A Survey on Neural Question Generation: Methods, Applications, and Prospects

Shasha Guo[1*], **Lizi Liao**[2], **Cuiping Li**[1], **Tat-Seng Chua**[3]
[1]Renmin University of China
[2]Singapore Management University
[3]National University of Singapore
{guoshashaxing, licuiping}@ruc.edu.cn, lzliao@smu.edu.sg, dcscts@nus.edu.sg

Abstract

In this survey, we present a detailed examination of the advancements in Neural Question Generation (NQG), a field leveraging neural network techniques to generate relevant questions from diverse inputs like knowledge bases, texts, and images. The survey begins with an overview of NQG's background, encompassing the task's problem formulation, prevalent benchmark datasets, established evaluation metrics, and notable applications. It then methodically classifies NQG approaches into three predominant categories: **structured NQG**, which utilizes organized data sources, **unstructured NQG**, focusing on more loosely structured inputs like texts or visual content, and **hybrid NQG**, drawing on diverse input modalities. This classification is followed by an in-depth analysis of the distinct neural network models tailored for each category, discussing their inherent strengths and potential limitations. The survey culminates with a forward-looking perspective on the trajectory of NQG, identifying emergent research trends and prospective developmental paths. Accompanying this survey is a curated collection of related research papers, datasets and codes[1], providing an extensive reference for those delving into NQG.

1 Introduction

Question Generation (QG) represents a crucial and complex task within the domain of natural language processing (NLP). Its objective is to automatically generate questions from various sources such as knowledge bases [Kumar *et al.*, 2019a; Xiong *et al.*, 2022; Liang *et al.*, 2023], natural language texts [Tuan *et al.*, 2020; Pan *et al.*, 2020; Pan *et al.*, 2021a], and images [Chen *et al.*, 2021; Xie *et al.*, 2021; Chen *et al.*, 2023a]. The task has garnered substantial interest in the research community, attributable to its wide-ranging applications. Notably, QG serves as a means for data augmentation, enhancing the corpus of training data for question-answering (QA) tasks, thereby refining QA models [Chen *et al.*, 2023b; Guo *et al.*, 2022]. Additionally, it plays a vital role in intelligent tutoring systems by generating diverse questions from educational materials, aiding in evaluating and fostering student's learning [Zhao *et al.*, 2022; Gonzalez *et al.*, 2023]. Furthermore, QG contributes significantly to conversational systems, enabling them to initiate more engaging and dynamic human-machine interactions [Saeidi *et al.*, 2018; Ling *et al.*, 2020]. In the realm of fact verification, QG is pivotal in creating training claims to augment the effectiveness of verification models [Pan *et al.*, 2021b; Zhang and Gao, 2023].

The ascendance of deep neural networks [Vaswani *et al.*, 2017; Shen *et al.*, 2018] has prompted a paradigm shift in QG methodologies. The field has progressively transitioned from rule-based approaches to neural network-based (NN-based) methods [Bi *et al.*, 2020; Pan *et al.*, 2021a; Chen *et al.*, 2023b]. Predominantly, these NN-based approaches follow the Sequence-to-Sequence (Seq2Seq) framework, utilizing various encoder-decoder architectures to refine question generation. However, a critical limitation of these models is their reliance on extensive training data, a challenge exacerbated by the typically small size of benchmark datasets in QG, leading to potential overfitting issues.

The emergence of pre-trained language models (PLMs), such as T5 [Raffel *et al.*, 2020] and BART [Lewis *et al.*, 2020], represents a significant advancement. These models, pre-trained on extensive corpora, possess a wealth of semantic knowledge, which significantly enhances performance in various NLP tasks upon fine-tuning. Hence, PLMs effectively address the challenge faced by previous NN-based models in QG, obviating the need for training models from scratch. This development has established the pre-training-fine-tuning framework as the dominant paradigm in QG, achieving unprecedented state-of-the-art (SOTA) results.

With the continuous scaling of PLMs in terms of parameter size and training corpus volume, the field has witnessed the evolution of large language models (LLMs) such as ChatGPT [2] and Llama2 [3]. These models surpass PLMs in semantic richness, offering remarkable improvements across a wide array of NLP tasks [Liu *et al.*, 2023; Nan *et al.*, 2023]. Consequently, the research focus has shifted towards leveraging

[*] Work was done during an internship at SMU.

