Int J Artif Intell Educ
https://doi.org/10.1007/s40593-023-00374-x

ARTICLE

Review on Neural Question Generation for Education Purposes

Said Al Faraby[1] · **Adiwijaya Adiwijaya**[1] · **Ade Romadhony**[1]

Accepted: 4 October 2023

Abstract

Questioning plays a vital role in education, directing knowledge construction and assessing students' understanding. However, creating high-level questions requires significant creativity and effort. Automatic question generation is expected to facilitate the generation of not only fluent and relevant but also educationally valuable questions. While rule-based methods are intuitive for short inputs, they struggle with longer and more complex inputs. Neural question generation (NQG) has shown better results in this regard. This review summarizes the advancements in NQG between 2016 and early 2022. The focus is on the development of NQG for educational purposes, including challenges and research opportunities. We found that although NQG can generate fluent and relevant factoid-type questions, few studies focus on education. Specifically, there is limited literature using context in the form of multi-paragraphs, which due to the input limitation of the current deep learning techniques, require key content identification. The desirable key content should be important to specific topics or learning objectives and be able to generate certain types of questions. A further research opportunity is controllable NQG systems, which can be customized by taking into account factors like difficulty level, desired answer type, and other individualized needs. Equally important, the results of our review also suggest that it is necessary to create datasets specific to the question generation tasks with annotations that support better learning for neural-based methods.

Keywords Question generation · Neural methods · Educational application · Key content identification · Usefulness

✉ Adiwijaya Adiwijaya
adiwijaya@telkomuniversity.ac.id

Said Al Faraby
saidalfaraby@telkomuniversity.ac.id

Ade Romadhony
aderomadhony@telkomuniversity.ac.id

[1] School of Computing, Telkom University, Bandung, Indonesia

Published online: 31 October 2023

Introduction

Active retrieval is an effective learning technique with long-term effects. It involves processes of reconstructing knowledge stored in the brain. One form of active retrieval activity is formal questioning, such as during exams, or informal questioning integrated with classroom learning or carried out independently by students (Karpicke, 2012). It is also known that certain types of questions accelerate learning more than others. Many terms have become associated with these types of advanced questions, such as conceptual questions (Bugg and McDaniel, 2012), deep questions (Graesser et al., 2010), inference questions (Sundbye et al., 1987; Kispal, 2008), and higher-order questions (Renaud and Murray, 2003; Bloom, 1956). For example, inference questions may require the connection of several parts of a text (text-connecting) to obtain an answer (Chikalanga, 1992). High-level questions refer specifically to Bloom's taxonomy; they are questions that hone analysis, synthesis, and evaluation skills.

Questions are one of the essential instruments for educators and students (Chin and Osborne, 2008). Unfortunately, the skill of developing an advanced question is not innate, and many students have difficulty making such questions (Nappi, 2017). Previous research into Automatic Question Generation (AQG) has mainly used rule-based approaches, wherein specific rules are defined to transform input text into questions (Ali et al., 2010; Mitkov and Le An, 2003; Heilman and Smith, 2010; Mostow and Chen, 2009). Rules-based approaches are intuitive for short and simple text input but encounter difficulty when handling increasingly long and complex inputs. To address these limitations, previous research has introduced interesting solutions, such as utilizing semantic maps derived from lengthy reading texts and selecting specific segments to generate questions using rule-based methods (Jouault et al., 2016).

However, with the advent of Deep Learning-based approaches, there has been a significant shift towards using neural methods for AQG, commonly referred to as NQG. Deep learning models have shown empirical superiority over rule-based approaches in terms of automatic evaluation, which measures the similarity between the generated questions and reference questions (Du et al., 2017). Over the last five years, deep learning has become the primary method for AQG, driving the advancements in the field (Zhou et al., 2017; Zhao et al., 2018; Dong et al., 2019; Zhang et al., 2021).

NQG has been used for various purposes, such as data augmentation for question-answering systems (QAS), drafting quizzes for educational purposes, evaluating factual consistency in automatic summarization, developing conversational agents, and information-seeking. In this review, we focus on the use of NQG for educational purposes. Kurdi et al. (2020) conducted a similar review but did not limit the methods considered. Most reviewed literature used template-based methods; indeed only nine studies used statistical methods, including neural methods. Several other related literature reviews did not focus on NQG (Zhang et al., 2021; Das et al., 2021; Amidei et al., 2018). Table 1 summarizes previous literature reviews and explains how they are complemented by the present review.

In this review, our primary focus is on the utilization of neural methods in AQG within the context of education. After conducting an initial survey, we observed that while there is a substantial body of literature addressing NQG, literature specifically tailored to educational purposes is relatively limited. Therefore, in this review, we

Table 1 Previous related literature reviews and how they are complemented by this review

No	Paper	Year	Gap
1	(Zhang et al., 2021)	2021	Comprehensively summarizes the development of neural question generation in general, but does not consider educational purposes. Challenges impeding the practical use of NQG for education and solutions progress are not outlined.
2	(Das et al., 2021)	2021	Briefly discusses the creation of questions and evaluation of answers, focusing on the use of automated question generation and evaluation for online learning. Most of the literature reviewed uses the rule-based model; only a few studies are neural-based.
3	(Kurdi et al., 2020)	2020	Comprehensive review of question generation for educational purposes, covering studies from 2015 to early 2019 but does not focus on neural-based methods. Important developments have been made since 2019.
4	(Amidei et al., 2018)	2018	Focuses only on the evaluation of automated question generation and does not explicitly address it from the perspective of educational goals.

aim to explore how NQG techniques can be adapted to meet the specific requirements of the educational field. To achieve this, we aim to answer the following research questions:

- RQ1: What is the current state-of-the-art in NQG research?
- RQ2: How can existing NQG methods be customized to address educational needs?
- RQ3: What are the gaps in NQG research for educational purposes, and how can future research bridge these gaps?

The remainder of this review is structured as follows: Section 2 provides a theoretical foundation on questions from the perspective of cognitive science, helping readers understand the distinct criteria that set educational questions apart. In Section 3, we elaborate on our review methodology. Section 4 delves into the main contributions of this review, while Section 5 concludes the review and outlines potential avenues for future research. Finally, Section 6 highlights the limitations of this review.

Background Theory

Before delving into the literature review on Neural Question Generation (NQG) for educational purposes, it is crucial to establish key definitions and concepts related to NQG and questions in the field of education. These definitions and concepts will provide a clear understanding of the domain and create a foundation for the following discussion on NQG in the education sector.

Neural Question Generation (NQG)

Automatic Question Generation (AQG) refers to the process of automatically generating *questions* from a given input *context*. Neural Question Generation (NQG) is a specific form of AQG that utilizes *neural network*-based models for question generation.

Neural-based methods, such as sequence-to-sequence models and transformers, leverage deep learning techniques to automatically learn patterns and relationships in the input context. These models can capture complex dependencies and generate questions that may not have been explicitly defined in the rules. They are data-driven and can be adapted to different domains with sufficient training data, making them versatile and capable of handling a wide range of contexts. However, there are some drawbacks to NQG. Firstly, these models often require extensive training on large datasets of human-generated questions. Secondly, the training and execution of NQG models can be computationally expensive. Lastly, controlling the types of questions generated can be challenging due to the black-box nature of these models.

On the other hand, rule-based methods rely on predefined sets of rules or patterns to generate questions. These rules are often handcrafted and require expert knowledge to design. While rule-based methods can produce accurate and controlled question patterns, they may struggle to handle more diverse and complex contexts that were not anticipated during rule design. Additionally, rule-based systems can be time-consuming and challenging to update or modify when new contexts arise.

In Question Generation (QG), there are three essential components: context, question, and answer.

- **Context**. The context refers to the input text or information from which the question needs to be generated. This is the term commonly used in NQG literature. It serves as the source of information that the QG system uses to understand the content and generate relevant questions. Besides the most common context, which is natural text, other contexts can include knowledge graphs, keywords, structured text (e.g., SQL, formulas, tables), and/or images. This 'context' is different from 'discourse context' which refers to the surrounding language or text that helps in understanding the meaning and interpretation of a specific word, sentence, or passage.
- **Question**. While questions are predominantly generated in natural text format, various factors differentiate them. These factors include the type of question, such as *factoid* or *non-factoid*, the level of complexity, such as *single-hop* or *multi-hop* (*hop* denotes the process of navigating between information pieces to find an

answer), the structure of the question, such as *standalone* or *sequence*, and the format of the question, such as *free text* or *multiple choice*. In addition to these evident variations, subtle distinctions may also arise, such as the intended purpose or application of the questions.

- **Answer**. The answer represents the information that the generated question seeks to elicit or inquire about. In some cases, the answer may be explicitly provided to the QG system, in this case, it is called *answer-aware QG*. On the other hand, when the system does not utilize answers during the question generation process, it is referred to as *answer-unaware QG*. While the presence of an answer is optional in the QG process, it can be highly beneficial in creating more specific and relevant questions.

