Diversifying Question Generation over Knowledge Base via External Natural Questions

Shasha Guo[1, 2], Jing Zhang[1, 2 *], Xirui Ke[1, 2], Cuiping Li[1, 2], Hong Chen[1, 2]
[1]School of Information, Renmin University of China, Beijing, China
[2]Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education
{guoshashaxing, zhang-jing, kexirui, licuiping, chong}@ruc.edu.cn

Abstract

Previous methods on knowledge base question generation (KBQG) primarily focus on refining the quality of a single generated question. However, considering the remarkable paraphrasing ability of humans, we believe that diverse texts can express identical semantics through varied expressions. The above insights make diversifying question generation an intriguing task, where the first challenge is evaluation metrics for diversity. Current metrics inadequately assess the aforementioned diversity. They calculate the ratio of unique n-grams in the generated question, which tends to measure duplication rather than true diversity. Accordingly, we devise a new diversity evaluation metric, which measures the diversity among top-k generated questions for each instance while ensuring their relevance to the ground truth. Clearly, the second challenge is how to enhance diversifying question generation. To address this challenge, we introduce a dual model framework interwoven by two selection strategies to generate diverse questions leveraging external natural questions. The main idea of our dual framework is to extract more diverse expressions and integrate them into the generation model to enhance diversifying question generation. Extensive experiments on widely used benchmarks for KBQG show that our approach can outperform pre-trained language model baselines and text-davinci-003 in diversity while achieving comparable performance with ChatGPT.

Keywords: Diversifying Question Generation, Knowledge Base, Diversity Metric

1. Introduction

Knowledge Base Question Generation (KBQG) is an essential task, which focuses on generating natural language questions based on a set of formatted facts extracted from a knowledge base (KB). In recent years, KBQG has attracted substantial research interest due to its wide range of applications. For example, in education, KBQG can generate numerous questions from course materials, aiding in assessing students' grasp of the content and enhancing self-learning (Liu et al., 2021; Wang et al., 2021b). Furthermore, in industry, KBQG can encourage machines to actively ask human-like questions in human-machine conversations (Ling et al., 2020; Zeng and Nakano, 2020). Additionally, KBQG can augment training data to boost the quality of question answering (QA) tasks (Chen et al., 2023; Guo et al., 2022).

Recently, pre-trained language models (PLMs)-based methods (Chen et al., 2020; Ke et al., 2021; Xiong et al., 2022) have achieved advanced performance on KBQG. Despite the success of these models, they conform to the one-to-one encoder-decoder paradigm and concentrate on improving the quality of the generated question, resulting in insufficient diversity. In fact, human communication exhibits a remarkable ability to paraphrase, which means humans can express the same semantics in different surface forms, such as words, phrases, and grammatical patterns. Figure 1 gives an illustration of several diverse questions (those generated by ours), expressing the same semantics from the identical input triples of KBs. Intuitively, we think that the diversity of texts should be that texts expressing the same semantics have different forms of expression.

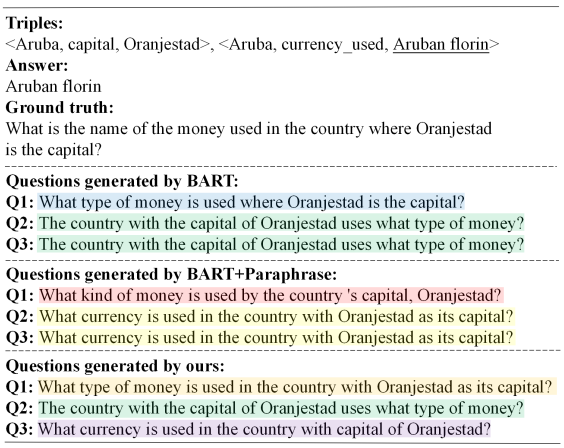

Figure 1: Example questions generated by BART, BART+Paraphrase, and our approach on the WebQuestions (WQ) dataset. Given a set of triples with the answer(underlined), each method returns top-3 questions, where the various surface forms are marked in different colors.

However, current evaluation metrics about relevance and diversity deviate from the above obser-

*Corresponding author.

LREC-COLING 2024, pages 5096–5108
20-25 May, 2024.

vation. To measure the relevance between the generated question and the ground truth, BLEU (Papineni et al., 2002a) and ROUGE (Lin and Och, 2004) have been proposed, which simply calculate the ratio of common n-grams in two texts without considering semantics. Unlike the above metrics for computing n-grams similarity, BERTScore (Zhang et al., 2020) further measures token-level semantic similarity between two texts. Nonetheless, we believe that sentence-level semantic similarity between the two texts is more important. As for the diversity of the generated question, metrics like Distinct-n (Li et al., 2016) measure the percentage of unique n-grams within the question itself. It can be viewed as a measure of the duplication of n-grams in the generated question, which does not conform with the diversity defined above.

Towards these evaluation metrics, existing PLMs-based models fail to mimic humans to produce diverse questions. For one thing, the metrics for assessing diversity are inappropriate. For another, these models strive to make the generated question similar to the ground truth question, which limits and narrows down the search space when decoding the output. Therefore, a natural solution is to expand the search space by increasing the ground truth question. To illustrate such an idea, we first perform a pilot study (Cf. the detailed settings in Section 3.2) by conducting one preliminary experiment that augments the ground truth questions with automatically paraphrased questions. We make the observation that **injecting paraphrased questions results in a more diverse set of generated questions.** However, due to the limited capability of the paraphrase model, the paraphrased questions exhibit only slight vocabulary variations compared to the ground truth question, which leaves the potential for further exploration.

Inspired by the above insights, in this paper, we propose a novel diversity evaluation metric called $Diverse@k$, which measures the diversity among the top-k generated questions for each instance while ensuring their relevance to the ground truth. Additionally, we investigate the use of a wide range of external natural questions (Kwiatkowski et al., 2019) to enhance the diversity of question generation. We believe that large-scale external corpus, instead of being restricted to a small amount of training data, can provide a wider and more diverse range of linguistic and semantic expressions.

To extract rich expressions and squeeze them into the generation model, we design two dual models, namely the forward model and the backward model, to transform the formatted facts like triples into the natural question and vice versa respectively. We further design two simple yet effective selection strategies of pseudo pairs to interweave the two models. The first selection strategy is imposed on the output of the backward model. Given an external question and the outputted triples, we first calculate the sentence-level semantic score (Gao et al., 2021) between this external question and the question generated from the triples using the forward model. Then we utilize the semantic score to discern reliable pseudo pairs, which can help select diverse natural expressions that still maintain semantic similarity to the training data. The second strategy is applied to the output of the forward model. Given the triples from the training data and the outputted top-k generated questions, semantic relevance and diversity scores are used together to sift out similar but different questions for each instance, improving the capacity of the backward model for dealing with the external questions. As a result, as the pseudo data flows through the two models, a considerable variety of natural questions resembling the training data are assembled gradually, which far exceeds the paraphrased questions.