[1]https://github.com/PersistenceForever/Neural-Question-Generation-Survey-List

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

[3]https://ai.meta.com/llama/

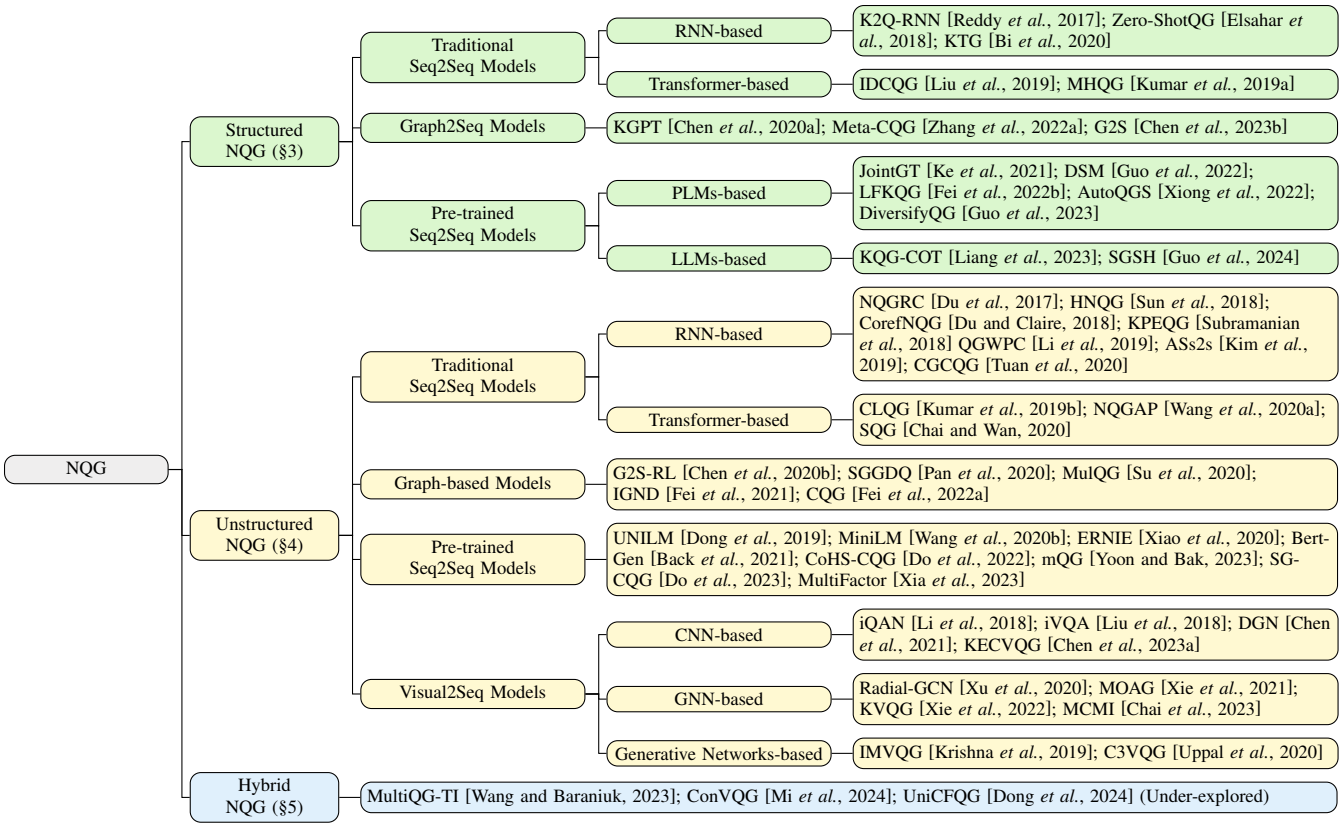

Figure 1: The taxonomy of NQG. We classify NQG into three types based on input modalities: Structured NQG, which deals with structured data; Unstructured NQG, which handles unstructured data; and Hybrid NQG, which integrates both structured and unstructured data.

LLMs for QG, aiming to capitalize on their advanced semantic understanding [Liang *et al.*, 2023]. A pivotal aspect in this context is In-Context Learning (ICL), a unique capability of LLMs, which can be effectively harnessed through well-designed prompts to generate the desired questions.

Existing surveys on QG primarily concentrate on traditional Seq2Seq models and on models that generate questions from text. For instance, Pan *et al.* [2019b] mainly review traditional Seq2Seq QG models predating 2019, while Zhang *et al.* [2022b] offer an expansive overview of both traditional and pre-trained Seq2Seq models for question generation from text. To the best of our knowledge, our survey is the first to provide a comprehensive review of NQG across various input modalities, including knowledge bases, texts and images.

Moving forward, we first outline essential background settings for NQG in Section 2. We then introduce a new ontology for NQG, delving into structured, unstructured, and hybrid NQG in Sections 3, 4 and 5, respectively. Furthermore, we present a thorough prospect on several promising research directions for future studies on NQG in Section 6.

2 Background

2.1 Problem Formulation

The NQG task aims to automatically generate textual questions from diverse input modalities, such as knowledge bases, texts, and images, which we denote as $\mathcal{X}$. Given an input $\mathcal{X}$, and optionally a specific target answer A, the objective of the NQG task is to learn a mapping function f_θ to generate a textual question Q. This is achieved by optimizing the model parameter θ to maximize the conditional likelihood $P_\theta\left(Q|\mathcal{X}, A\right)$. Formally, the NQG task can be described as:

$$f_\theta : (\mathcal{X}, A) \rightarrow Q, \tag{1}$$

where f_θ generates a textual question $Q =< q_1, q_2, ..., q_n >$ comprising of a sequence of word tokens q_i. Each token q_i is selected from a predefined vocabulary $\mathcal{V}$. In this survey, the model f_θ is realized in various neural network architectures, encompassing Recurrent Neural Networks (RNN), Transformer, PLMs, or even LLMs.

2.2 Popular Datasets

We summarize popular datasets in NQG tasks, as shown in Table 1. Since QG can be viewed as a dual task of QA, QG datasets are typically derived from QA datasets. To enhance clarity, we classify these datasets based on their input types, encompassing those derived from knowledge bases, natural language text, and visual sources.

Knowledge Base-based Datasets. We show three popular datasets for knowledge base question generation (KBQG). Firstly, WebQuestions (WQ) [Kumar *et al.*, 2019a] comprises 22,989 instances from WebQuestionsSP [Yih *et al.*, 2016] and ComplexWebQuestions [Talmor and Berant, 2018], both of

Problem	Dataset	Question Types	#Questions	#Ent. / #Doc. / #Ima.	#Avg.Tri. / #Avg.Que.
KBQG	WQ [Kumar *et al.*, 2019a]	Multi-hop	22,989	25,703	5.8
	PQ [Zhou *et al.*, 2018]	Multi-hop	9,731	7,250	2.7
	GQ [Gu *et al.*, 2021]	Multi-hop	64,331	32,585	1.4
TQG	SQuAD [Rajpurkar *et al.*, 2016]	Factoid	97,888	20,958	4.67
	MS MARCO [Nguyen *et al.*, 2016]	Factoid	3,563,535	1,010,916	3.53
	NewsQA [Trischler *et al.*, 2017]	Factoid	119,633	12,744	9.39
	HotpotQA [Yang *et al.*, 2018]	Multi-hop	112,779	5,000	8
	CoQA [Reddy *et al.*, 2019]	Conversational	127,000	8,000	10
VQG	VQA [Antol *et al.*, 2015]	Factoid	369,861	204,721	3
	VQG [Mostafazadeh *et al.*, 2016]	Commonsense	25,000	5,000	5

Table 1: Summary of popular datasets for neural question generation. #Questions represents the total number of questions. #Ent., #Doc., and #Ima. denote the total number of entities, documents, and images in KBQG, TQG, and VQG, respectively. #Avg.Tri. is the average number of triples in each question for KBQG. #Avg.Que. denotes the average number of questions per document in TQG or per image in VQG.

which are benchmarks for knowledge base question answering (KBQA). These benchmarks contain questions, answers, and corresponding SPARQL queries. Following [Kumar *et al.*, 2019a], they convert a SPARQL query to a subgraph. Consequently, each instance in the WQ dataset includes subgraphs, answers, and questions. In addition, PathQuestions (PQ) [Zhou *et al.*, 2018] is constructed using two subsets of Freebase. Notably, in PQ, the KB subgraph forms a path between the topic entities and answer entities, typically spanning connections of 2-hop or 3-hop. Furthermore, GrailQA (GQ) [Gu *et al.*, 2021] is a large-scale, high-quality KBQA dataset. Each question is associated with an S-expression, which can be interpreted as a logical form.