Together, these three components form the basis of Question Generation, where the context serves as the input, the question is generated as the output, and the answer represents the information sought through the question.

Educational Purposes

The function of questions for educational purposes is to recall prior knowledge, test comprehension, and hone critical thinking (Tofade et al., 2013). Both students and teachers can use questions for these purposes. Students can use questions to help form knowledge during the learning process (Chin and Osborne, 2008), such as clarifying the meaning and significance of questions. In addition, students can use questions for independent study in order to better understand subject materials. Teachers can use questions to evaluate students' learning, explore a subject further, or provoke discussion. These goals necessarily influence the criteria for the required QG system.

Types of Questions

Based on the Question-Answer relationship, question types can be differentiated depending on the *type of answer expected*, which is determined using *question words*. For example, questions that start with *where* usually expect the answer to be a *location*, whereas questions beginning with *when* are about *time*. Besides *wh*-questions, there are also *yes/no* questions, for example starting with *be*, *do*, or modal verb words (e.g., *can, should, will*).

What kind of questions are used for educational purposes? To answer this question, we review the literature from cognitive sciences and educational psychology. Graesser and colleagues have conducted extensive research on questions across the fields of psychology, education, and artificial intelligence (Graesser and Person, 1994; Graesser et al., 2010). Their studies have identified sixteen common question types in education, distinguished by the *type of expected answers*. Furthermore, questions can also be categorized based on the *cognitive processes* required to obtain answers, as outlined in *Bloom's taxonomy*. Each question type in Bloom's and Graesser's taxonomies can be further classified into three common difficulty levels: low, medium, and high, as shown in Table 2. It's important to note that the table focuses on displaying the list

Table 2 Relation of Bloom's taxonomy to the type of comprehension and purpose of the question and the level of difficulty

Level	Type of question (Graesser and Person, 1994)	Type of cognitive process (Bloom, 1956)
Low	Verification	Recognition
	Disjunctive	Recall
	Concept completion	
	Example	
Medium	Feature specification	Comprehension
	Quantification	
	Definition	
	Comparison	
High	Interpretation	Application
	Causal antecedent	Analysis
	Causal consequence	Synthesis
	Goal orientation	Evaluation
	Instrumental/Procedural	
	Enablement	
	Expectation	
	Judgmental	

of question types and their respective difficulty levels, rather than directly equating individual question types between Graesser and Bloom.

In addition to the previous groupings, questions can also be categorized based on comprehension levels, including *literal*, *inferential*, and *evaluative* levels (Basaraba et al., 2013). Literal questions involve directly recalling explicitly stated information from the text, while inferential questions require interpreting implicit information. Evaluative questions, on the other hand, require evaluating information based on personal knowledge and experience. Furthermore, inference questions can be categorized in various ways, for example, the division into lexical (e.g., pronominal inference), propositional (e.g., logical spatio-temporal), and pragmatics (e.g., elaborative motivational), as proposed by Chikalanga (1992). Overall, these categorizations provide valuable insights into different question types and their cognitive demands during the learning process.

Review Methods

Literature Search and Evaluation

This section describes the procedures undertaken for collecting all candidate literature in this study, including keywords, search engines, and the criteria used to eliminate candidates.

Search Methods

Searches were carried out sequentially using Google Scholar, ACL, ACM, DBLP, and Springer databases, as well as by considering citations made by or to several articles collected. We used the keyword *"question generation"*. We deliberately use fairly general keywords instead of more specific ones such as *"question generation for education"* or *"neural question generation"* in order to collect as many publications as possible and subsequently make further selections. We set the search period to start in 2016 because, based on our initial research, we found that the first study on NQG was published that year. Our search results were sorted by relevance, and we retrieved all articles until reaching the page where the results ceased to be relevant. As a final step, we followed the "cited by" of more than ten of the most cited papers to ensure that important studies were not overlooked.

Evaluation

Upon retrieving the search results, we conducted an evaluation of each publication to ascertain its relevance to our research question. To do this, we skimmed each paper unless it was clear that it could be determined only from the title. We considered all papers related to automatic NQG without limitation to a particular domain. We employed six exclusion criteria in a sequential order to filter the literature, as outlined below. Any literature that met one or more of these criteria was excluded from further consideration:

1. *Full-text cannot be accessed*. During the process of collecting literature from search engines using a reference management tool[1], some publications were not retrievable as PDFs, potentially due to metadata errors or restricted access. To address this issue, we conducted manual searches to overcome problems stemming from metadata errors. Publications to which we did not have access were excluded from the review.
2. *Not question generation*. Upon skimming through the obtained publications, we discovered that some of them did not pertain to the creation of questions.
3. *Not automatic generation*. Some studies did create questions but did not involve the automatic generation of questions by a system (e.g., questions created by students). These studies were also excluded to maintain the focus of the discussion.
4. *Not neural-based method*. To ensure a more concentrated analysis, studies that did not utilize neural-based methods were also excluded.
5. *Secondary study*. Additionally, secondary studies (e.g., survey papers) were excluded to avoid redundancy and maintain a focus on original research.
6. *Low quality*. Finally, papers that did not meet the general criteria for a research paper, such as having unclear or incomplete writing (e.g., lacking a results section), were also excluded.

[1] paperpile.com

Information Extraction

After gathering all the relevant publications, we extracted a pre-defined list of information from each paper to create a meaningful taxonomy and organize the review results. This list of information includes the title, year, type of publication, publisher, applications supported by question generation, datasets, the type of questions generated, generation methods, and human evaluation of the generated questions.

Results

Included Literature

Following the search and evaluation process outlined previously, a total of 302 unique papers were obtained by searching keywords on various platforms: 201 from Google Scholar, 34 from ACL, 16 from ACM, and 51 from DBLP. An additional 88 studies were identified by exploring the 'cited by' articles from the previously obtained papers, bringing the total to 390 papers. After applying the evaluation criteria described earlier, 224 articles were deemed to meet the criteria for inclusion in our review. These articles will be further analyzed and discussed in the subsequent review section.

The main reason for article exclusion was the absence of neural-based methods (i.e., methods that employ variants of neural networks, such as Recurrent Neural Network and Transformer) utilized for question construction (79 articles). The second criterion for exclusion was the unavailability of PDF access to certain articles. A total of 44 articles met this criterion comprising 27 conference papers, 10 journal articles, 4 theses, 2 book chapters, and 1 preprint manuscript. The remaining exclusions were due to being unrelated to question generation (13 articles), not involving automatic generation (7 articles), being secondary studies (4 articles), and exhibiting low quality (1 article). Furthermore, we have excluded 17 articles that specifically focus on Visual Question Generation (VQG). The rationale behind this exclusion is that VQG typically involves an additional step of converting images to text before applying similar methods as question generation from text to generate questions. Given the similarity in methodology and the distinct focus on visual input, we have opted to exclude these articles from our review.

Rate of Publication

Since their introduction in 2016 (Mostafazadeh et al., 2016; Serban et al., 2016), the number of published studies on NQG has generally grown each year (2016=2, 2017=10, 2018=29, 2019=49, 2020=60, 2021=70, 2022=4), with the exception of 2022 (due to the cutoff date of our literature search). This indicates that the topic of QG is still open for further research. This is still true when the search criteria are narrowed to literature closely related to education (11 papers, of which 6 were published in 2021). This number is very small compared to the number of non-neural QG literature related to education (Kurdi et al., 2020). Thus, research, and development opportunities for NQG for education are still very much needed. Hopefully, this review

can assist researchers in identifying challenges, knowledge gaps, and directions for further research.

Publication Venue

In terms of conference proceedings, Empirical Methods in Natural Language Processing (EMNLP) is the most frequently encountered venue. Machine Reading Question Answering (MRQA) stands out as the most common choice for workshops. As for journals, the International Journal of Artificial Intelligence in Education (IJAIED) and Transactions of the Association for Computational Linguistics (TACL) are widely utilized in the field of NQG. For a more complete list, please see the Supplementary Materials[2].

Indeed, a total of 23 studies included in our review were published on Arxiv, which is an open-access preprint repository. We included articles from Arxiv intentionally. This was done for two main reasons. Firstly, some of these articles have high citations but are only published on that platform. Secondly, certain articles had metadata errors, resulting in them being unsearchable or unavailable in reference management tools. In addition, incorporating Arxiv papers allowed us to gather a comprehensive collection of studies. However, it is important to note that these papers have not yet undergone formal peer review. As such, we encourage readers to explore these studies independently and assess their quality and significance.