Contributions. (1) To the best of our knowledge, we are the first to propose the diversity among top-k generated questions for each instance, ensuring their relevance to the ground truth, and design a novel metric to measure the diversity. (2) We present a dual model framework interwoven by two selection strategies to assemble a variety of diverse questions from external natural questions, enabling diverse expressions to be injected into the generation model. (3) Extensive experiments show our model consistently exhibits superior diversity. It surpasses pre-trained language model baselines and text-davinci-003 (Ouyang et al., 2022), while achieving comparable performance with ChatGPT[1].

2. Rethinking Diversity Evaluation

A metric is essential to evaluate a generation model's capacity to produce diverse questions from identical input, as illustrated in Figure 1.

Distinct-n. Previous works (Shao et al., 2021; Wang et al., 2021a) mostly use $Distinct\text{-}n$ (Li et al., 2016), *i.e.*, $Distinct\text{-}n = \frac{|unique\ n\text{-}grams|}{|total\ n\text{-}grams|}$, to calculate the diversity score of the generated text. Some works assess $Distinct\text{-}n$ in instance-level (Jia et al., 2020; Shao et al., 2021), while others treat all instances as a whole to compute the unique n-grams in the total n-grams of all instances (Wang et al., 2021a; Zhou et al., 2021). Clearly, neither of them is appropriate, as $Distinct\text{-}n$ actually focuses on the proportion of unique n-grams, which appears more akin to evaluating duplication than diversity.

Diverse@k. We propose $Diverse@k$ as a new metric to assess the diversity of the top-k generated questions for each instance. The main idea is to

[1] https://openai.com/blog/chatgpt

sum the pairwise diversity of the top-k generated questions. Specifically, given two generated questions S_i and S_j with the ground truth question S, $Diverse@k$ is defined as:

$$
\begin{aligned}
Diverse@k &= \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Diverse(S_i, S_j), \\
Diverse(S_i, S_j) &= \frac{|\mathcal{T}_i - \mathcal{T}_j| + |\mathcal{T}_j - \mathcal{T}_i|}{|\mathcal{T}_i \cup \mathcal{T}_j|}, \\
& R(S_i, S) \geq \alpha \text{ and } R(S_j, S) \geq \alpha
\end{aligned} \tag{1}
$$

where $\mathcal{T}_i$ and $\mathcal{T}_j$ are the sets of tokens in S_i and S_j, respectively, so $Diverse(S_i, S_j)$ measures their differences. Then we sum $Diverse(S_i, S_j)$ of all pairwise top-k generated questions to represent the diversity of the instance. Clearly, mere summation fails to accurately reflect the quality of the generated questions. Consequently, we impose constraints on semantic similarity to guarantee the relevance of each generated question. Specifically, we use simCSE (Gao et al., 2021), a popular semantic relevance metric, to measure the relevance score between S_i and the ground truth question S and denote it as $R(S_i, S)$. α is the threshold of the relevance score. We set it as **70%** (Cf. the detailed explanation in Section 4.2(3)) to filter out irrelevant questions. Moreover, we evaluate the correlation between $Diverse@k$ and human evaluation using the Pearson correlation coefficient. Table 6 reports the results on $Diverse@3$ (*i.e.*, **0.935**) and $Diverse@5$ (*i.e.*, **0.949**) respectively. Based on the results, we conclude that $Diverse@k$ aligns closely with human evaluation, highlighting its rationality.

3. Approach

3.1. Problem Definition

KBQG aims to generate questions given a set of triple facts represented as a subgraph. Formally, given a dataset $\mathcal{D} = \{(G_i, q_i)\}_{i=1}^{N}$, where G_i represents a subgraph consisting of a set of connected triples and q_i signifies the corresponding question, the objective of KBQG is to learn a function f with parameter θ to map from G_i to q_i, *i.e.*, $f_\theta : G_i \rightarrow q_i$.

3.2. Pilot Study

We conduct a preliminary experiment by paraphrasing target questions to show their positive effect on diversifying question generation.

Modeling. We employ t5-large-paraphraser-diverse-high-quality[2], an advanced paraphrase model to paraphrase ground truth questions in the training data automatically. We create three paraphrased questions (q_i^1, q_i^2, q_i^3) for each ground truth question q_i. We use BART-base[3] (Lewis et al., 2020) as the backbone of our generation model, as prior research (Chen et al., 2023) has demonstrated that BART results in state-of-the-art question generation performance. We fine-tune BART on the paraphrased dataset and denote the method as BART+Paraphrase (abbreviated as B+P).

Model	Diverse@10
BART	21.50
BART+Paraphrase	**29.05**
Gain	**7.55**

Table 1: Performance comparison between BART and BART+Paraphrase for KBQG on the WQ dataset (%).

Observation. Table 1 illustrates the evaluation results measured by our proposed metric $Diverse@k$ (*i.e.*, $Diverse@10$) . The results show that questions generated by injecting paraphrased ground truth are more diverse than those generated solely from the original ground truth, indicating that paraphrasing has a positive effect on enhancing diversifying question generation. Since these paraphrased questions contain much richer semantic patterns and expressions than the ground truth, the generation model can learn from them to obtain more diverse question expressions.

Discussion. Above we propose a simple but effective approach to construct one-to-many instances to expand the searching space of the generation model. Despite the advantages of these paraphrased questions, they only exhibit minor differences from the target questions and are limited in scale due to the inadequate capacity of the paraphrase model. A straightforward method is to make efforts to design promising paraphrase models, but we explore another way to acquire diversity by leveraging external natural questions. Compared with paraphrased questions, external natural questions can cover a much broader range of semantic patterns and language expressions. Moreover, natural questions are more human-like, while paraphrased questions are relatively rigid and mechanical. In view of this, we attempt to employ external natural questions to diversify question generation.