Text-based Datasets. We introduce five classical benchmark datasets for text-based question generation (TQG). Firstly, Stanford Question Answering Dataset (SQuAD) [Rajpurkar *et al.*, 2016] is a typical reading comprehension dataset, consisting of QA pairs. These pairs are created by crowd workers using Wikipedia articles. The answers to these questions are text segments extracted from the corresponding reading passages within these articles. Secondly, MicroSoft MAchine Reading COmprehension (MS MARCO) [Nguyen *et al.*, 2016] is a comprehensive real-world dataset for reading comprehension. Each question receives a response from a crowdsourced worker, ensuring that each answer is human-generated. Thirdly, NewsQA [Trischler *et al.*, 2017] is a challenging machine comprehension dataset. Crowd workers provide questions and their corresponding answers based on news articles. The answers are composed of specific text spans extracted directly from the related news articles. Fourthly, HotpotQA [Yang *et al.*, 2018], a multi-hop QA dataset, consists of QA pairs sourced from Wikipedia. To create these pairs, crowd workers are presented with a variety of contextual supporting documents. They are instructed to formulate questions that require reasoning across these documents. Following this, they answer the questions by identifying and extracting pertinent text spans from the given context. Lastly, CoQA [Reddy *et al.*, 2019], a large-scale conversational QA dataset, contains 127,000 QA pairs derived from 8,000 conversations. These pairs are based on text passages spanning seven diverse domains.

Visual-based Datasets. We present two widely used datasets for visual question generation (VQG). Initially, VQA [Antol *et al.*, 2015], a classical VQG benchmark, consists of images along with corresponding questions and answers. Notably, due to the unavailability of answers for the VQA test set, the validation set is commonly utilized as a proxy for test set evaluation. Subsequently, VQG COCO [Mostafazadeh *et al.*, 2016] showcases naturally formulated and engaging questions that are based on common sense reasoning. These human-annotated questions originate from the Microsoft common objects in context dataset.

2.3 Evaluation

Given the inherent complexity and diversity in human evaluation, we mainly focus on automatic evaluation for NQG. We present automatic metrics across three categories, including n-grams-based, diversity, and semantic similarity metrics.

N-gram-based Metrics. We showcase three classical evaluation metrics that assess the n-gram similarity between the ground-truth and the generated questions.

- **BLEU.** BiLingual Evaluation Understudy (BLEU) [Papineni *et al.*, 2002] metric evaluates the average n-gram precision against the reference text, applying a penalty for excessively short text. BLEU-n calculates the proportion of the common n-grams between the generated question and the ground-truth.
- **ROUGE.** Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [Lin, 2004] focuses on recall, measuring the ratio of n-grams from the ground-truth question that are also present in the generated question.
- **METEOR.** Metric for Evaluation of Translation with Explicit ORdering (METEOR) [Banerjee and Lavie, 2005] offers a more comprehensive assessment than BLEU. METEOR is calculated based on the harmonic mean of the unigram precision and recall, providing a balanced evaluation.

Diversity Metrics. The diversity metric is essential for tasks involving diversified question generation. A widely recognized metric, **Distinct-**n [Li *et al.*, 2016], calculates the proportion of unique n-grams in the generated question.

Semantic Similarity Metrics. Beyond word-level comparison, it is vital to assess sentence-level comparison, particularly about semantic similarity. **BERTScore** [Zhang *et al.*,

2020], a widely-used metric, utilizes pre-trained contextual embeddings from BERT to compare words between generated and ground-truth questions, calculating their similarity using cosine similarity.

2.4 Applications

NQG tasks have emerged as a versatile tool in various applications, each demonstrating its unique value and potential. We examine four key applications:

Question Answering. This application revolves around deriving answers from data sources like Wikipedia and knowledge bases in response to natural language questions. The effectiveness of QA models is largely dependent on the richness of available QA pairs. Manual labeling, a standard method for creating QA datasets, is both resource-intensive and time-consuming, which often limits the size of the datasets. QG, serving as a dual task to QA, significantly enhances the capabilities of QA systems by producing vital training data. For example, Guo *et al.* [2022] substitute the original questions in the WebQuestionSP dataset with questions generated by the proposed QG model, demonstrating that the questions produced by the QG model are quite close to the original questions. In practical terms, this means more efficient information retrieval from digital assistants and more robust responses in customer service chatbots.

Intelligent Tutoring. Personalized education is a rapidly growing field, and intelligent tutoring systems stand at its forefront. The ability to generate custom questions tailored to a student's learning material and progress is invaluable. QG facilitates this by creating diverse and level-appropriate questions, thereby offering a more dynamic and responsive learning experience. For instance, the Bull2Sum system, developed by Gonzalez *et al.* [2023] not only generates relevant questions but also contributes to a substantial educational dataset. In practical application, this means students receive more engaging, varied, and effective learning tools, leading to better understanding and retention of material.

Conversational Systems. In the realm of interactive technology, conversational systems such as virtual assistants and customer support chatbots rely heavily on QG to maintain engaging and relevant dialogues. QG enhances these systems' ability to ask contextually related and engaging questions, thereby significantly improving user interactions. The concept of conversational question generation (CQG) introduced by Pan *et al.*[2019a] underscores the importance of context-aware and history-informed questioning in making these interactions more natural and smooth.

Fact Verification. In an era of information overload, fact verification is essential, particularly in journalism, legal investigation, and content moderation on social media. The ability of QG to generate (evidence, claim) pairs marks a significant advancement, as it automates the creation of training data for fact-checking models. The approach presented by Pan *et al.*[2021b] showcases how QG can streamline the validation of claims against available evidence, thereby enhancing the efficiency and accuracy of fact-checking operations.

3 Structured Neural Question Generation

Structured NQG is designed to create pertinent questions based on structured data sources. Within these, the knowledge base stands out as the most typical data source. This section primarily focuses on *knowledge base question generation* (KBQG). KBQG generates questions based on a set of facts from a KB subgraph, where each fact is usually represented as a triple. As shown in Figure 1, we classify KBQG models into three categories based on their architectural design: *Traditional Seq2Seq models*, *Graph2Seq models*, and *Pre-trained Seq2Seq models*.

Traditional Seq2Seq Models

Previous studies predominantly follow the sequence-to-sequence (Seq2Seq) framework, wherein the linearized subgraph is initially inputted into an encoder to derive its representation, followed by employing a decoder to produce the question from this representation. We categorize Seq2Seq models into two types based on the encoder and decoder used, namely **RNN-based** and **Transformer-based**.