Neural Question Generation for Education

As stated above, although many studies discuss NQG, few are specifically related to education as indicated by special evaluations or special datasets. Therefore, this section does not compare all studies using the same criteria but rather highlights some supporting parts used for education purposes. In this section, we present eleven subsections covering applications to evaluations, where we discuss the significant findings from the reviewed papers. To facilitate readers in locating specific papers relevant to each subsection, we have compiled lists of papers into several tables available in the Supplementary Materials.

Applications of NQG

Most of the early research on NQG was motivated by the limited data available to train QAS models (Serban et al., 2016; Yang et al., 2017; Reddy et al., 2017), so synthetic questions were needed in addition to those created manually. The following motivation includes educational purposes, e.g., reading comprehension (Zhou et al., 2017; Du et al., 2017). Creating questions and answers manually requires a significant amount of effort. This task can be challenging for novice students. Thus, an automated system capable of generating questions can assist teachers while also helping students learn independently through active retrieval.

[2] https://www.github.com/saidalfaraby/Review-QG-for-Education

New practical applications of QG have since been investigated. Conversational AI is one such area in which generated questions play an essential role in helping agents provide interesting interactions, clarify information, and request new information (Sekulić et al., 2021; Wang et al., 2018; Qi et al., 2020). Last is summarization, wherein QG systems are used to evaluate factual consistency in summarization results, identify and select which parts of an input are eligible for questioning, and summarize questions from QA forums into compelling titles (Guo et al., 2018; Wang et al., 2020; Gao et al., 2020).

Existing Work on NQG for Education Purposes

In this section, we present the studies that we consider to have specific educational goals, as shown by their use of datasets, methods, question types, or special evaluations closely related to education. Table 3 summarizes the relevant studies along with their rationales.

Relevance to Education Purposes

Three aspects were used to categorize NQG literature according to its educational relevance:

Relevant Dataset We consider a dataset to be specific for educational purposes if it was collected from educational sources or if there is a contribution from experts (e.g., teachers) in its preparation. Some specific datasets used in the NQG literature include OpenStax (Rice University, 1999), a university textbook, RACE (Lai et al., 2017), which was taken from an English reading test in China, LearningQ (Chen et al., 2018), which was compiled from materials and questions provided by instructors on an online learning platform, and TQA (Kembhavi et al., 2017), which was compiled using school textbooks. Further explanation of these datasets is provided in Section 4.4.

Relevant Question Types Most questions generated in the literature are factoids, which is reasonable considering that popular datasets such as SQuAD and HotpotQA contain mostly factoid questions. For this reason, we searched the literature that explicitly describes question types with a greater focus on education. Several studies (Cao and Wang, 2021; Steuer et al., 2021; Stasaski et al., 2021) were found to use the same types of questions derived from cognitive science as presented in Section 2.

Relevant Evaluation NQG systems are commonly evaluated using automatic metrics such as BLEU, METEOR, and ROUGE-L, as well as crowd-sourced human evaluations that assess fluency, relevance, and answerability. While these criteria are important, they do not fully encompass an essential aspect of questions in an educational context: their usefulness in supporting learning. Useful questions go beyond mere factual recall and instead possess the ability to assess students' knowledge, promote comprehension, and stimulate critical thinking (Tofade et al., 2013). Hence, we solely consider evaluations in papers as relevant to education if they explicitly measure the usefulness of the generated questions in any capacity. Further explanation on evaluation is presented in Section 4.11.

Table 3 Literature closely related to educational purposes

Literature	Question type	Evaluation method	Educational relevant by
Wang et al. (2018)	Factoid	Evaluators: crowdsourced Criteria: - fluency - relevance - human-like	Evaluation on textbooks data
Willis et al. (2019)	Factoid	Evaluators: domain expert Criteria: matching extracted keywords	Evaluation by education experts
Cao and Wang (2021)	- Verification - Disjunctive - Concept - Extent - Example - Comparison - Cause - Consequence - Procedural - Judgmental	Evaluators: 3 annotators Criteria: - diversity (type, syntax, answer content) - content quality (appropriateness, answerability, scope)	Question types from cognitive science
Steuer et al. (2021)	Definitional	Evaluators: experts Criteria: - Horbach scheme (Horbach et al., 2020)	- Dataset from textbooks - Question types from cognitive science - Evaluation by education experts
Jia et al. (2021)	English Exam-like (including factoid, cause, consequence)	Evaluators: 3 annotators Criteria: - fluency - relevancy - answerability	Dataset from school exam

Table 3 continued

Literature	Question type	Evaluation method	Educational relevant by
Steuer et al. (2020)	Factoid	Evaluators: 2 Annotators Criteria: - Grammar - Answerability - Usefulness for learning	Dataset from school exam
Cheng et al. (2021)	Factoid	Evaluators: crowdsourced 50 people Criteria: - usefulness for learning - linguistic comprehensibility - matching answers	Evaluation measures usefulness for learning
Murakhovs'ka et al. (2021)	Factoid and non-factoid (including Yes-No)	Evaluators: authors Criteria: - fluency - relevancy	Question types beyond factoid
Krishna and Iyyer (2019)	Specific (factoid), General, and Yes-No (non-Factoid)	Evaluators: crowdsourced Criteria: - fluency - relevancy - answerability	Question types beyond factoid
Stasaski et al. (2021)	Cause and Consequence	Evaluators: crowdsourced Criteria: - matching question type (cause or consequence) - matching answers	Dataset from textbooks
Qu et al. (2021)	Non-Factoid	Evaluators: 3 evaluators Criteria: - fluency - relevance - answerability	Dataset from school exam

Tasks, Domains, and Applications Related to Education

In this study, we categorize educational tasks that can be facilitated by the current advancements in NQG research into three types: (1) *generating questions for reading comprehension*, (2) *generating word problems*, and (3) *generating questions for conversation or interaction modes*.

Question Generation from Reading Materials

Generating Questions from reading materials is based on a reading, whereupon the answers to those questions are found directly or indirectly in the text. This task is common in education, such as questions at the end of a book chapter to review readers' understanding of some critical concepts or questions in language tests that examine the ability to remember facts or draw inferences from given reading materials.

In the field of Neural Question Generation (NQG), researchers have not only relied on non-domain-specific texts like Wikipedia articles but have also explored diverse reading materials from various domains. In the context of education, datasets have been obtained from sources such as children's storybooks (Yao et al., 2022), school science textbooks (Stasaski et al., 2021), and specific subjects extracted from university textbooks (Wang et al., 2018; Steuer et al., 2021). This broad range of reading materials allows for more comprehensive and domain-specific question generation research in the educational context.

Word Problem Generation

In contrast to generating questions for reading comprehension, where the target answers (e.g., word, span) are typically derived from a larger input context, the task of generating word problems involves using the target answers as the main input context. These target answers are often represented in the form of equations (Zhou and Huang, 2019; Cao et al., 2021; Wang et al., 2021; Liu et al., 2021) or structured languages like SQL queries (Guo et al., 2018; Yu and Jiang, 2021). An example question derived from an equation is depicted in Fig. 1. The target answer to the generated question is formulated using the same equation used as the context.

Context (Equation):
$y = 3 * x;\ y - 20 = 100$

Question:
The teacher said if you multiply my age by 3, then subtract 20, the result is 100. How old is the teacher?.

Fig. 1 Word problem example from Cao et al. (2021)

Topic: Daffy Duck, Origin and History
Question1: What is the origin of Daffy Duck?
Answer1: first appeared in Porky's Duck Hunt
Question2: What was he like in that episode?
Answer2: assertive, unrestrained, combative
Question3: Was he the star?
Answer3: No, barely more than an unnamed bit player in this short
Question4: Who was the star?
Answer4: No answer

Fig. 2 An example of open domain question in a conversational setting from Choi et al. (2018)

Conversation Question Generation

NQG systems can also be used in interactive and open-domain situations. For example, a series of questions can be asked, in which each relates to the previous question and its response. When a question in a conversation receives a "No answer" response, it indicates that further follow-up questions should be avoided, and instead, the direction of the questioning should be altered. These systems can be deployed in chat apps or learning discussion forums to augment student-teacher interactions in the process of knowledge formation and topic exploration. An example of the data used for this task can be seen in Fig. 2.

Dataset

The availability of datasets is a crucial prerequisite for producing neural-based QG. Datasets that were widely used during the early days of NQG research were specifically aimed at improving QA systems, such as SQuAD, MS Marco (Bajaj et al., 2016), and NewsQA (Trischler et al., 2016). These datasets come from Wikipedia and news articles, and can thus be considered in the general domain. For a more detailed review of the various datasets devoted to QA, see Bai and Wang (2021). In this subsection, we discuss some of the datasets utilized in education-related NQG studies. For a comprehensive list of datasets found in the NQG literature, please refer to the Supplementary Materials.