3.3. Model Overview

In this work, we leverage external natural questions denoted as $\mathcal{D}_\mathcal{Q} = \{Q_j\}_{j=1}^{M}$ to diversify question generation, where Q_j has no corresponding subgraph. Figure 2 illustrates the overview of our proposed approach. At a high level, our approach

[2]https://huggingface.co/ramsrigouthamg/t5-large-paraphraser-diverse-high-quality

[3]https://huggingface.co/facebook/bart-base

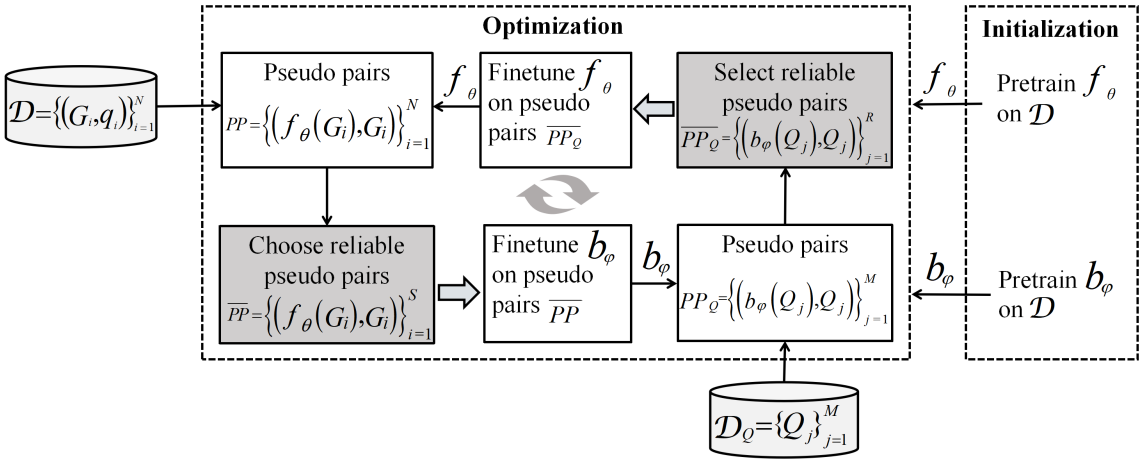

Figure 2: Overview of our proposed method. The forward model f_θ and the backward model b_φ are pre-trained on training data $\mathcal{D}$ and then iteratively fine-tuned on reliable pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^{R}$ and $\{(f_\theta(G_i), G_i)\}_{i=1}^{S}$ respectively, which are filtered by our proposed selection strategies.

consists of two steps including **initialization** and **optimization**. Motivated by the concept of back translation (Hu et al., 2022; Zhu et al., 2020), we propose a backward model to help the forward model capture diverse question expressions. First, the forward model f_θ and the backward model b_φ are pre-trained on the training data $\mathcal{D}$ to obtain a good initialization point. Next, we employ $\mathcal{D}_\mathcal{Q}$ to optimize f_θ and b_φ simultaneously. Specifically, we use b_φ to generate corresponding triple-based sequences of $\mathcal{D}_\mathcal{Q}$ to train f_θ. By doing this, f_θ can capture more diverse expressions from $\mathcal{D}_\mathcal{Q}$. Then, we use the above optimized f_θ to generate questions for the triple-based sequences in training data $\mathcal{D}$, which are then used to train b_φ. By doing this, the pseudo questions of the training data $\mathcal{D}$ can provide b_φ with more diverse expressions than the ground truth. Thus, the optimized b_φ can deal with various external questions to further offer more diverse pseudo data for f_θ. With continuous iterative training, f_θ and b_φ constantly gather substantial diverse questions for the training data, which are subsequently injected into f_θ along with each iteration to acquire diversity.

3.4. Step 1: Initialization

We use BART (Lewis et al., 2020) to instantiate the forward model f_θ. Concretely, we linearize each subgraph G_i into a triple-based sequence, where each triple is separated by the special token "</s>". Then, we input the sequence into BART to generate a corresponding question. We finally fine-tune f_θ on $\mathcal{D}$ by maximizing the probabilities of generating all gold questions, *i.e.*,

$$\mathcal{L}_f^{(0)} = \max_\theta \sum_{i=1}^{N} \log P_\theta\left(q_i|G_i\right) \tag{2}$$

Similarly, we also employ BART to instantiate the backward model b_φ. Specifically, each question q_i is fed into b_φ to generate a triple-based sequence. Then, we fine-tune b_φ on $\mathcal{D}$ by maximizing the probabilities of generating all gold triple-based sequences, *i.e.*,

$$\mathcal{L}_b^{(0)} = \max_\varphi \sum_{i=1}^{N} \log P_\varphi\left(G_i|q_i\right) \tag{3}$$

3.5. Step 2: Optimization

In this section, we explain how to iteratively fine-tune f_θ and b_φ on $\mathcal{D}_\mathcal{Q}$ and $\mathcal{D}$. Through this iterative fine-tuning, f_θ and b_φ accumulate a wide range of diverse patterns and expressions from $\mathcal{D}_\mathcal{Q}$, which endows the forward model with diversity.

Forward Model. We explain how to fine-tune the forward model f_θ with external natural questions $\mathcal{D}_\mathcal{Q} = \{Q_j\}_{j=1}^{M}$. Because questions in $\mathcal{D}_\mathcal{Q}$ do not have corresponding triple-based sequences, we use the backward model b_φ to generate them and construct the pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^{M}$. Then, the forward model f_θ is fine-tuned on these pseudo pairs by maximizing the probabilities of generating the external questions, *i.e.*,

$$\mathcal{L}_f = \max_\theta \sum_{j=1}^{M} \log P_\theta\left(Q_j|b_\varphi(Q_j)\right) \tag{4}$$

where $b_\varphi(Q_j)$ is an abbreviation for $b_\varphi(G_i|Q_j)$. Since b_φ is pre-trained on training data $\mathcal{D}$, $b_\varphi(Q_j)$ generally follows the patterns of triple-based sequences in $\mathcal{D}$. External questions $\mathcal{D}_\mathcal{Q}$ provide more semantic patterns and expressions, enabling f_θ to generate more diverse questions.

Backward Model. We elaborate how to fine-tune the backward model b_φ based on training data $\mathcal{D} = \{(G_i, q_i)\}_{i=1}^N$. Beyond the gold question q_i of G_i, we employ the forward model f_θ to generate the new question, namely $f_\theta(G_i)$. Because f_θ is fine-tuned on external natural questions $\mathcal{D}_\mathcal{Q}$, $f_\theta(G_i)$ could have different expressions from ground truth q_i. We organize $f_\theta(G_i)$ and G_i into the pseudo pairs $\{(f_\theta(G_i), G_i)\}_{i=1}^N$. Then, the backward model b_φ is fine-tuned on these pseudo pairs by maximizing the probabilities of generating all gold triple-based sequences, *i.e.*,

$$\mathcal{L}_b = \max_{\varphi} \sum_{i=1}^{N} \log P_\varphi \left(G_i | f_\theta(G_i)\right) \tag{5}$$

where $f_\theta(G_i)$ is an abbreviation for $f_\theta(q_i|G_i)$. Various $f_\theta(G_i)$ empower b_φ the ability to deal with a variety of external questions.