RNN-based. Serban *et al.* [2016] first use a recurrent neural network (RNN) with an attention mechanism to map the KB facts into corresponding natural language questions. Subsequently, Reddy *et al.* [2017] develop an RNN-based approach, K2Q-RNN, for generating questions from a specific set of keywords. Specifically, K2Q-RNN first utilizes a question keywords extractor to derive a set of keywords from entities in KB. It then employs an RNN encoder to transform these keywords into a representation. Finally, it decodes this representation to produce the output question sequence. To improve the generalization for KBQG, Elsahar *et al.* [2018] propose an encoder-decoder framework leveraging extra textual contexts of triples. Concretely, a feed forward architecture encodes the input triple and a set of RNN to encode textual context. A decoder is equipped with triple and textual context attention modules and a copy mechanism to generate questions. Despite these advances, two classical challenges remain: limited information and semantic drift. To solve these, Bi *et al.* [2020] design a novel model, KTG, which integrates a knowledge-augmented fact encoder and a typed decoder within a reinforcement learning framework, enhanced by a grammar-guided evaluator. The encoder processes entities, relations, and auxiliary knowledge to create an augmented fact representation used by the decoder for question generation, while the evaluator assesses each question's grammatical similarity to the ground-truth, providing feedback that continually refines the encoder-decoder module.

Transformer-based. Considering the limitations of RNN in terms of effectiveness and efficiency in handling larger contexts, Transformer [Vaswani *et al.*, 2017] emerges as a robust and effective alternative, offering a promising solution. Kumar *et al.* [2019a] propose a novel model to generate complex multi-hop, difficulty-controllable questions from subgraphs. The encoder encodes the difficulty level of the given subgraph, subsequently enabling the decoder to generate questions that are tailored to this specific difficulty level. To effectively generate questions that not only articulate the specified predicate but also correlate with a definitive an-

swer, Liu *et al.* [2019] utilize a range of diverse contexts and design an answer-aware loss. A context-augmented fact encoder, equipped with multi-level copy mechanisms, effectively captures the diversified information across contexts and fact triples, thereby mitigating the issue of inaccurate predicate expression. Furthermore, the use of an answer-aware loss function ensures that the generated questions align with definitive answers by applying cross-entropy between the words in the question and those that denote the answer type.

Graph2Seq Models

The intricate structural information within a KB is crucial for generating high-quality questions in KBQG. However, previous approaches fall short in effectively capturing this rich structural information, as they merely linearize a KB subgraph into a sequence of triples and use RNN / Transformer models to learn its embeddings. Motivated by this, Chen *et al.* [2023b] propose a novel Graph-to-Sequence (Graph2Seq) model, which first utilizes a bidirectional gated graph neural network-based (BiGGNN-based) encoder to encode the KB subgraph, followed by decoding the output question using an RNN-based decoder, equipped with a node-level copy mechanism. Zhang *et al.* [2022a] propose a meta-learning framework, Meta-CQG, to address the data imbalance issue. They initially utilize graph-level contrastive learning to train a graph retriever, followed by retrieving similar subgraphs using cosine similarity across graph embeddings. Subsequently, they employ a meta-learning approach to train a generator tailored to each input subgraph by learning the potential features of retrieved similar subgraphs.

Pre-trained Seq2Seq Models

Pre-trained language models (PLMs), pre-trained on a large-scale corpus, possess rich semantic knowledge, which can boost the performance of downstream KBQG tasks through fine-tuning. With the continuous expansion in parameter size and training corpus volume, the field has witnessed the evolution of large language models (LLMs). In light of this, we categorize pre-trained Seq2Seq models for KBQG into two groups: those based on traditional PLMs (**PLMs-based**) and those utilizing the more advanced LLMs (**LLMs-based**).

PLMs-based. BART [Lewis *et al.*, 2020] and T5 [Raffel *et al.*, 2020], two widely recognized PLMs, are built upon the encoder-decoder framework. A significant challenge in adapting these PLMs to KBQG tasks lies in bridging the semantic gap. This is primarily because PLMs are originally pre-trained on unstructured text, which contrasts with the structured nature of KB. Additionally, another key challenge involves effectively capturing the structural information inherent in KB.

To address the above challenges, Ke *et al.* [2021] introduce three pre-training tasks including graph-enhanced text reconstruction, text-enhanced graph reconstruction, and graph-text embedding alignment to explicitly build the connection between knowledge graphs and text sequences. Additionally, they develop a structure-aware semantic aggregation module at each Transformer layer to aggregate contextual information following the graph structure. Guo *et al.* [2022] address the issue of the semantic gap by converting the input subgraph into a linear sequence of triples, achieved through the concatenation of relational triples. Furthermore, they explicitly incorporate structural information as input. This incorporation involves directly inserting special tokens "⟨H⟩", "⟨R⟩", and "⟨T⟩" before the head entity, relation, and tail entity in each fact triple, respectively. This approach effectively clarifies the relationships between entities, ensuring a more coherent representation within the model. Likewise, Guo *et al.* [2023] add special tokens indicative of structured information at the beginning of each element within every triple in KB. Xiong *et al.* [2022] focus on directly generating questions from SPARQL, aimed at covering complex operations. They first execute the SPARQL query on a KB to retrieve a corresponding subgraph. This subgraph is then linearized, serving as input for an auto-prompter, which generates the prompt text. Subsequently, this prompt text, along with the original SPARQL query, are used as inputs for a QG generator. This process ultimately results in high-quality question generation, effectively bridging the gap between non-natural language SPARQL and the natural language question.

LLMs-based. Despite the success of PLMs-based models on KBQG [Fei *et al.*, 2022b; Guo *et al.*, 2022; Guo *et al.*, 2023], their effectiveness hinges on extensive fine-tuning using large training datasets. However, the creation of labeled datasets is costly and time-consuming. Hence, researchers are increasingly focusing on few-shot KBQG tasks to mitigate these challenges [Xiong *et al.*, 2022]. Recently, LLMs, like ChatGPT[4] and Llama2 [5], have exhibited exceptional capabilities in various few-shot and zero-shot tasks. This emerging insight inspires researchers to investigate few-shot KBQG tasks, leveraging the capabilities of LLMs. The key challenge is to design effective prompts that prompt LLMs to generate the targeted questions for KBQG. Liang *et al.* [2023] propose KQG-COT framework, which involves first employing LLMs (i.e., text-davinci-003) to generate ideal questions for KBQG using chain-of-thought (COT). To be specific, KQG-COT first identifies suitable logical forms from the unlabeled data pool, meticulously evaluating their attributes. Following this, a specialized prompt is developed to steer LLMs in creating complex questions derived from these chosen logical forms. Guo *et al.* [2024] develop a fine-grained prompting approach named SGSH. Concretely, SGSH involves training a learnable skeleton generator which then uses the generated skeleton to create skeleton-based prompts, effectively stimulating LLMs to generate desired questions.