- *OpenStax* (Rice University, 1999) is a collection of textbooks created by domain experts and widely used for education. However, this dataset is not specifically designed for question-answering or generation, which typically requires triplets of ⟨context, answer, question⟩. Therefore, it is often necessary to perform additional processing steps, for example, segmenting the context into smaller parts and extracting an answer using a rule-based approach and heuristics as in Steuer et al. (2021) and Wang et al. (2018). In the absence of reference questions in the datasets, it is possible to train an NQG model using alternative datasets, such as SQuAD (Stanford Question Answering Dataset), and then generate questions for OpenStax.

- *RACE (ReAding Comprehension from Examinations)* (Lai et al., 2017) is a question-answering dataset derived from reading comprehension exams conducted in middle and high school English lessons in China. Therefore, the context, questions, and answers were created by domain experts. However, it is worth noting that the questions in RACE primarily focus on assessing comprehension and English language skills, rather than requiring higher-order cognitive abilities. Additionally, RACE includes general-style questions such as "What would be the best title for the paragraph?", which are less interesting for QG as they are not tied to specific reading articles and can be generated using rule-based methods. To make the dataset more suitable for QG, (Jia et al., 2021) performed filtering by exclusively considering questions that are specific to a particular reading article.
- *LearningQ* (Chen et al., 2018) is composed of questions submitted to online learning forums by both instructors and students. Questions from instructors were prepared to access knowledge, whereas questions from students usually seek information or confirm material. Many questions require higher-order cognitive skills, in contrast to RACE, which tests more on literal comprehension and understanding of English. Unfortunately, the annotations available consist only of context and questions. The document context is longer than other average datasets, such as SQuAD (10 x longer), and there is no keyword annotation or target answer. To overcome the unavailability of answers, (Steuer et al., 2020) performed answer selection based on word categories, then used a model trained with other datasets (e.g., SQuAD) to generate questions.
- *TQA* (Kembhavi et al., 2017) is a dataset compiled from a middle school science textbook. Its context includes text, images, and diagrams. Questions and answers were taken from workbooks and quizzes available on the corresponding website, plus crowdsourced questions for context involving diagrams. The questions in this dataset require complex reasoning and parsing of diagrams and figures, which remains difficult for current QA or QG systems. As a result, (Stasaski et al., 2021) used only the textual context to generate cause-and-effect questions.

Features of Ideal Datasets

In conclusion, when considering the development and training of educational question generation models, it is imperative to prioritize the incorporation of specific features within the dataset. These features not only contribute to the quality of the models but also align with the overarching objectives of enhancing the educational question generation process.

- Focus on higher-level questions: Emphasize the process of generating higher-level questions rather than just seeking answers, similar to question-answering tasks.
- Include diverse and comprehensive educational contexts: Encompass a wide range of subjects and topics to analyze question characteristics across different educational domains.
- Optional answers with depth: If answers are included, they should go beyond short and extractive responses, providing more in-depth and informative content.

- Highlights for key content selection: For longer contexts, incorporate highlights or specific portions that contribute to generating corresponding questions. This aids in training key content selection models and simplifies neural model training.
- Provide shorter context for human verification: Utilize the output of key content selection as a shorter, more focused context for the question generator. This enables easier human verification of question relevance.
- Question type annotations: Include question type annotations relevant to educational purposes. This allows controlled question generation and systematic training for models to produce specific types of questions accurately.

Context

Based on the reviewed literature, we found that many types of context are used in NQG studies. Here, we categorize the most prominent contexts into three groups: natural text, structured text, and images. Table 4 summarizes the representative methods applied to each category.

Most NQG studies use natural language texts, including individual sentences, paragraphs, or multi-paragraphs. However, longer inputs can pose challenges as they contain both relevant and irrelevant information, leading to specificity issues in question generation. Researchers have proposed various techniques to address this, such as modifying RNN-based models, using multiple stages of encoding, and employing transformer models, which have shown better results than RNN-based approaches.

Structured text includes formats like tables, formulas, and SQL queries. For example, some studies use tables accompanied by natural text as context, converting them to flat text for input into NQG models. An example of such conversion is provided by Fig. 3. In the case of formula data, Some studies converted formulas using templates before they were used as input to NQG. SQL queries can be inputted directly or transformed into templates. Although SQL is the primary structured language used in NQG, other programming languages, like Python, can also be explored for question generation tasks.

In addition to Visual Question Answering (VQA) research, QG from natural images has been studied. Methods for processing images as contexts in NQG include using outputs of specific layers of deep learning architecture models, or using natural text generated by image captioner models. However, there is limited research on technical images like diagrams and charts in AQG literature. Datasets with non-natural images, such as TQA, FigureQA (Kahou et al., 2017), PlotQA (Methani et al., 2019), and ChartQA (Masry et al., 2022), provide opportunities for future development in NQG models that involve technical images.

The following section presents the challenges and currently available solutions for processing educational context to produce valuable questions.

Key Content Identification from Long Contexts

Complex questions are usually generated from long contexts, for example, in the case of the HotpotQA dataset (Yang et al., 2018), where the context consists of several

Table 4 Types of context and processing methods

Context	Types	Representation methods
Natural text	Paragraph	Modified copy and attention mechanism (Zhao et al., 2018)
		Multi-stage processing (Chan and Fan, 2019; Tuan et al., 2020)
		Dynamic vocabulary (Kumar et al., 2020)
		Transformer (Lopez et al., 2021)
	Multi-paragraph / document	Semantic graph (Pan et al., 2020; Jia et al., 2021; Cao and Wang, 2021)
		Key content selection (Steuer et al., 2021)
Structured information	Table	Linearization (Pandraju and Mahalingam, 2021)
	Formula	Equation template (Zhou and Huang, 2019; Cao et al., 2021)
		Symbolic graph (Liu et al., 2021)
	Structured language (SQL)	RNN-encoder (Guo et al., 2018)
		Template (Yu and Jiang, 2021)
Images		Encoding using CNN-based model VGG Net (Mostafazadeh et al., 2016), CNN (Krishna et al., 2019) Image captioner model (White et al., 2021)

related paragraphs. Questions were then created such that they required readers to look at several paragraphs to answer them correctly. This is commonly referred to as *multi-hop* questions. In such cases, limiting or cutting out context merely based on

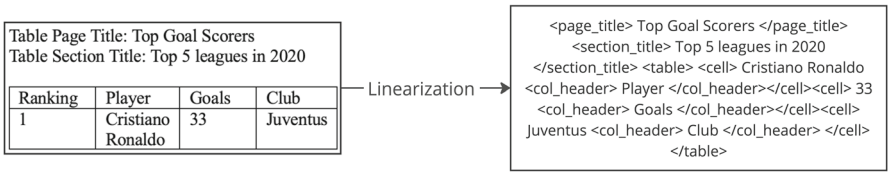

Fig. 3 Illustration of table conversion into flat text

individual sentences or paragraphs may limit the type and complexity of questions that can be generated, ultimately limiting its usefulness in learning.

In the field of education, particularly in the task of generating questions from reading material, the text involved is often longer than a single paragraph. This is especially evident in datasets such as TQA and LearningQ, where the lessons encompass an average word count exceeding 1000 words. Consequently, research that incorporates inputs spanning multiple paragraphs is crucial in order to effectively address the challenges posed by longer educational texts and ensure the generation of accurate and relevant questions. The two main techniques used in the literature entail representing context into a structured representation, such as a semantic graph (Cao and Wang, 2021; Jia et al., 2021; Pan et al., 2020), or changing to a shorter text by segmentation (Steuer et al., 2021), summarization (Dugan et al., 2022), or content selection (Du and Cardie, 2017). To take different parts of a text that have certain relationships (e.g., comparison, elaboration, evidence, etc.), discourse analysis can also be applied, but this is still rarely found due to the complexity of conducting discourse analysis, especially at the document level (Desai, 2021).

Long text inputs are challenging not only for NQG but also in other fields of NLP. In these cases, solutions that have succeeded in other fields may also be applied to NQG. For example, apart from changing the context, it is interesting to investigate the use of transformer models capable of accepting longer inputs (e.g., 4000 tokens), such as Longformer (Beltagy et al., 2020) and Big Bird (Zaheer et al., 2020).