3.6. Reliable Pseudo Pairs Selection

During optimization, the forward model and the backward model are fine-tuned with pseudo pairs. To better assemble reliable diverse expressions for the training data, we propose two simple yet effective selection strategies.

First, we aim to extract diverse expressions from external questions to enrich the training data. To achieve this, we design the first strategy based on simCSE to filter the pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^M$ produced by the backward model b_φ. We calculate the simCSE score between Q_j and $f_\theta(b_\varphi(Q_j))$, where the latter is the question generated by the forward model f_θ taking $b_\varphi(Q_j)$ as the input. The more similar Q_j and $f_\theta(b_\varphi(Q_j))$ are, the more reliable $b_\varphi(Q_j)$ is because the forward model is easier to identify. Through this strategy, pseudo pairs with similar formatted facts but different expressions are selected to train the forward model further.

Subsequently, we need to further augment the extraction patterns for the backward model to enlarge the scale of selected external questions in subsequent steps. Therefore, to expand the search space of b_φ, we design the second selection strategy based on both SimCSE and our designed $Diverse(S_i, S_j)$. Specifically, for the pseudo pairs $\{(f_\theta(G_i), G_i)\}_{i=1}^N$ generated by the forward model, we first use simCSE to compute the relevance between $f_\theta(G_i)$ and q_i. This allows us to retain $f_\theta(G_i)$ with high semantic relevance, effectively filtering out noisy and irrelevant generated questions. The threshold of relevance score is set as 70% (Cf. the detailed explanation in Section 4.2(3)). Then, we calculate the diverse score $Diverse(f_\theta(G_i), q_i)$ to select the highest scoring pseudo pair $(f_\theta(G_i), G_i)$ for fine-tuning the backward model. The forward model provides a more varied expression for the formatted facts in the training data, indicating that the backward model can be fine-tuned to handle more external inquiries.

Algorithm 1: Our Proposed Approach

Input: $\mathcal{D} = \{(G_i, q_i)\}_{i=1}^N$, $\mathcal{D}_\mathcal{Q} = \{Q_j\}_{j=1}^M$.
Output: θ of the forward model f_θ, φ of the backward model b_φ.

1: Instantiate f_θ and b_φ via BART;
2: Pretrain f_θ on $\mathcal{D}$ via Eq. (2);
3: Pretrain b_φ on $\mathcal{D}$ via Eq. (3);
4: **for** each iteration **do**
5: **for** each epoch **do**
6: Generate M pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^M$;
7: Generate M pseudo questions $\{\hat{Q}_j\}_{j=1}^M$ = $\{f_\theta(b_\varphi(Q_j))\}_{j=1}^M$ on $\{b_\varphi(Q_j)\}_{j=1}^M$;
8: Select R reliable pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^R$ if semantic score between $\hat{Q}_j$ and Q_j is greater than 0.7;
9: Update θ based on $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^R$ via Eq. (4);
10: **end for**
11: **for** each epoch **do**
12: Generate N pseudo pairs $\{(f_\theta(G_i), G_i)\}_{i=1}^N$;
13: Choose S reliable pseudo pairs $\{(f_\theta(G_i), G_i)\}_{i=1}^S$ on semantic relevance and diversity between $f_\theta(G_i)$ and q_i;
14: Optimize φ based on $\{(f_\theta(G_i), G_i)\}_{i=1}^S$ via Eq. (5);
15: **end for**
16: Finetune f_θ on $\mathcal{D}$ via Eq. (2);
17: Finetune b_φ on $\mathcal{D}$ via Eq. (3);
18: **end for**
19: Return θ and φ.

The two selection strategies encourage two dual models to benefit by associating together and complementing each other. We summarize the whole procedure by Algorithm 1. In it, line 1 initializes the forward model f_θ and the backward model b_φ via BART. Lines 2-3 pretrain f_θ and b_φ based on $\mathcal{D}$ to get a good initialization point. Lines 4-15 iteratively train f_θ and b_φ util convergence. Line 6 generates the pseudo triple-based sequence $b_\varphi(Q_j)$ about Q_j using b_φ, and constructs M pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^M$. Line 7 adopts f_θ to generate M pseudo questions $\{\hat{Q}_j\}_{j=1}^M$ on $\{b_\varphi(Q_j)\}_{j=1}^M$. Line 8 chooses reliable pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^R$ if the semantic score between $\hat{Q}_j$ and Q_j is greater than the threshold 0.7 (Cf. the detailed explanation in Section 4.2(3)) . Line 9 optimizes the parameter θ of f_θ based on pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^R$. Line 12 employs f_θ to generate the pseudo question $f_\theta(G_i)$ about G_i, on which N pseudo pairs $\{(f_\theta(G_i), G_i)\}_{i=1}^N$ are constructed. Line 13 sifts

out S pseudo pairs $\{(f_\theta(G_i), G_i)\}_{j=1}^{S}$ according to semantic relevance and diverse of $f_\theta(G_i)$ and q_i, which guarantees that they are similar but different. Line 14 updates the parameter φ of b_φ on S reliable pseudo pairs $\{(f_\theta(G_i), G_i)\}_{i=1}^{S}$. To ensure that the performance after each iteration remains consistent with the original training dataset $\mathcal{D}$, we further fine-tune the forward model f_θ and the backward model b_φ based on $\mathcal{D}$ in lines 16 and 17.

4. Experimental Evaluation

4.1. Experimental Settings

Datasets. We evaluate our proposed approach on two widely used benchmark datasets WebQuestions (WQ) and PathQuestions (PQ) (Zhou et al., 2018). Specifically, WQ combines 22,989 instances from WebQuestionsSP (tau Yih et al., 2016) and ComplexWebQuestions (Talmor and Berant, 2018), which are further divided into 18989/2000/2000 for training/validating/testing. PQ contains 11,793 instances that are partitioned into 9793/1000/1000 for training/validating/testing.

Evaluation Metrics. We assess the generated questions from two aspects: **relevance** and **diversity**. For each instance, we evaluate the top-3, top-5, and top-10 generated questions respectively. We assess the relevance of generated questions in terms of semantics (*i.e.*, the meaning of the text). Specifically, we adopt simCSE (Gao et al., 2021) to calculate sentence-level semantic relevance between the generated question and the ground truth. Besides, we also report the token-level relevance using BLEU (Papineni et al., 2002b), which computes the ratio of the common n-grams between the generated question and the ground truth question. We evaluate the diversity of generated questions using our proposed $Diverse@k$, which measures the diversity among top-k generated questions for each instance while ensuring their relevance to the ground truth. In addition, we also report Distinct-n (Li et al., 2016) (abbreviated as Dist-n), which calculates the proportion of unique n-grams in the generated question. For QA, we use Hits@1 to evaluate whether the top-1 predicted answer is accurate and report the F1 score since some questions have multiple answers. For human evaluation, we invite three graduate students to measure the diversity and relevance of the generated questions.