4 Unstructured Neural Question Generation

Unstructured NQG focuses on producing textual questions derived from unstructured data sources such as texts and visual images. Accordingly, our exploration will specifically focus on two distinct types: Text-based Question Generation (**TQG**) and Visual Question Generation (**VQG**).

4.1 TQG

As shown in Figure 1, TQG models are primarily divided into three types: *Traditional Seq2Seq models*, *Graph-based Mod-*

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

[5]https://ai.meta.com/llama/

els, and *Pre-trained Seq2Seq models*.

Traditional Seq2Seq Models

Most TQG models adhere to the Seq2Seq framework. This framework first employs an encoder to compress the input text into low-dimensional vectors that retain the essential semantic meanings. Subsequently, a decoder is employed to generate questions based on these condensed vectors. We divide TQG models into RNN-based and Transformer-based models according to their backbone architecture.

RNN-based. Du *et al.* [2017] first apply RNN to TQG tasks, leveraging an attention mechanism to enable the decoder to concentrate on the most pertinent segments of the input text. To decide which information to focus on when generating questions, most Seq2Seq models leverage the answer position features to incorporate the answer spans. For example, Sun *et al.* [2018] contend that context words near the answer are more apt to be answer-relevant. Hence, they explicitly encode the positional proximity of these context words to the answer by position embedding and a position-aware attention mechanism. Nevertheless, Li *et al.* [2019] think the proximity-based approach does not always work. Therefore, they devise a more generalized model, which exploits answer-relevant relations to facilitate the faithfulness of the generated question.

However, when dealing with long documents as the input context, these models face increased difficulty in effectively exploiting relevant content while avoiding irrelevant information. To solve this issue, Du and Claire [2018] suggest integrating coreference knowledge into the encoder to improve the model's ability to identify entities across different sentences, thereby enhancing the quality of question generation. Tuan *et al.* [2020] apply multi-stage attention to focus on crucial segments of the document that are pertinent to the answer, leveraging them to facilitate the generation of questions.

Transformer-based. Due to the inherent sequential nature of RNN, RNN-based models face significant computational costs and struggle with long-range dependency issues. Fortunately, the Transformer effectively addresses these challenges, resulting in the widespread adoption of Transformer-based models for TQG tasks. Kumar *et al.* [2019b] develop a cross-lingual model designed to enhance QG for a primary language by utilizing resources from a secondary language. Wang *et al.* [2020a] regard the answer as the hidden pivot for QG. Specifically, they first generate the hidden answer according to the paragraph. Subsequently, they merge this paragraph with the derived pivot answers to generate the question. Chai and Wan [2020] present a semi-autoregressive approach for generating sequential questions. Concretely, they segment the target questions into various groups, and then simultaneously generate each group of closely related questions.

Graph-based Models

Traditional Seq2Seq models struggle to capture the inherent structure of context, including syntax and semantic relationships. In contrast, graph neural networks (GNNs), with their inherent advantage in graph structure, are more adept at understanding and expressing the relationships between entities or sentences. Consequently, researchers are increasingly adopting GNN-based models for TQG tasks [Chen *et al.*, 2020b; Su *et al.*, 2020; Pan *et al.*, 2020; Fei *et al.*, 2021; Fei *et al.*, 2022a]. In general, most approaches initiate by constructing a graph from the input context, followed by utilizing a GNN to learn the graph representation from the constructed text graph effectively. Subsequently, this representation is fed into the decoder to produce the question. For instance, Chen *et al.* [2020b] first create two types of passage graph from the input text, *i.e.*, syntax-based static graph and semantics-aware dynamic graph. Following this, they introduce an innovative bidirectional gated graph neural network, designed to effectively learn the passage graph embeddings from the assembled text graph. Pan *et al.* [2020] focus on generating deep questions, wherein they initially extract key information from the passage to organize it as a semantic graph. Subsequently, they propose an attention-based gated graph neural network to capture the dependency relations of the semantic graph. Fei *et al.* [2021] propose a novel model, in which a relational-graph encoder is introduced for encoding dependency relations within passages, accompanied by an iterative GNN-based decoder. This decoder is specifically designed to capture structural information throughout each step of the generation process. Fei *et al.* [2022a] first construct an entity graph from the input documents, followed by utilizing a graph attention network to extract key entities. Furthermore, they introduce a controlled Transformer-based decoder, enhanced with a flag tag, to ensure the inclusion of these key entities in the generated questions.

Pre-trained Seq2Seq Models

Pre-trained language models, through pre-training on vast textual corpora, acquire an extensive range of linguistic knowledge, which can significantly enhance the performance of downstream tasks. While PLMs exhibit remarkable proficiency in processing natural language text, there still exist several challenges for TQG tasks. Primarily, PLMs are not specifically trained on TQG datasets, leading to their reduced proficiency in TQG tasks. Additionally, PLMs rely on generic self-supervised learning tasks, which are not tailored for TQG tasks, resulting in suboptimal performance in TQG tasks.

To address these challenges, researchers are increasingly focusing on fine-tuning PLMs for specific downstream tasks and adapting their neural architectures accordingly [Dong *et al.*, 2019; Wang *et al.*, 2020b; Xiao *et al.*, 2020; Do *et al.*, 2022; Xia *et al.*, 2023]. For example, Dong *et al.* [2019] present a unified pre-trained language model (UNILM), which is distinctively optimized across three distinct types of language modeling tasks, including unidirectional, bidirectional, and sequence-to-sequence prediction. Back *et al.* [2021] propose a novel pre-training approach tailored specifically for QG tasks. This approach intensively focuses on the answer, aiming to generate contextually relevant sentences containing missing answers. By doing this, it aims to learn more effective representations that are highly optimized for the question generation task. Do *et al.* [2023] present a new framework comprising two modules for generating conversational questions: "what-to-ask" and "how-to-ask". The "what-to-ask" module constructs a semantic graph to extract underlying rationale and selects the relevant answer span. The "how-to-ask" module uses a classifier to iden-

tify the appropriate question type. Subsequently, the framework fine-tunes the T5 [Raffel *et al.*, 2020] model on the tailored dataset to produce conversational questions. Xia *et al.* [2023] introduce phrase-enhanced Transformer, an effective model that capitalizes on the strengths of powerful PLMs. This method creatively integrates phrase selection probabilities from the encoder into the decoder, significantly enhancing the quality of question generation.

