. National CCF Conference on Natural Language Processing and Chinese Computing*, 2017, pp. 662-671.

[7] V. Harrison and M. Walker, “Neural generation of diverse questions using answer focus, contextual and linguistic features,” arXiv preprint arXiv:1809.02637, 2018.

[8] W. J. Zhou, M. H. Zhang, and Y. F. Wu, “Question-type driven question generation,” arXiv preprint arXiv:1909.00140, 2019.

[9] Y. Zhao, X. C. Ni, Y. Y. Ding, and Q. F. Ke, “Paragraph-level neural question generation with maxout pointer and gated self-attention networks,” in *Proc. the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3901-3910.

[10] S. Bi, X. Y. Cheng, Y. F. Li, Y. Z. Wang, and G. L. Qi, “Knowledge-enriched, type-constrained and grammar-guided question generation over knowledge bases,” arXiv preprint arXiv:2010.03157, 2020.

[11] S. Schuster, R. Krishna, A. Chang, F. F. Li, and D. Christopher Manning, “Generating semantically precise scene graphs from textual descriptions for improved image retrieval,” in *Proc. the Fourth Workshop on Vision and Language*, 2015, pp. 70-80.

[12] Y. Cheng, S. Y. Li, B. Liu, R. H. Zhao, S. J. Li, C. H. Lin, and Y. F. Zheng, “Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting,” arXiv preprint arXiv:2105.11698, 2021.

[13] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proc. the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.

[14] A. Lavie and A. Agarwal, “METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments,” in *Proc. the Second Workshop on statistical Machine Translation*, 2007, pp. 228-231.

[15] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” *Text Summarization Branches Out*, 2004, pp. 74-81.

[16] D. Alihosseini, E. Montahaei, and M. S. Baghshah, “Jointly measuring diversity and quality in text generation models,” in *Proc. the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 2019, pp. 90-98.

[17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[18] J. Kasai, N. Pappas, H. Peng, J. Cross, and A. Noah Smith, “Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation,” arXiv preprint arXiv:2006.10369, 2020.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Q. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” arXiv preprint arXiv:1910.10683, 2019.

[20] H. Wu, J. Y. Mao, Y. F. Zhang, Y. N. Jiang, L. Li, W. W. Sun, and W. Y. Ma, “Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations,” in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6609-6618.

[21] S. Schuster, R. Krishna, A. Chang, F. F. Li and D. Christopher Manning, “Generating semantically precise scene graphs from textual descriptions for improved image retrieval,” in *Proc. the Fourth Workshop on Vision and Language*, 2015, pp. 70-80.

Zhifei Xu was borned in 2000 of China. Zhifei Xu is a senior student in Cornell College and his bachelor degree is computer science major and applied mathematics minor in Cornell College, Mount Vernon, Iowa state and United State. Zhifei Xu will graduate in 2022. His main interests in computer science are data processing, database applications, machine learning, natural language processing, image processing and so on.