Baselines. We compare two types of baselines: pre-trained language models-based (PLMs-based) and large language models-based (LLMs-based). Among PLMs-based baselines, **T5** (Raffel et al., 2020) and **BART** (Lewis et al., 2020), the state-of-the-art PLMs for text generation, are fine-tuned for KBQG. Concretely, we linearize each subgraph G_i into a triple-based sequence and then feed the sequence into BART and T5 to generate top-3, top-5, and top-10 questions. **JointGT** (Ke et al., 2021) proposes three novel pre-training tasks to learn graph-text alignments and develops a structure-aware semantic aggregation module inserted into the transformer layer to retain the graph structure. **T5+Paraphrase**, **BART+Paraphrase** and **JointGT+Paraphrase** (abbreviated as **T5+P**, **B+P** and **JointGT+P**) are T5, BART, and JointGT trained on the original and the paraphrased questions. Specifically, we first apply a popular paraphrase model, *i.e.*, t5-large-paraphraser-diverse-high-quality, to paraphrase the ground truth questions and then create three paraphrased questions for each ground truth question. Then, we fine-tune T5, BART, and JointGT using these paraphrased and original questions. For LLMs-based baselines, we compare two strong baselines, *i.e.*, **ChatGPT**[4] and text-davinci-003 (Ouyang et al., 2022) (abbreviated as **Davinci003**).

Pre-training. We use BART-base to instantiate the forward model f_θ and the backward model b_φ. For pre-training them, we set the learning rate as 5e-5, the batch size as 8, the beam size as 5, the patience as 5, and the maximum epochs as 20 for early stopping.

Fine-tuning. We iteratively fine-tune the forward model f_θ and the backward model b_φ with the same training settings as the pre-training process, but use the pre-trained BART-base as the backbone and train the two models on the generated reliable pseudo pairs $\{(b_\varphi(Q_j), Q_j)\}_{j=1}^{R}$ and $\{(f_\theta(G_i), G_i)\}_{i=1}^{S}$ respectively. We fix the number of iterations to 2 and 1 for the datasets WQ and PQ, which are chosen based on the results of the validation set.

Code Implementation. We implement our method using Pytorch and conduct all experimental evaluations on a server. This server is configured with a single Nvidia RTX A6000 GPU (48 GB) and equipped with 256 GB memory. Our code is available at GitHub[5].

4.2. Overall Evaluation

Table 2 and Table 3 report the overall evaluation results on PQ and WQ respectively. Bold formatting denotes the best results while underlining signifies the second best. According to the results, we conclude that: **(1) Injecting paraphrased questions can contribute to diversifying KBQG.** T5+P, B+P, and JointGT+P exhibit better diversity than their corresponding models without fine-tuning on paraphrased questions, demonstrating paraphrasing

[4] https://openai.com/blog/chatgpt

[5] https://github.com/RUCKBReasoning/DiversifyQG

Model	Top-3 Questions				Top-5 Questions				Top-10 Questions			
	simCSE	BLEU-1	Diverse@3	Dist-1	simCSE	BLEU-1	Diverse@5	Dist-1	simCSE	BLEU-1	Diverse@10	Dist-1
T5	87.04	52.35	22.50	34.67	86.75	52.30	25.57	23.67	86.16	52.47	29.80	14.67
BART	94.07	78.76	18.32	33.97	**93.37**	**77.76**	20.70	22.55	**92.06**	**76.21**	24.53	13.84
JointGT	**94.11**	**78.93**	18.66	33.96	93.28	77.74	21.26	22.69	91.99	76.05	25.22	14.01
T5+P	85.58	48.61	24.81	36.82	85.44	49.50	27.66	25.36	85.24	50.41	31.10	15.37
B+P	89.97	68.51	24.58	37.96	89.92	68.75	26.39	25.12	89.55	68.47	28.88	14.79
JointGT+P	90.09	68.89	24.44	38.18	89.97	68.75	26.76	25.28	89.56	68.30	28.91	14.66
Davinci003	77.06	39.33	28.14	38.95	76.86	39.32	30.18	27.27	76.94	39.38	31.46	16.97
ChatGPT	77.17	33.58	**29.87**	39.59	77.18	33.59	**32.04**	**28.68**	77.25	33.75	34.38	**18.04**
Ours	85.63	59.71	28.12	**40.62**	85.06	59.17	31.60	28.39	84.40	58.46	**35.85**	17.75

Table 2: Overall evaluation on PQ (%).

Model	Top-3 Questions				Top-5 Questions				Top-10 Questions			
	simCSE	BLEU-1	Diverse@3	Dist-1	simCSE	BLEU-1	Diverse@5	Dist-1	simCSE	BLEU-1	Diverse@10	Dist-1
T5	75.80	42.11	21.36	41.19	75.88	42.51	24.14	28.21	75.83	42.85	28.02	17.13
BART	82.42	51.64	16.88	42.46	82.30	51.59	18.87	29.50	82.02	51.18	21.50	18.22
JointGT	**82.64**	**52.01**	16.37	41.92	**82.52**	**51.90**	18.27	29.14	**82.24**	**51.61**	20.65	17.82
T5+P	77.97	42.76	22.42	42.27	78.04	43.10	25.53	29.26	78.04	43.48	29.78	17.92
B+P	81.24	48.77	22.97	41.31	81.16	48.74	25.63	28.30	81.03	48.74	29.05	17.06
JointGT+P	81.09	48.28	23.71	41.56	81.10	48.37	26.17	28.41	80.88	48.27	29.85	17.28
Davinci003	71.68	33.62	24.32	42.95	71.75	33.75	26.51	30.50	71.68	33.61	29.21	19.10
ChatGPT	74.54	33.21	**28.88**	**42.96**	74.38	33.14	**31.38**	**30.82**	74.40	33.09	**33.81**	**19.47**
Ours	80.58	49.95	25.17	42.52	80.55	49.94	28.05	29.57	80.31	49.85	31.33	18.27

Table 3: Overall evaluation on WQ (%).