Multimodal and Multi-domain Question Generation

- *Multimodal*. The context of the education domain usually consists of several modalities (e.g., tables, natural images, diagrams, and videos). For example, the TQA dataset includes text, natural images, and diagrams, whereas the SciQ dataset (Welbl et al., 2017) only takes context in the form of text during crowdsourcing, even though the original sources are various school and college-level science books, most of which also contain images and diagrams. While the use of multimodal approaches is crucial for the widespread adoption of QG systems in the field of education, there is currently a limited number of NQG studies that incorporate more than one modality as context. However, considering the prevalence of multimodal Question Answering (QA) studies (Ilievski and Feng, 2017; Singh et al., 2021) and the increasing availability of multimodal datasets such as TQA and MultimodalQA (Talmor et al., 2021), it is expected that future studies will delve into multimodal QG research in greater detail.
- *Multi-domain*. To generate fluent, specific, and useful questions, a specific vocabulary, and language model are usually needed for each different subject. Although syntactic patterns of questions may be similar, the types of target answers and the syntactic structure used can be different between subjects (Yu and Jiang, 2021). For this reason, particular datasets for each subject are needed to produce specific questions. Several datasets and research on various subjects use texts from the world of education, among others: TQA (life science, earth science, physical science), SciQA (physics, chemistry, biology), (Cao and Wang, 2021) (science, history, political science), (Steuer et al., 2021) (anatomy, biology chemistry, physics,

psychology, sociology), and Kulshreshtha et al. (2021) (machine learning, biomedical).

Although some datasets cover a variety of subjects (e.g., Biology, Chemistry), there are still limited studies that explicitly consider the subject as an integral part of the question generation model. In many cases, the existing literature focuses on general question generation without explicitly addressing specific subjects or domains. Instead, subject relevance is only implicitly inferred from the content of the context (Willis et al., 2019). Further research is needed to explore and incorporate subjects into question generation models in order to enhance the applicability and effectiveness of NQG systems in various educational domains. Moreover, in scenarios where labeled datasets are available for certain subjects but not for others, techniques such as *domain adaptation* can be employed to transfer the knowledge acquired from labeled datasets to unlabeled datasets in different subjects (Kulshreshtha et al., 2021).

Incorporating External Knowledge

In addition to utilizing information from the input context, several studies have also shown that integrating external knowledge can improve the quality of the questions generated. The following example illustrates ways in which external knowledge can be used to increase the clarity of questions.

- **Context**: Gelora Bung Karno is a football stadium located in Jakarta
- **Target Answer**: Jakarta
- **Question (only from context)**: Where is Bung Karno Stadium located?
- **Question (with external knowledge ⟨Jakarta, is-a, city⟩)**: In which city is the Bung Karno Stadium located?

Xin et al. (2021) empirically showed that general domain external knowledge, namely Conceptnet (Speer et al., 2016) and Wordnet (Miller, 1995) could be integrated to improve the quality of questions developed. One reason for integrating external knowledge is that questions developed by humans from context may use some synonyms of the words in the context. Another motivation for using external knowledge is to overcome rare and unknown words in domains with limited training data (Delpisheh, 2020).

For the purposes of education, external knowledge can help to increase the complexity of the generated questions. This increase relates to the inference process that can arise by replacing words with synonyms or harnessing common sense reasoning. Furthermore, using domain-specific external knowledge will enrich various aspects of questions, such as introducing new related facts (Lewis et al., 2020) that can stimulate exploration. Some examples of domain-specific knowledge bases are provided by Abu-Salih (2020). More information regarding the integration of knowledge that focuses on text generation is provided by Yu et al. (2022).

Answer

Longer context allows more and different questions to be generated. The use of answers as input to QG systems can lead to the generation of more specific and contextually relevant questions. As a result, the generated questions can be more informative and precise. QG systems that use answers as one of their inputs are also termed *answer-aware* QGs; this is the most widely used approach in the reviewed literature. However, since answers are not always available, some studies aim to generate answers through *answer extraction* or *answer generation* approaches; these are then used as inputs to answer-aware QGs. When pairs of answers and questions are produced at the same time, these are termed *joint question-answer generation* (Cui et al., 2021). Finally, there is an option not to use the answer at all (Du et al., 2017).

In addition to the availability of answers, an answer's characteristics can also influence the model and the types of questions generated. Characteristics of answers can be divided based on several dimensions, including:

- *Extractive vs abstractive*. Extractive answers comprise one or more word spans selected from the context such that the words and their orders are the same as in the context. If the answer is generated by humans or systems such that the words and their arrangement may differ from the context, it is considered abstractive.
- *Inside vs outside*. Whether or not the answer can be found by simply understanding the context. An example of an answer that cannot be found in context is when looking for new information, such as an exploratory question on a topic or an interview question.
- *Short vs long*. Short answers consist of a few words or phrases, whereas long answers can consist of one or more sentences. The length of the answer can provide information about the types of questions that are more likely to be generated. For example, based on Graesser and Person (1994), questions involving verification, disjunctive, concept completion, feature specification, and quantification usually expect short answers. On the other hand, longer answers are more likely for questions with types of example, definition, comparison, interpretation, causal antecedent, causal consequence, goal orientation, instrumental, procedural, enablement, expectation, and judgment.

Generation from Long and Abstractive Answers

NQG research with short and extractive answers, often called factoids, has been widely conducted; however, for educational purposes, long, and abstract answers are more desirable. Few studies have used long and abstract answers. One obstacle facing such approaches is the limited dataset, considering that long and abstract answer annotations are difficult to produce effectively and require significant effort (Cao and Wang, 2021; Mishra et al., 2020).

A dataset whose answers are abstract and long with high-level question types is ELI5 (Fan et al., 2019), compiled from the QnA Reddit forum, to which contexts from various documents related to questions and answers have been added. The average length of the answers in ELI5 is 6.6 sentences or 130 words. Unfortunately, no NQG

literature from our review list has utilized this dataset. Cao and Wang (2021) also used Reddit's QnA forum as a dataset source, so the answers are also abstract and lengthy. However, the created dataset was not accompanied by context; therefore, the questions are based only on the answers. A similar approach was performed by Mishra et al. (2020) who used long answers from the Natural Questions dataset as input without any supporting documents as contexts.

In contrast, certain studies have adopted a combination of diverse datasets with different types of answer formats. Murakhovs'ka et al. (2021) demonstrated that models trained on various datasets, each containing different types of answers, yielded improved performance compared to models trained on a single dataset. Meanwhile, (Yuan et al., 2022) developed a model that could adapt to different types of answers without requiring modifications to the underlying model architecture. They further applied a lifelong learning approach, enabling the model to continually learn and incorporate new information over time. These studies highlight the benefits of incorporating diverse datasets and accommodating various answer types in improving the performance and adaptability of question generation models.

Answer Extraction/Generation

It would be more practical if QG systems not only generated questions but also their respective answers. Thus, the only input required would be a document without the need for the user to input answers. For example, in education, the input can be in the form of a lesson, and the system can output several pairs of questions and answers with pedagogical value. In addition to increasing the specificity of the question, determining the answer before constructing the question can also increase the *explainability* of the QG system. Observers can see the output at each stage in order to facilitate the evaluation and improvement of the system. Especially in the world of education, the selection of answers is an important stage because of the need to identify useful answers for the learning topics, not just any answers from which questions can be made.

Most studies used extraction, i.e., taking words from a context. Only one study generated an answer whose words may differ from the context (Willis et al., 2019). All identified literature produced relatively short answers, namely one or more phrases that may not be consecutive or in the form of a span that usually predicts the start and end indexes. Answers can also be generated through supervised learning if a dataset with answer annotations is available; if this is not the case, one can nonetheless take advantage of unsupervised learning using linguistic rules and heuristics. Finally, few studies have extracted, or generated answers that pay attention to pedagogical aspects, especially answers that are suitable for high-level questions.

Input Features

This subsection explains the input features used in various NQG systems. Understanding the various input features used in NQG systems is essential for optimizing question generation performance.

- *Word Features*. This is the main input feature in the QG system, commonly utilizing pre-trained word embeddings that transform words or sentences into numerical vectors.
- *Linguistics Features*. Various linguistic features can also empirically improve the performance of a QG system which is usually measured by an increase in automatic metrics, such as BLEU. Linguistic features include *parts of speech*, *named entities*, and the *word case* (Zhou et al., 2017).
- *Coreference*. An input context of several sentences or paragraphs may consist of pronouns. To relate information about the same entity and support the production of complex questions, coreference can be an important part of input features. For example, as done by Du and Cardie (2018), they inserted antecedents of each pronoun into the input text and then added the coreference position feature to inform the location of a pronoun and its antecedent.
- *Answer Information*. Information about target answers is essential to generate specific questions. Several methods can be used, among others, by providing an additional feature in the form of an answer position if the answer is part of the input context. The answer position can be represented by a BIO scheme or by the relative distance of each word to the answer words (Sun et al., 2018). For answers contained in several different sentences, one can add an answer information feature in the form of node embeddings from a graph encoded by a *graph neural network* (Cheng et al., 2021). Another option is to separate answers from the input context, encode the answer with a different encoder, and replace or mask the answer words in the input context using a special token (Kim et al., 2019). This process helps to avoid the resulting questions containing the answer words. If using a pre-trained model, the answer sequence is added after the input context sequence with a special token as a separator, for example, a [SEP] token (Chan and Fan, 2019).
- *Graph*. Graph structure can add semantic information that ordinary word sequences cannot represent. Several types of graph representation have been used for long inputs and to produce deep questions where answers are obtained from several parts of the text, including dependency and semantic role labeling (Pan et al., 2020). After building a graph consisting of nodes and edges, encoding is done using a graph neural network variant to obtain the final node embeddings, which are then used as an input feature. Several studies have extracted ⟨subject, relation, object⟩ information to build a reasoning chain to select important and related content from various parts of a text, thereby developing complex questions (Cheng et al., 2021; Wang et al., 2020).