4.2 VQG

VQG can be regarded as a dual task of visual question answering. As illustrated in Figure 1, VQG models are typically divided into three main categories, including **CNN-based**, **GNN-based**, and **Generative Networks-based**.

CNN-based. Most VQG models typically employ a convolutional neural network (CNN) to encode an image and a RNN to encode an answer, both merging into an intermediate representation. This representation is then decoded to generate a question. For example, Liu *et al.* [2018] devise a multi-model attention module to dynamically identify regions in the image that are relevant to the answer. To generate difficulty-controllable questions, Chen *et al.* [2021] introduce a difficulty control mechanism in the decoder, utilizing a difficulty variable to regulate the complexity of the questions generated. Chen *et al.* [2023a] present a knowledge-enhanced causal visual question generation (KECVQG) model, which addresses the inherent bias in previous VQG models by employing a causal approach and knowledge integration to generate more accurate and unbiased questions from images.

GNN-based. One key challenge of VQG is to focus on answer-related regions during question generation. To solve the issue, researchers propose several approaches to perform explicit region selection. For instance, Xu *et al.* [2020] perform explicit object-level cross-modal interaction by identifying a core answer area and constructing an answer-related graph convolutional network (GCN) graph structure. Xie *et al.* [2021] leverage a co-attention network and a graph network to identify and relate key objects in an image to a target answer, thereby generating more comprehensive questions. However, previous approaches rely solely on semantic features to identify regions related to the answer, leading to potential biases and overlooking complex relations between objects. Given this, Chai *et al.* [2023] utilize contrastive learning to integrate semantic knowledge with regional representations and leverage a relation-level interaction scenario to consider multiple types of relations among regions and answers.

Generative Networks-based. To overcome the limitation of existing VQG models to produce generic and uninformative questions, Krishna *et al.* [2019] introduce a novel method that maximizes the mutual information between the image, the expected answer, and the generated question. This is achieved by employing a Variational Auto-Encoder (VAE) framework and utilizing two distinct latent spaces, enhancing the diversity and relevance of the questions generated. Observing that previous VQG models often rely heavily on answers, leading to overfitting and a lack of creativity. Uppal *et al.* [2020] propose a category-specific, cyclic training approach. This innovative method employs weak supervision and structured latent spaces, enabling the generation of diverse and relevant questions based on categories, thereby eliminating the need for ground-truth answers.

5 Hybrid Neural Question Generation

Hybrid NQG aims to generate textual questions based on both structured and unstructured data sources [Wang and Baraniuk, 2023; Mi *et al.*, 2024; Dong *et al.*, 2024], where multimodal is common. Compared with the previous single-modal question generation, hybrid question generation is more prevalent in real-life scenarios, especially in the education field. The primary challenge in hybrid question generation lies in effectively integrating information across diverse data sources or modalities. As pioneers in the field, Wang and Baraniuk [2023] first investigate multi-modal question generation from images and texts, proposing a novel and effective PLMs-based approach that surpasses the performance of ChatGPT. Dong *et al.* [2024] propose a unified framework for generating contextual questions (CQG) and factoid questions (FQG), which addresses current methods' limitations in structural and contextual information. Specifically, they introduce shared task modules for cross-domain learning and task-specific modules that integrate external knowledge for CQG and enhance contextual understanding for FQG, demonstrating advanced performance. Despite these promising results, hybrid NQG remains under-explored, with significant potential for further innovation and improvement.

6 Conclusion and Future Directions

This paper provides a thorough overview of Neural Question Generation (NQG) in various modalities. We first introduce popular datasets, classical evaluation metrics, and four prominent applications. We then explore prevalent methods for modeling diverse inputs, including structured NQG, unstructured NQG, and hybrid NQG. Despite the notable achievements of NQG models, several challenges remain, suggesting promising directions for future research.

Proactive Question Generation. Previous studies predominantly focus on producing reactive questions based on the provided inputs. However, the ability to proactively tailor question generation to meet specific user requirements and achieve pre-defined targets is essential in real-world applications. Intelligent tutoring systems serve as a prime example, engaging students with tailored interactions. These systems carefully design a series of exercises aimed at specific objectives, gradually guiding step by step from simple to advanced towards the targeted goal, in order to enhance students' understanding of specific concepts. Although the benefits and practical applications are evident, the field of proactive question generation remains under-explored. This underlines its significant potential as a promising field for future research.

Multi-modal Question Generation. Current question generation tasks primarily concentrate on single-modal question generation, such as KBQG, TQG, and VQG. Nevertheless, in practical scenarios, the significance of multi-modal question generation is on the rise. This is particularly evident in the educational field, where numerous scientific questions require

an understanding of both visual images and textual descriptions. To the best of our knowledge, multi-modal question generation remains in a very early stage, with few studies having been conducted in this field [Wang and Baraniuk, 2023], as detailed in Section 5. Given the advanced capabilities of vision-language pre-trained models like CLIP, a compelling research direction is to develop effective strategies for leveraging VL-PLMs in multi-modal question generation.

Controllable Question Generation. The capacity to control specific aspects of question generation holds significant applications, such as creating different difficulty levels [Kumar *et al.*, 2019a] and types of questions for intelligent tutoring systems (ITS). This potential paves the way for future research, particularly in exploring how these customized elements can enhance personalized learning experiences within ITS. Additionally, certain studies overlook subjective human factors such as sentiment and style, focusing instead on the influence of the input and the target answer. Yet, these human elements are critical in influencing the process of question generation. Thus, upcoming studies need to investigate approaches that correspond to distinct human behaviors and preferences.

Automatic Evaluation Metrics for Generation. Widely used metrics such as BLEU and ROUGE assess question quality by measuring the lexical overlap between the generated question and the ground-truth. Yet, these metrics can potentially penalize well-formed questions that diverge in lexical similarity from the ground-truth questions, indicating a limitation in capturing question validity. Accordingly, a more reasonable metric for assessing question quality would consider key factors such as question answerability [Mohammadshahi *et al.*, 2023], consistency with the context provided, and containing a sufficient amount of information content. Meanwhile, the advancement of diversity metrics is critical, particularly due to the significant diversity capabilities demonstrated by LLMs. The popular diversity metric, Distinct-n, emphasizes the ratio of unique n-grams but it is overly simplistic. Hence, a comprehensive assessment of diversity can be conducted from multiple perspectives, including semantic diversity [Guo *et al.*, 2023], syntactic diversity, and thematic diversity. This highlights the need for innovative metrics to accurately evaluate the diverse aspects of question quality.