can increase richer semantic patterns and expressions than original data, which helps to produce diverse questions. **(2) Our approach surpasses PLMs-based baselines in diversity and matches or even outperforms LLMs-based baselines, which demonstrates the effectiveness of leveraging external natural questions.** Although the three PLMs-based enhanced baselines considering paraphrased questions obtain better performance than their base version, their performance is constrained by the limited ability of the paraphrase model. Alternatively, our approach introduces external questions that cover a much broader range of semantic patterns and expressions. Furthermore, we note that our method generally outperforms Davinci003 in diversity and is comparable to ChatGPT. LLMs (such as ChatGPT) inherently possess richer semantic knowledge compared to PLMs (like BART, the backbone of our approach), leading to higher diversity scores. However, our method surpasses Davinci003, highlighting the effectiveness of our approach. **(3) Our approach achieves comparable performance to PLMs-based baselines in relevance and surpasses LLMs-based baselines.** We observe that our approach achieves slightly lower semantic scores in terms of simCSE than the best PLMs-based baselines, but outperforms LLMs-based baselines. Despite the difference between our method and the best baseline, the simCSE scores of our method already meet the relevance criterion. SimCSE scores of our approach are all greater than 80%, which indicates the generated questions are very relevant to the gold questions. We conduct an experiment to verify this fact. Specifically, we randomly select 1000 question pairs from Quora Question Pairs[6], with each pair annotated as semantically relevant. We then calculate the average simCSE score among these question pairs, and the result is 70.33%. In fact, the ideal model should excel in diversity metrics while maintaining balanced relevance scores. Our goal is to enhance diversity while ensuring relevance remains within an acceptable range.

Model	WQ		
	Diverse@3	Diverse@5	Diverse@10
Ours	**25.17**	**28.05**	**31.33**
Ours (w/o ss_f)	24.33	27.09	30.59
Ours (w/o ss_b)	24.13	26.81	30.09

Table 4: Ablation studies of two selection strategies (%).

4.3. Ablation Studies

4.3.1. Effect of Selection Strategy on f_θ

To demonstrate the effectiveness of our proposed selection strategy in the forward model f_θ, we remove it (*i.e.*, w/o ss_f) and use all pseudo pairs to train the forward model. Table 4 reports the results in diversity. We observe that removing the strategy results in relative declines of 3.34% in $Diverse@3$, 3.42% in $Diverse@5$, and 2.36% in $Diverse@10$,

[6] http://qim.fs.quoracdn.net/quora_duplicate_questions.tsv

respectively. These results indicate that pseudo pairs dissimilar to the training data can be noisy and hurt the performance of the forward model.

4.3.2. Effect of Selection Strategy on b_φ

We explore whether reliable pseudo pairs in training the backward model b_φ can improve the performance. We delete the selection strategy for choosing relevant and diverse pseudo pairs (*i.e.*, w/o ss_b) but utilize all pseudo pairs to fine-tune the backward model. According to the results shown in Table 4, we observe that removing the selection strategy leads to relative declines of 4.13% in $Diverse@3$, 4.42% in $Diverse@5$, and 3.96% in $Diverse@10$, respectively. These results indicate that reliable pseudo pairs can improve the performance of the forward model.

4.4. Positive Impact on QA Tasks

We conduct experiments on WebQSP (tau Yih et al., 2016), a widely-adapted KBQA dataset with 2,848 (question, answer) training instances, to evaluate two typical KBQA models, GRAFT-Net (Sun et al., 2018) and NSM (He et al., 2021). Since 1,409 (question, answer) pairs in the training data of WebQSP coincide with those in WQ, we can extract their corresponding subgraphs from WQ. With these subgraphs, KBQG models can produce their corresponding questions. For evaluation, we augment WebQSP with questions generated by B+P and our proposed model and denote the new dataset as "Augment by B+P" and "Augment by Ours" respectively. On these augmented datasets, we train GRAFT-Net and NSM and compare their performance with the same models trained on the original WebQSP (*i.e.*, Real).

From the results presented in Table 5, we conclude: **(1) The generated (question, answer) pairs can be viewed as a method of data augmentation for KBQA**, as both GRAFT-Net and NSM trained on datasets augmented by different KBQG models (*i.e.*, B+P and Ours) can enhance their QA performance when trained on the original dataset. **(2) Our model generates questions that significantly outperform those generated by B+P,** because KBQA models that are trained on the dataset augmented by ours outperform B+P.

4.5. Human Evaluation

We randomly choose 50 instances $\mathcal{S}_{50} = \{(G_i, q_i)\}_{i=1}^{50}$ from the test set of the WQ dataset and then evaluate whether top-3 and top-5 generated questions for each instance have different surface forms while ensuring their relevance to the ground truth.

Model	GRAFT-Net		NSM	
	Hits@1	F1	Hits@1	F1
Real	0.677	0.616	0.724	0.663
Augment by B+P	0.676	0.624	0.732	0.673
Augment by Ours	**0.688**	**0.629**	**0.739**	**0.681**

Table 5: QA performance on the augmented QA dataset.

Model	Top-3 Questions		Top-5 Questions	
	Diversity	Relevance	Diversity	Relevance
BART	3.45	**4.25**	3.56	**4.18**
B+P	3.67	4.05	3.85	4.02
Ours	**3.98**	3.96	**4.21**	3.89
Pearson	0.935	-	0.949	-

Table 6: Human evaluation results on WQ.

Concretely, we first invite three graduate students to score the relevance between generated questions and the ground truth. Then the three students judge the diversity of the generated questions for each instance and average the diversity. Finally, we average the scores of three students for our approach and two PLMs-based baselines BART and B+P. The diversity and relevance are scored on a five-point Likert scale, where 1-point indicates poor diversity and relevance, and 5-point represents perfect diversity and relevance. Table 6 reports the results, which shows that our approach can produce more diverse questions than other baselines while achieving respectable performance in terms of relevance with the baselines. Additionally, we use the Pearson correlation coefficient to evaluate the correlation between $Diverse@k$ and human evaluation. Table 6 reports the result of the Pearson correlation. We observe that our devised metric $Diverse@k$ is highly consistent with human evaluation, which demonstrates its rationality.