Question

Questions for reading comprehension were the most frequently found in the studies reviewed herein; however, there are actually many different types of situations and objectives in which different types of questions might be used, for example clarifying questions (Cao et al., 2019), pure information-seeking (Qi et al., 2020), questions to increase interaction (Wang et al., 2018), interview questions (Su et al., 2018),

CV screening questions (Shi et al., 2020), and questions with pedagogical objectives (Krishna and Iyyer, 2019).

Relatively few studies on NQG have mentioned that the questions they use are related to pedagogical purposes. Here, we try to summarize the types of questions related to the question function for education as described in Section 2, namely to recall prior knowledge, test comprehension, and hone critical thinking. The following categorizations are not mutually exclusive.

- *Factoid questions* are the most prevalent type of questions in Question Generation (QG) studies. They are designed to retrieve specific factual information, such as names, locations, or time-related details. Simple factoid questions can be generated by selecting a concise answer from the given context and rearranging the words in the context to form a question that targets the chosen answer. This approach allows for the creation of straightforward questions that directly inquire about the selected information. Although there are many rule-based methods for generating this type of question, the use of neural methods shows better results (Du et al., 2017). As a state-of-the-art model on the factoid dataset SQuAD, ERNIE-GEN (Xiao et al., 2019), a pre-trained seq2seq transformer model, showed the best BLEU4 score. A comparison between various models for the SQuAD dataset is provided in Zhang et al. (2021).
- *Multi-hop questions*. Another popular type of question in NQG research, in addition to simple factoid questions, is the multi-hop questions. These have become better explored since the emergence of the HotpotQA dataset (Yang et al., 2018). The essence of this type of question is that it requires connecting different parts of the text to obtain an answer, sometimes referred to as *text-connecting*. Text-connecting is a form of inference applied to questions, enabling the generation of questions that necessitate the integration and synthesis of information from multiple passages or contexts (Chikalanga, 1992). A common step to generate this type of question is to represent the input in a more structured form, such as a graph. Then, the graph is encoded with a variant of graph neural networks (GNNs). Next, some content from the graph is selected as suitable for generating multi-hop questions (Pan et al., 2020). Cheng et al. (2021) constructed multi-hop questions iteratively, from 1-hop to n-hop, using two types of decoders. These types of questions can help to improve comprehension skills. Unfortunately, although the questions require a text-connecting process, current research only deals with short answers so that they can be classified as factoid questions.
- *Non-factoid questions*. Non-factoid questions usually require long and abstract answers. Although supporting datasets, such as MS MARCO, and RACE, can be used, NQG studies focusing explicitly on generating non-factoid questions remain limited and are mainly implicitly learned from datasets. Current challenges involve the generation of abstract answers and the effective use of long abstract answers (Qu et al., 2021).
- *Question types from cognitive science*. Questions related to educational goals focus on high-level questions that can stimulate critical and abstract thinking. Such questions cannot be determined only from the question words. For example, a question starting with "why" is not necessarily high-level. The answer may be

easily found by considering only one cause-and-effect sentence. For this reason, it is necessary to use a more specific classification, for example, those from cognitive science as proposed by Graesser et al. (2010). Cao and Wang (2021) used ten types of questions derived from cognitive science and explicitly attempted to generate questions for each type by adding templates and examples as inputs. Krishna and Iyyer (2019) divided questions into specific and general types, before controlling the question type based on the length of the given answers. Steuer et al. (2021) focused on generating one type of question, definitional questions, by choosing sentences that contain definitions using rules and heuristics.

System Objectives

The development of the NQG system through each of its components has been discussed above, namely Context, Answer, and Question. This section will discuss what the objectives of the NQG system aim to achieve in various studies, together with the approaches, and methods used. These objectives also provide the basis for determining the evaluation mechanism for the NQG system. The objectives of the NQG system can be divided into five, namely: naturalness, usefulness, diversity, controllability, and personalization. In addition, we arrange the order according to their priorities for educational applications.

Naturalness

Most reviewed studies addressed objectives in fluency, relevance, and answerability. Considering that these are basic objectives that are almost always present, we have merged them into *naturalness* for simplicity. Fluency is related to the use of appropriate language, relevance requires that the question relates to the context, and answerability requires that the question can be answered from the context provided, or, if the answer is provided, the suitability of the answer is measured. Answerability may be optional, for example, in pure information seeking. Feature engineering, such as using linguistic features (e.g., NER, PoSTag) or replacing certain words with placeholders, can also help to improve naturalness. Another approach is to select or generate more context-specific information to help generate questions more relevant to both context and answers. Using external knowledge can help to paraphrase questions (in the manner of humans) such that they are more natural and readily understood. Many studies also made improvements to the encoding and decoding processes. Finally, other components may be added, such as the integration of Reinforcement Learning, multi-task learning, question rankers, and others.

Usefulness

In addition to naturalness, questions related to education aim for information that is not only relevant to their context but also central to the topic of the context in order to be useful for learning. Several studies have demonstrated the importance of learning objectives to help students and teachers focus on important material

(FitzPatrick et al., 2015; Osueke et al., 2018; Laneuville and Sikora, 2015). Horbach et al. (2020) also found that questions generated by domain experts were more useful for learning (evaluated by the complexity of deriving answers, targeting central information, and applicability) than those generated by crowdsourcing.

Chen et al. (2019) evaluated nine strategies for choosing important sentences, such as sentences at the beginning of the paragraph, sentences with novel words that are not found in other sentences, and sentences with low similarity (e.g., the most different) to other sentences. Heuristics were also used by Yao et al. (2022) to select keywords that are important for understanding stories, such as characters, feelings, and causal relationships. In addition to using rule-based approaches, sentences can also be chosen by models trained on datasets labeled by assuming that sentences containing answers are important sentences (Du and Cardie, 2017).

The above methods do not use external information to guide the selection of important content. As such, the model either treats all possible inputs from different topics equally or only distinguishes them implicitly through knowledge based on training data. For contexts derived from educational materials, the presence of learning objectives in each lesson can be used as external information to increase usefulness. These learning objectives can be used at the content selection stage, as in Shimmei and Matsuda (2021), or determine answers, and construct questions simultaneously such that both of them are in accordance with the learning objectives. In addition, as an alternative, one can use the words in the index textbook to denote the most important concepts (Steuer et al., 2021).

Diversity and Controllability

Zhang and Zhu (2021) split diversity into two categories, namely global, and local. Global diversity is the measured diversity of all outputs produced by the NQG system from all contexts, whereas local diversity is focused on the variety of questions generated from the same context. Consequently, the system can output more than one question for the same context. In general, diversity, and specificity are interrelated. When a question is very specific to the context, the diversity will increase, and *vice versa*.

To ensure global diversity, (Yu and Jiang, 2021) utilized a variety of question templates to avoid generating questions that are too general, and similar. Wang et al. (2018) aimed to increase the specificity of the questions generated through a typed decoder, with the objective of having the resulting questions include the topic word presented in the context. Topic words are crucial elements that contribute to the specificity and focus of a question. These words help to identify and target the key subject or topic of the question. On the other hand, interrogative words (such as "what," "where," "when," etc.) and ordinary words (like articles and prepositions) play a complementary role in ensuring grammaticality and coherence in the question structure.

The most popular method to produce local diversity is to vary additional inputs, and then generate a question for each different additional input. Additional inputs can be varied, including their target answers, keywords, question words, and additional contexts. Another method is through the model itself, for example, by sampling during inference (Sultan et al., 2020), or during training through a variational decoder.

In addition to being able to generate multiple diverse questions from a single context, introducing a mechanism to control the type of questions generated by a system can greatly enhance its practicality. Rather than generating all possible questions, the system can be tailored to produce specific types of questions as needed. Based on the studies reviewed, such control can be achieved through various approaches. One approach is by inputting a question word or question type, which guides the system to generate questions of a particular category. Another aspect that can be controlled is the difficulty level of the questions, wherein a difficulty level value can be specified to ensure the generation of questions that match the desired level of complexity.