Acknowledgments

This work is supported by the 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 gratefully acknowledge the support provided by the China Scholarship Council Scholarship Fund. We sincerely thank all reviewers for their valuable feedback.

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015.

[Back *et al.*, 2021] Seohyun Back, Akhil Kedia, Sai Chetan Chinthakindi, Haejun Lee, and Jaegul Choo. Learning to generate questions by learning to recover answer-containing sentences. In *ACL*, 2021.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, 2005.

[Bi *et al.*, 2020] Sheng Bi, Xiya Cheng, Yuan-Fang Li, Yongzhen Wang, and Guilin Qi. Knowledge-enriched, type-constrained and grammar-guided question generation over knowledge bases. In *COLING*, 2020.

[Chai and Wan, 2020] Zi Chai and Xiaojun Wan. Learning to ask more: Semi-autoregressive sequential question generation under dual-graph interaction. In *ACL*, 2020.

[Chai *et al.*, 2023] Zi Chai, Xiaojun Wan, Soyeon Caren Han, and Josiah Poon. Visual question generation under multi-granularity cross-modal interaction. In *MMM*, 2023.

[Chen *et al.*, 2020a] Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. Kgpt: Knowledge-grounded pre-training for data-to-text generation. In *EMNLP*, 2020.

[Chen *et al.*, 2020b] Yu Chen, Lingfei Wu, and Mohammed J Zaki. Reinforcement learning based graph-to-sequence model for natural question generation. In *ICLR*, 2020.

[Chen *et al.*, 2021] Feng Chen, Jiayuan Xie, Yi Cai, Tao Wang, and Qing Li. Difficulty-controllable visual question generation. In *APWeb-WAIM*, 2021.

[Chen *et al.*, 2023a] Jiali Chen, Zhenjun Guo, Jiayuan Xie, Yi Cai, and Qing Li. Deconfounded visual question generation with causal inference. In *ACM MM*, 2023.

[Chen *et al.*, 2023b] Yu Chen, Lingfei Wu, and Mohammed J. Zaki. Toward subgraph-guided knowledge graph question generation with graph neural networks. *TNNLS*, 2023.

[Do *et al.*, 2022] Xuan Long Do, Bowei Zou, Liangming Pan, Nancy F. Chen, Shafiq R. Joty, and Ai Ti Aw. Cohs-cqg: Context and history selection for conversational question generation. In *COLING*, 2022.

[Do *et al.*, 2023] Xuan Long Do, Bowei Zou, Shafiq R. Joty, Anh Tran Tai, Liangming Pan, Nancy F. Chen, and Ai Ti Aw. Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation. In *ACL*, 2023.

[Dong *et al.*, 2019] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, 2019.

[Dong *et al.*, 2024] Chenhe Dong, Ying Shen, Shiyang Lin, Zhenzhou Lin, and Yang Deng. A unified framework for contextual and factoid question generation. *TKDE*, 2024.

[Du and Claire, 2018] Xinya Du and Claire. Harvesting paragraph-level question-answer pairs from wikipedia. In *ACL*, 2018.

[Du *et al.*, 2017] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *ACL*, 2017.

[Elsahar *et al.*, 2018] Hady Elsahar, Christophe Gravier, and Frederique Laforest. Zero-shot question generation from knowledge graphs for unseen predicates and entity types. In *NAACL*, 2018.

[Fei *et al.*, 2021] Zichu Fei, Qi Zhang, and Yaqian Zhou. Iterative gnn-based decoder for question generation. In *EMNLP*, 2021.

[Fei *et al.*, 2022a] Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuanjing Huang. CQG: A simple and effective controlled generation framework for multi-hop question generation. In *ACL*, 2022.

[Fei *et al.*, 2022b] Zichu Fei, Xin Zhou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Lfkqg: A controlled generation framework with local fine-tuning for question generation over knowledge bases. In *COLING*, 2022.

[Gonzalez *et al.*, 2023] Hannah Gonzalez, Liam Dugan, Eleni Miltsakaki, Zhiqi Cui, Jiaxuan Ren, Bryan Li, Shriyash Upadhyay, Etan Ginsberg, and Chris Callison-Burch. Enhancing human summaries for question-answer generation in education. In *BEA@ACL*, 2023.

[Gu *et al.*, 2021] Yu Gu, Sue Kase, Michelle Vanni, Brian M. Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond I.I.D.: three levels of generalization for question answering on knowledge bases. In *WWW*, 2021.

[Guo *et al.*, 2022] Shasha Guo, Jing Zhang, Yanling Wang, Qianyi Zhang, Cuiping Li, and Hong Chen. Dsm: Question generation over knowledge base via modeling diverse subgraphs with meta-learner. In *EMNLP*, 2022.

[Guo *et al.*, 2023] Shasha Guo, Jing Zhang, Xirui Ke, Cuiping Li, and Hong Chen. Diversifying question generation over knowledge base via external natural questions. *CoRR*, 2023.

[Guo *et al.*, 2024] Shasha Guo, Lizi Liao, Jing Zhang, Yanling Wang, Cuiping Li, and Hong Chen. Sgsh: Stimulate large language models with skeleton heuristics for knowledge base question generation. *arXiv preprint arXiv:2404.01923*, 2024.

[Ke *et al.*, 2021] Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. In *ACL*, 2021.

[Kim *et al.*, 2019] Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. Improving neural question generation using answer separation. In *AAAI*, 2019.

[Krishna *et al.*, 2019] Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *CVPR*, 2019.

[Kumar *et al.*, 2019a] Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. Difficulty-controllable multi-hop question generation from knowledge graphs. In *ISWC*, 2019.

[Kumar *et al.*, 2019b] Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. Cross-lingual training for automatic question generation. In *ACL*, 2019.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, 2016.

[Li *et al.*, 2018] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *CVPR*, 2018.

[Li *et al.*, 2019] Jingjing Li, Yifan Gao, Lidong Bing, Irwin King, and Michael R. Lyu. Improving question generation with to the point context. In *EMNLP*, 2019.

[Liang *et al.*, 2023] Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Weining Qian, and Yunshi Lan. Prompting large language models with chain-of-thought for few-shot knowledge base question generation. In *EMNLP*, 2023.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[Ling *et al.*, 2020] Yanxiang Ling, Fei Cai, Honghui Chen, and Maarten de Rijke. Leveraging context for neural question generation in open-domain dialogue systems. In *WWW*, 2020.

[Liu *et al.*, 2018] Feng Liu, Tao Xiang, Timothy M. Hospedales, Wankou Yang, and Changyin Sun. IVQA: inverse visual question answering. In *CVPR*, 2018.