5. Related Work

5.1. Diversifying Question Generation.

Recently, PLMs-based methods (Chen et al., 2020; Guo et al., 2024, 2022; Ke et al., 2021; Xiong et al., 2022; Zhang et al., 2022) have been increasingly applied to automatically generate questions. Despite the success of PLMs-based models on KBQG, they concentrate on improving the quality of a single generated question but lack diversity. It is well known that diversity can make the generated question look more natural and human-like. Currently, diversifying text generation has attracted the interest of researchers and can be broadly categorized into two approaches: model enhancement and data augmentation. The former mainly concentrates on modifying the model architecture or revising loss functions. For example, Elangovan et al. (2023) utilize self-attention-based keyword selection to pro-

duce headlines that are diverse yet semantically consistent. Wang et al. (2020) use a continuous latent variable to model the content selection process and explicitly model question types using Conditional Variational Auto-encoder (CVAE) to diversify question generation. Shao et al. (2021) further inject a control algorithm into CVAE to balance the diversity and accuracy of the generated question. Zhang and Zhu (2021) design two loss functions to estimate the distribution of keywords in questions, and generate the question based on them. However, these methods do not apply to our task due to the differences in their settings compared to ours. In addition, some researchers also explore data augmentation-based methods. For instance, Su et al. (2020) introduce non-conversational text to diversify dialogue generation. Jia et al. (2020) create new (source, target) pairs by a simple back translation method to generate human-like questions. Our proposed method can be viewed as a data augmentation approach that leverages external natural questions to enhance diversity. Although Su et al. (2020) have also investigated the effectiveness of introducing external data in dialogue generation, our definition of diversity differs from theirs. They consider diversity as the distinction among all the instances as a whole, whereas we focus on the diversity of top-k generated results of each instance while ensuring their relevance to the ground truth. Furthermore, we carefully devise two simple and effective reliable pseudo pairs selection strategies on top of the dual model framework.

5.2. Diversity Evaluation Metrics.

For diversifying text generation tasks, diversity evaluation is a core step. Early studies have proposed some popular diversity evaluation metrics, such as Distinct-n (Li et al., 2016) and Self-BLEU (Zhu et al., 2018). For Distinct-n, it is a widely-used metric in various generation tasks, such as text generation (Shao et al., 2021; Jia et al., 2020) and story generation (Guan et al., 2021). To be concrete, Distinct-n is calculated as the number of distinct tokens divided by the total number of tokens, which makes it more like a measure of the duplication of n-grams rather than diversity. For Self-BLEU, it first computes the BLEU (Papineni et al., 2002b) score of all instance pairs and then takes their average. It is worth noting the smaller the Self-BLEU score, the more diverse it is. Obviously, Distinct-n and self-BLEU are inappropriate to measure the diversity explored in this paper since the two metrics focus on the ratio of distinct n-grams. Meanwhile, the two diversity metrics ignore semantic relevance to the ground truth, which is the key to assessing diversity. In view of this, we propose a novel diversity metric called $Diverse@k$, which measures the diversity among multiple generated questions for each instance while ensuring their relevance.

6. Conclusion

This work conducts pilot studies on diversifying KBQG. We rethink the diversity of questions and suppose that diversity should be that questions expressing the same semantics have different forms of expression. In light of this, we design a novel diversity evaluation metric $Diverse@k$ that measures the diversity among the top-k generated questions for each instance while guaranteeing relevance to the ground truth. Furthermore, we propose a dual framework with two simple yet effective selection strategies to generate diverse questions leveraging external natural questions. Experimental results demonstrate the superiority of our method.

7. Limitations

In our approach, we introduce external natural questions to diversify question generation, which can generate questions with different expressions, since these natural questions cover a wider range of semantic patterns and expressions. However, for instances with simple expressions, the paraphrasing-based method may achieve better performance. For example, the ground truth is "*What religion in Australia that influenced Arthur Schopenhauer?*", the paraphrasing-based approach generates "*What **faith** in Australia **inspired** Arthur Schopenhauer?*". Our method generates "***What is the religion** in Australia that influenced Arthur Schopenhauer?* ". We observe that the paraphrasing-based approach rewrites "**religion**" to "**faith**" and rewrites "**influenced**" to "**inspired**", but our method only rewrites "**What religion**" to "**What is the religion**", because the paraphrasing-based method focuses on words while ours focuses more on the structure of the sentences. Therefore, when the sentence's expression is not so diverse, the paraphrasing-based method may be well suited. We could study how to improve both word-level and structure-level diversity in the future.

8. Acknowledgments

This work is supported by National Key Research & Develop Plan (2023YFF0725100) and the National Natural Science Foundation of China (62322214, U23A20299, 62076245, 62072460, 62172424, 62276270). This work is supported by Public Computing Cloud, Renmin University of China. We deeply appreciate the insightful feedback provided by all reviewers.

9. Bibliographical References

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 8635–8648.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2023. Toward subgraph-guided knowledge graph question generation with graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12.

Venkatesh Elangovan, Kaushal Maurya, Deepak Kumar, and Maunendra Sankar Desarkar. 2023. Divhsk: Diverse headline generation using self-attention based keyword selection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1879–1891.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6894–6910.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 6394–6407.

Shasha Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. 2024. A survey on neural question generation: Methods, applications, and prospects. *arXiv preprint arXiv:2402.18267*.

Shasha Guo, Jing Zhang, Yanling Wang, Qianyi Zhang, Cuiping Li, and Hong Chen. 2022. Dsm: Question generation over knowledge base via modeling diverse subgraphs with meta-learner. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 4194–4207.

Gaole He, Yunshi Lan, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Improving multi-hop knowledge base question answering by learning intermediate supervision signals. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM 21*, pages 553–561.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2022. DEEP: denoising entity pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 1753–1766.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2020. How to ask good questions? try to leverage paraphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 6130–6140.

Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 2526–2538.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2016*, pages 110–119.

Chin-Yew Lin and FJ Och. 2004. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir workshop*, pages 1–8.

Yanxiang Ling, Fei Cai, Honghui Chen, and Maarten de Rijke. 2020. Leveraging context for neural question generation in open-domain dialogue systems. In *Proceedings of the Web Conference 2020, WWW 2020*, pages 2486–2492.

Tianqiao Liu, Qiang Fang, Wenbiao Ding, Hang Li, Zhongqin Wu, and Zitao Liu. 2021. Mathematical word problem generation from commonsense knowledge graph and equations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 4225–4240.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th Conference on Neural Information Processing Systems, NeurIPS 2022*, pages 27730–27744.

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

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

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Huajie Shao, Jun Wang, Haohong Lin, Xuezhou Zhang, Aston Zhang, Heng Ji, and Tarek F. Abdelzaher. 2021. Controllable and diverse text generation in e-commerce. In *Proceedings of the Web Conference 2021, WWW 2021*, pages 2392–2401.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 7087–7097.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W. Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4231–4242.

Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. 2021a. Diversifying dialog generation via adaptive label smoothing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 3507–3520.

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020. Diversify question generation with continuous content selectors and question type modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2134–2143.

Zichao Wang, Andrew S. Lan, and Richard G. Baraniuk. 2021b. Math word problem generation with mathematical consistency and problem context constraints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 5986–5999.

Guanming Xiong, Junwei Bao, Wen Zhao, Youzheng Wu, and Xiaodong He. 2022. Autoqgs: Auto-prompt for low-resource knowledge-based question generation from SPARQL. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM 2022*, pages 2250–2259.