The definition of difficulty levels varies across different studies. Several approaches have been employed to determine the difficulty of a question:

- *Answerability*: Some studies assess difficulty based on whether a QA system can correctly answer the question. If a question is challenging for a QA system to answer accurately, it is considered more difficult (Gao et al., 2019).
- *Textual complexity*: Difficulty can be determined by the number of different parts or passages of text that need to be considered to answer the question. Questions that require integrating information from multiple sources or passages are generally considered more difficult (Cheng et al., 2021).
- *Answer length*: Another criterion for assessing difficulty is the length of the answer. Questions with longer and more detailed answers may be considered more difficult than those with short and concise answers (Murakhovs'ka et al., 2021).

By incorporating these control mechanisms, NQG systems can provide more targeted and customized question generation, catering to specific requirements and educational objectives.

Personalization

Developments related to personalized learning also include personalized assessment (Bulger, 2016; Baylari and Montazer, 2009), where questions are tailored to the needs of each student. Thus, systems that select or generate questions must consider information about each student, such as their background, or learning progress.

Few NQA studies have directly integrated the personal aspects of target users into the QG model. Srivastava and Goodman (2021) modeled the student state through historical sequences of questions and responses from the student, then used the student state and a certain difficulty level as inputs for QG. Stewart and Mihalcea (2021) modeled users in online question and answer forums based on three dimensions, namely expertise (i.e., expert vs. novice), response time (i.e., fast vs. slow), and also location (i.e., US vs. non-US). The QG model then received user categories *via* discrete or continuous inputs.

As an alternative, personalization can also be achieved through separate modeling between the QG model and the model used to predict or track the state of each user, provided that the QG model can generate different types of questions at once or can be controlled by inputting specific question categories.

Explainability

In the context of question generation, explainability refers to the degree to which the process of generating questions can be understood and interpreted by humans. It involves transparency and clarity in the mechanisms and decisions made by the question generation system, allowing users to comprehend how and why specific questions are generated. Explainability is a desirable objective for educational QG.

While we have not come across a paper specifically focusing on building and discussing the explainability of an NQG (Neural Question Generation) system, there are some general approaches that can be utilized. For instance, the use of a *saliency heatmap* (Danilevsky et al., 2020) based on *attention weight* can provide insights. Additionally, decomposing the QG process into multiple stages, each producing an intermediate output utilized by a question constructor, can enhance explainability. For example, the content-selection stage accepts long text as input to produce shorter text. Subsequent keyword identification processes highlight keywords in context and the answer extraction stage generates the target of the question. Finally, inputs in the form of control variables, such as question type and difficulty level, are considered. For example, (Cheng et al., 2021) showed that their system was more explainable by determining the number of hops as an input to control the difficulty level of generated questions. All intermediate outputs and the control variables were then fed into a question construction stage to generate questions. By observing those detailed inputs, users can anticipate the kinds of questions that should be generated and evaluate the quality of the systems by comparing the results and expectations.

Design of QG Systems for Education Purposes

Having discussed all aspects of NQG for education purposes, we here illustrate the essential stages of the process by considering the objectives that currently matter most to educational applications: usefulness, diversity, and controllability. Personalization is indeed crucial, but achieving it relies on the ability to diversify and control the output of NQG. Regarding explainability, in addition to the difficulty in its production, many use cases can still derive benefits from the outputs without requiring a detailed explanation.

Figure 4 illustrates the three main processes relevant to education-oriented QG systems: key content identification, answer identification, and question construction. Neural-based methods, such as RNN-based or Transformer-based models, are commonly used for question construction. Key content identification can involve sentence selection, subgraph selection, or summarization, with attention to topic importance and the desired question type. To promote diversity, key content identification should generate multiple sub-contexts from the original context, based on different target answers or question types. Answer identification models can be extraction-based, employing rule-based or statistical methods, or generation-based, typically using neural techniques. Clues regarding topic importance, such as learning objectives or essential concepts, should ideally be provided. Finally, the question construction stage can

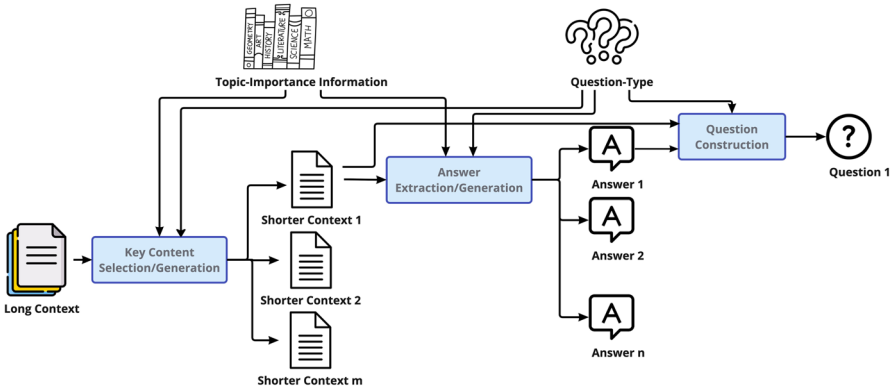

Fig. 4 Illustration of QG systems for education considering aspects of usefulness, diversity, and controllability

be controlled in terms of difficulty level or other specifications to ensure desired outcomes.

Evaluation

In this subsection, we will explore how the automated process of generating questions can impact the learning process, complementing the known usefulness of manually crafted questions. Additionally, we will discuss the evaluation methods used in existing studies, which typically involve assessing the quality of the NQG systems through automatic and human evaluations of the generated outputs.

Impact of Automated Question Generation on Learning

For quite some time, it has been established that questions, traditionally formulated by human experts, play a vital role in enhancing learning. These questions facilitate reflection, strengthen memory trace, and promote long-term retention of acquired information. However, manually generating questions requires considerable time and cognitive effort, often leading to questions being provided after larger sections of content rather than for each small part, like each page of reading. With the advancements in AQG, it becomes more feasible to generate questions efficiently and effectively. Here, we delve into the various ways automated question generation enhances the learning process.

- Active Engagement: AQG can be seamlessly integrated into websites or reading software, transforming the reading experience from passive to interactive. By generating questions on-the-fly based on the current page, readers are prompted to engage actively with the material (Syed et al., 2020).
- Personalized Learning: AQG allows for the customization of questions based on individual learning needs. Learners can receive questions tailored to their profi-

ciency level, enabling them to focus on areas that need improvement (Srivastava and Goodman, 2021).

- Promoting Critical Thinking: Finally, well-designed automated questions can prompt students to think critically, analyze information, and apply knowledge to real-world scenarios. Moreover, beyond answering the automated questions, students can also learn how to formulate critical questions themselves by observing the examples provided, leading to an enhancement in their ability to ask more insightful and effective questions (Hofstein et al., 2005).

Automatic Evaluation

Automated evaluation metrics commonly used in QG are adopted from other text generation tasks. The metrics include BLEU, METEOR, and ROUGE-L. These metrics only consider the lexical similarity between the model output and one or more references. Mathur et al. (2020) showed that an increase of 1-2 points in BLEU exhibits a positive correlation with an increase in human judgment only 50% of the time. Indeed, (Sultan et al., 2020) empirically demonstrated that systems producing lower BLEU or ROUGE-L scores were perceived better when the resulting questions were used in the downstream task (QA), indicating that the systems are more useful. To overcome this issue, (Kane et al., 2020) proposed a model-based automatic evaluation method for text generation that can measure similarity beyond the lexical level; however, this method assumes a robust neural language model, which does not necessarily exist for certain languages or domains.

Unfortunately, the abovementioned metrics are even more problematic for QG due to: (i) the lack of datasets with multiple references; and (ii) the one-to-many nature of QG, which means that several entirely different questions can be generated from one context. In order to overcome these two challenges, (Rodrigues et al., 2021) compiled a special corpus for evaluating the QG model, which has an average of 26 different questions for each context. The construction of such a dataset is highly useful and needs to be extended to other domains.

Human Evaluation

Due to the shortcomings of those automated metrics, many studies also report the results of manual evaluations by humans, which generally cover three main criteria: naturalness (fluency, relevance, and answerability), difficulty (Du et al., 2017; Chan and Fan, 2019; Bi et al., 2021), and helpfulness (Cheng et al., 2021; Sekulić et al., 2021). Unfortunately, the lack of uniformity in the methods employed makes it challenging to directly compare manual evaluations across different studies. The evaluation approaches often involve experts, crowdsourced participants, or the authors themselves, who assess the generated questions based on specific criteria and assign scores accordingly. Furthermore, some studies solicit preferences from evaluators regarding sets of questions from different systems, including comparisons with human-generated questions.

For educational purposes, (Horbach et al., 2020) introduced a tiered human evaluation scheme that includes a total of nine questions. The proposed scheme provides

a comprehensive assessment of the quality of the generated questions, taking into account their relevance to education. It incorporates two key measurements: the complexity involved in obtaining the answer and the importance of the questions in relation to the topic covered in the input context. This comprehensive evaluation approach provides valuable insights into the effectiveness of the generated questions in educational settings. They also highlighted the suitability of question datasets for educational purposes when created by domain experts. This approach ensures that the questions are curated with pedagogical expertise, contributing to their effectiveness in educational settings.