[Liu *et al.*, 2019] Cao Liu, Kang Liu, Shizhu He, Zaiqing Nie, and Jun Zhao. Generating questions for knowledge bases via incorporating diversified contexts and answer-aware loss. In *EMNLP-IJCNLP*, 2019.

[Liu *et al.*, 2023] Chao Liu, Xuanlin Bao, Hongyu Zhang, Neng Zhang, Haibo Hu, Xiaohong Zhang, and Meng Yan. Improving chatgpt prompt for code generation. *CoRR*, 2023.

[Mi *et al.*, 2024] Li Mi, Syrielle Montariol, Javiera Castillo Navarro, Xianjie Dai, Antoine Bosselut, and Devis Tuia. Convqg: Contrastive visual question generation with multimodal guidance. In *AAAI*, 2024.

[Mohammadshahi *et al.*, 2023] Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. RQUGE: reference-free metric for evaluating question generation by answering the question. In *Findings of ACL*, 2023.

[Mostafazadeh *et al.*, 2016] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *ACL*, 2016.

[Nan *et al.*, 2023] Linyong Nan, Yilun Zhao, Weijin Zou, Narutatsu Ri, Jaesung Tae, Ellen Zhang, and et al. Enhancing few-shot text-to-sql capabilities of large language models. *arXiv*, 2023.

[Nguyen *et al.*, 2016] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. A human generated machine reading comprehension dataset. In *NeurIPS*, 2016.

[Pan *et al.*, 2019a] Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. Reinforced dynamic reasoning for conversational question generation. In *ACL*, 2019.

[Pan *et al.*, 2019b] Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. Recent advances in neural question generation. *CoRR*, 2019.

[Pan *et al.*, 2020] Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. Semantic graphs for generating deep questions. In *ACL*, 2020.

[Pan *et al.*, 2021a] Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. Unsupervised multi-hop question answering by question generation. In *NAACL*, 2021.

[Pan *et al.*, 2021b] Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. Zero-shot fact verification by claim generation. In *ACL*, 2021.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

[Rajpurkar *et al.*, 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, 2016.

[Reddy *et al.*, 2017] Sathish Reddy, Dinesh Raghu, Mitesh M. Khapra, and Sachindra Joshi. Generating natural language question-answer pairs from a knowledge graph using a rnn based question generation model. In *EACL*, 2017.

[Reddy *et al.*, 2019] Siva Reddy, Danqi Chen, and Christopher D. Manning. A conversational question answering challenge. *TACL*, 2019.

[Saeidi *et al.*, 2018] Marzieh Saeidi, Max Bartolo, Patrick S. H. Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *EMNLP*, 2018.

[Serban *et al.*, 2016] Iulian Vlad Serban, Alberto García-Durán, Çaglar Gülçehre, Sungjin Ahn, Sarath Chandar, Aaron C. Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks. In *ACL*, 2016.

[Shen *et al.*, 2018] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*, 2018.

[Su *et al.*, 2020] Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. Multi-hop question generation with graph convolutional network. In *EMNLP*, 2020.

[Subramanian *et al.*, 2018] Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. Neural models for key phrase extraction and question generation. In *ACL*, 2018.

[Sun *et al.*, 2018] Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. Answer-focused and position-aware neural question generation. In *EMNLP*, 2018.

[Talmor and Berant, 2018] Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *NAACL-HLT*, 2018.

[Trischler *et al.*, 2017] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *ACL*, 2017.

[Tuan *et al.*, 2020] Luu Anh Tuan, Darsh J. Shah, and Regina Barzilay. Capturing greater context for question generation. In *AAAI*, 2020.

[Uppal *et al.*, 2020] Shagun Uppal, Anish Madan, Sarthak Bhagat, Yi Yu, and Rajiv Ratn Shah. C3VQG: category consistent cyclic visual question generation. In *ACM MM Asia*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wang and Baraniuk, 2023] Zichao Wang and Richard G. Baraniuk. Multiqg-ti: Towards question generation from multi-modal sources. In *ACL*, 2023.

[Wang *et al.*, 2020a] Bingning Wang, Xiaochuan Wang, Ting Tao, Qi Zhang, and Jingfang Xu. Neural question generation with answer pivot. In *AAAI*, 2020.

[Wang *et al.*, 2020b] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, 2020.

[Xia *et al.*, 2023] Zehua Xia, Qi Gou, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li, and Cam-Tu Nguyen. Improving question generation with multi-level content planning. In *EMNLP*, 2023.

[Xiao *et al.*, 2020] Dongling Xiao, Han Zhang, Yu-Kun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *IJCAI*, 2020.

[Xie *et al.*, 2021] Jiayuan Xie, Yi Cai, Qingbao Huang, and Tao Wang. Multiple objects-aware visual question generation. In *ACM MM*, 2021.

[Xie *et al.*, 2022] Jiayuan Xie, Wenhao Fang, Yi Cai, Qingbao Huang, and Qing Li. Knowledge-based visual question generation. *IEEE Trans. Circuits Syst. Video Technol.*, 2022.

[Xiong *et al.*, 2022] Guanming Xiong, Junwei Bao, Wen Zhao, Youzheng Wu, and Xiaodong He. Autoqgs: Auto-prompt for low-resource knowledge-based question generation from SPARQL. In *CIKM*, 2022.

[Xu *et al.*, 2020] Xing Xu, Tan Wang, Yang Yang, Alan Hanjalic, and Heng Tao Shen. Radial graph convolutional network for visual question generation. *TNNLS*, 2020.

[Yang *et al.*, 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

[Yih *et al.*, 2016] Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In *ACL*, 2016.

[Yoon and Bak, 2023] Hokeun Yoon and JinYeong Bak. Diversity enhanced narrative question generation for storybooks. In *EMNLP*, 2023.

[Zhang and Gao, 2023] Xuan Zhang and Wei Gao. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *AACL*, 2023.

[Zhang *et al.*, 2020] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020.

[Zhang *et al.*, 2022a] Kun Zhang, Yunqi Qiu, Yuanzhuo Wang, Long Bai, Wei Li, Xuhui Jiang, Huawei Shen, and Xueqi Cheng. Meta-cqg: A meta-learning framework for complex question generation over knowledge bases. In *COLING*, 2022.

[Zhang *et al.*, 2022b] Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. A review on question generation from natural language text. *TOIS*, 2022.

[Zhao *et al.*, 2022] Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. Educational question generation of children storybooks. In *ACL*, 2022.

[Zhou *et al.*, 2018] Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. An interpretable reasoning network for multi-relation question answering. In *COLING*, 2018.