Jie Zeng and Yukiko I Nakano. 2020. Exploiting a large-scale knowledge graph for question generation in food preference interview systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, pages 53–54.

Kun Zhang, Yunqi Qiu, Yuanzhuo Wang, Long Bai, Wei Li, Xuhui Jiang, Huawei Shen, and Xueqi Cheng. 2022. Meta-cqg: A meta-learning framework for complex question generation over knowledge bases. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*, pages 6105–6114.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, pages 1–43.

Zhiling Zhang and Kenny Q. Zhu. 2021. Diverse and specific clarification question generation with keywords. In *Proceedings of the Web Conference 2021, WWW 21*, pages 3501–3511.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2021. Learning from perturbations: Diverse and informative dialogue generation with inverse adver-

sarial training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 694–703.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020*, pages 1–18.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, pages 1097–1100.

10. Language Resource References

Alon Talmor and Jonathan Berant. 2018. *ComplexWebQuestions v.1*. Association for Computational Linguistics. PID https://www.tau-nlp.sites.tau.ac.il/compwebq.

Wen-tau Yih and Matthew Richardson and Christopher Meek and Ming-Wei Chang and Jina Suh. 2016. *WebQuestionsSP v.1*. Association for Computational Linguistics. PID https://www.microsoft.com/en-us/download/details.aspx?id=52763.

Mantong Zhou and Minlie Huang and Xiaoyan Zhu. 2018. *PathQuestion v.1*. Association for Computational Linguistics. PID https://github.com/zmtkeke/IRN.

A. Appendix

A.1. Case Study

We present top-3 questions for five instances generated by BART, B+P, and Ours on WQ in Table 7. Concretely, each approach returns top-3 generated questions, where the various surface forms for each instance are marked with different colors. We observe that the top-3 questions generated by ours are marked with three colors, but BART and B+P are mainly marked with two colors and a few with three colors. Based on the results, we conclude that top-3 questions generated by our model are more diverse than the baselines (*i.e.*, BART and B+P) because our approach introduces various external natural questions that cover a much broader range of semantic patterns and expressions so that it can benefit the diversifying question generation.

Task Description

Please generate a detailed and specific complex question using the information in the triples related to the answer. The question should include all relevant details from the triples.

Examples

Triples1: [Triples1] \n **Answer1:** [Answer1] \n
Question1: [Question1] \n

• • •

Triples5: [Triples5] \n **Answer5:** [Answer5] \n
Question5: [Question5] \n

Test Input

Triples: [Triples] \n **Answer:** [Answer] \n
Question:

Figure 3: The prompt for Large Language Models.

A.2. Prompt for Large Language Models

In our paper, we employ two advanced large language models (LLMs) as baselines, namely ChatGPT and text-davinci-003 (abbreviated as Davinci003). Our prompt design incorporates three key elements: the task description, illustrative examples, and the test input. As shown in Figure 3, the task description meticulously outlines the specifics of the task. For each test instance, we choose five representative examples to guide the model's generation.

Ground Truth	BART	B+P	Ours
Who is the coach of the team owned by Steve Bisciotti?	**Q1:** Who is the coach of the team owned by Steve Bisciotti? **Q2:** Team owner Steve Bisciotti 's sports team is coached by whom? **Q3:** Team owner Steve Bisciotti 's sports team is coached by whom?	**Q1:** Who is the head coach of the team owned by Steve Bisciotti? **Q2:** Who is the head coach of the team owned by Steve Bisciotti? **Q3:** Team owner Steve Bisciotti 's sports team is coached by whom?	**Q1:** Who is the coach of the team owned by Steve Bisciotti? **Q2:** Who is the current head coach of the team owned by Steve Bisciotti? **Q3:** Team owner Steve Bisciotti 's sports team is coached by whom?
People from the country with the national anthem Lofsöngur speak what language?	**Q1:** Which language is spoken in the country that has national anthem lofsöngur? **Q2:** What language is spoken in the country that has national anthem lofsöngur? **Q3:** What language is used in the country with national anthem lofsöngur?	**Q1:** Which language is spoken in the country with the national anthem lofsöngur? **Q2:** What spoken language is used in the country with national anthem lofsöngur? **Q3:** What spoken language was used in the country with national anthem lofsöngur?	**Q1:** What spoken language is used in the country with national anthem lofsöngur? **Q2:** People from the country that has the national anthem lofsöngur speak what language? **Q3:** Which language is spoken in the country with the national anthem lofsöngur?
What Canadian religion has a religious belief named Mahdi?	**Q1:** What religion with religious belief named Mahdi is recognized in Canada? **Q2:** What religion with religious belief named Mahdi is recognized in Canada? **Q3:** Which religion with religious belief in Mahdi is recognized in Canada?	**Q1:** In Canada, what faith Mahdi is recognised? **Q2:** In Canada, what faith Mahdi is recognized? **Q3:** What faith Mahdi is recognized in Canada?	**Q1:** What religion with religious belief Mahdi is recognized in Canada? **Q2:** Which of the major religions of Canada believes in Mahdi? **Q3:** What religion with religious belief Mahdi is in Canada?
What team with a mascot named K. C. Wolf did Warren Moon play for?	**Q1:** What team with a mascot named K. C. Wolf did Warren Moon play for? **Q2:** What team with a mascot named K. C. Wolf did Warren Moon play for in 2012? **Q3:** Which team with a mascot named K. C. Wolf did Warren Moon play for in 2012?	**Q1:** What team with a mascot named K. C. Wolf did Warren Moon play for? **Q2:** What team with a K. C. Wolf mascot did Warren Moon play for? **Q3:** Who did Warren Moon play for that has a mascot named K. C. Wolf?	**Q1:** In what team with a mascot named K. C. Wolf did Warren Moon play? **Q2:** Which team with a K. C. Wolf as a mascot did Warren Moon play for? **Q3:** Which team with a K. C. Wolf mascot did Warren Moon play for?
What stop motion film featured Miley Cyrus?	**Q1:** What stop motion movies has Miley Cyrus been in? **Q2:** What stop motion movies stared Miley Cyrus? **Q3:** Which stop motion movies stared Miley Cyrus?	**Q1:** What stop motion movies did Miley Cyrus appear in? **Q2:** What stop motion movies did Miley Cyrus play in? **Q3:** What stop motion film starred Miley Cyrus?	**Q1:** What movies that were filmed in stop motion was Miley Cyrus in? **Q2:** What movie featured Miley Cyrus and was filmed in stop motion? **Q3:** What stop motion movies did Miley Cyrus play?

Table 7: Comparison of top-3 generated questions on WQ, where the various surface forms are marked in different colors.