Conclusions and Future Work

This review has summarized the current literature concerning NQG and the application of NQG to education. We have described the development of NQG for educational purposes from various aspects such as context, answer, question, and evaluation. We have also discussed the challenges remaining in the current state-of-the-art. It can be concluded that the development of neural methods for QG has produced fluent and related questions from a limited length of context but has not yet become fully capable of producing high-level questions that are useful in educational contexts. Improvements to key content identification for long contexts and controllability of the type of questions generated will also increase usefulness. For more details, in the following, we present perspectives on future research:

More Specialized Datasets

Datasets specifically oriented toward QG for education are still very limited. Most large datasets are created for answering questions that focus on comprehension. There is a need for datasets that focus on high-level questions with annotations that make the learning process more straightforward using current deep learning methods. Although some datasets contain high-level questions, they have not been labeled with specific types.

Another direction for improvement lies in enhancing the effectiveness of automatic evaluation. To achieve this, the availability of more comprehensive datasets is crucial. For instance, datasets that include multiple reference questions for each context can be instrumental. In an ideal scenario, these datasets would encompass all possible questions that can be generated from a given context. By having a diverse range of reference questions, the evaluation process can better capture the quality and coverage of generated questions, leading to more reliable and informative assessments of NQG systems.

Identifying Key Content From Long and Diverse Contexts

Natural text, i.e., short text up to a maximum of one paragraph has been explored in great detail. Methods of identifying key content in much longer inputs remain to

be explored. Effective methods to achieve this should go beyond simply selecting a sentence, instead, they should combine various related information from different sentences or even paragraphs that lead to particular question types. One possible solution involves defining important elements of each question type. For example, text related to procedural questions can include objectives, methods, and equipment, while comparison questions may require objects being compared and the aspect of comparison. By extracting this information in a structured manner and utilizing graph-based techniques to store the information, additional modifications like merging similar entities, even from distant parts, can facilitate text-connecting reasoning-type of question. In addition, selecting specific graph paths can increase question specificity (e.g., asking for the equipment only), while still targeting meaningful information.

Generating Abstractive, Long, and Useful Answers

Currently, most target answers are extractive and short; in contrast, approaches producing both abstract and long answers are relatively few, including answers compiled from different parts of a text. Furthermore, identifying target answers that are relevant to the context and important to the topic also requires further research.

Combining Several Aspects of Questions

Some studies involve questions from cognitive science, in particular high-level questions. Studies on multi-hop but factoid questions are also common. Research that combines different aspects of questions will be able to improve the quality of the questions generated. Controllability is also crucial for improving the preciseness of model evaluation because it narrows down the candidate answer references that must be compared by automatic evaluation and improves accessibility during manual evaluation.

Limitations

The literature review has certain limitations. Firstly, the possibility of publication bias, potential exclusion of studies not accessible through selected databases, and variations in study quality that may impact the overall robustness of findings. Secondly, the review briefly mentions visual question generation but doesn't extensively explore specific methods for visual input. This indicates the need for a separate review focused solely on visual question generation. Lastly, the review identifies current limitations of deep learning techniques in question generation, such as challenges in ensuring explainability and handling abstractive answers, without proposing specific approaches to address them. To enhance deep learning-based question generation, future research should concentrate on innovative methodologies and techniques to overcome these challenges.

Funding This work was supported by Direktorat Riset, Teknologi, dan Pengabdian Kepada Masyarakat, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia (Grant number 156/E5/PG.02.00.PT/2022) .

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

References

Abu-Salih, B. (2020). Domain-specific knowledge graphs: A survey. arXiv:2011.00235

Ali, H., Chali, Y., Hasan, S.A. (2010). Automatic question generation from sentences. In *Actes de la 17e conférence sur le traitement automatique des langues naturelles. articles courts* (pp. 213–218). ATALA

Amidei, J., Piwek, P., Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th international natural language generation conference* (pp. 307-317). Association for Computational Linguistics

Bai, Y., Wang, D.Z. (2021). More than reading comprehension: a survey on datasets and metrics of textual question answering. arXiv:2109.12264 [cs.CL]

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., et al. (2016). MS MARCO: a human generated MAchine reading comprehension dataset. arXiv:1611.09268

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist. *Reading and Writing, 26*(3), 349–379.

Baylari, A., & Montazer, G. A. (2009). Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications, 36*(4), 8013–8021.

Beltagy, I., Peters, M.E., Cohan, A. (2020). Longformer: The long-document transformer. arXiv:2004.05150 [cs.CL]

Bi, S., Cheng, X., Li, Y.-F., Qu, L., Shen, S., Qi, G., Jiang, Y. (2021). Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 4645-4654). Association for Computational Linguistics

Bloom, B.S. (1956). Taxonomy of educational objectives : the classification of educational goals. *Cognitive Domain*

Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question selfgeneration and answering for remembering expository text. *Journal Education Psychology, 104*(4), 922–931.

Bulger, M. (2016). Personalized learning: The conversations we're not having. *Data and Society, 22*(1), 1–29.

Cao, S., & Wang, L. (2021). Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (vol. 1: Long papers)* (pp. 6424–6439). Association for Computational Linguistics

Cao, T., Zeng, S., Zhao, S., Mansur, M., Chang, B. (2021). Generating math word problems from equations with topic consistency maintaining and commonsense enforcement. In *Artificial neural networks and machine learning - ICANN 2021* (pp. 66–79). Springer International Publishing

Cao, Y.T., Rao, S., Daumé, H., III. (2019). Controlling the specificity of clarification question generation. In *WNLP@ ACL* (pp. 53–56). Association for Computational Linguistics.

Chan, Y.-H., & Fan, Y.-C. (2019). A recurrent BERT-based model for question generation. In *Proceedings of the 2nd workshop on machine reading for question answering* (pp. 154–162). Association for Computational Linguistics

Chen, G., Yang, J., Gasevic, D. (2019). A comparative study on Question-Worthy sentence selection strategies for educational question generation. In *Artificial intelligence in education*(pp. 59–70). Springer International Publishing.

Chen, G., Yang, J., Hauff, C., Houben, G.-J. (2018). Learningq: A largescale dataset for educational question generation. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12(1)

Cheng, Y., Ding, Y., Pascual, D., Richter, O., Volk, M., Wattenhofer, R. (2021). WikiFlash: Generating flashcards from wikipedia articles. In *AAAI 2021 workshop on AI education-35th AAAI conference on artificial intelligence (AAAI)*.

Cheng, Y., Li, S., Liu, B., Zhao, R., Li, S., Lin, C., Zheng, Y. (2021). Guiding the growth: Difficulty-Controllable question generation through Step-by- Step rewriting. Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (vol. 1: Long papers) (pp. 5968–5978)

Chikalanga, I. (1992). A suggested taxonomy of inferences for the reading teacher. *Reading in Foreign Language, 8*(2), 697–709.

Chin, C., & Osborne, J. (2008). Students' questions: a potential resource for teaching and learning science. *Studies in Science Education, 44*(1), 1–39.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-T., Choi, Y., Zettlemoyer, L. (2018). QuAC: Question answering in context. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2174–2184). Brussels, Belgium: Association for Computational Linguistics

Cui, S., Bao, X., Zu, X., Guo, Y., Zhao, Z., Zhang, J., Chen, H. (2021). OneStop QAMaker: extract question-answer pairs from text in a one-stop approach. arXiv:2102.12128 [cs.CL]

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P. (2020). A survey of the state of explainable AI for natural language processing. *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing* (pp. 447–459). Association for Computational Linguistics.

Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. (2021). Automatic question generation and answer assessment: a survey. *Research and Practice in Technology Enhanced Learning, 16*(1), 1–15.

Delpisheh, M. (2020). *Neural Question Generation with Transfer Learning and Utilization of External Knowledge (Unpublished doctoral dissertation)*. Toronto: York University.

Desai, T. (2021). Discourse parsing and its application to question generation (Unpublished doctoral dissertation). The University of Texas at Dallas.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. In *Advances in neural information processing systems* (vol. 32). Curran Associates, Inc.

Du, X., & Cardie, C. (2017). Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2067–2073)

Du, X., & Cardie, C. (2018). Harvesting paragraph-level Question-Answer pairs from Wikipedia. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (vol 1: Long papers) (pp. 1907–1917)

Du, X., Shao, J., Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension.In *Proceedings of the 55th annual meeting of the association for computational linguistics (vol. 1: Long papers)* (pp. 1342–1352). Association for Computational Linguistics.

Dugan, L., Miltsakaki, E., Upadhyay, S., Ginsberg, E., Gonzalez, H., Choi, D., Callison-Burch, C. (2022). A feasibility study of Answer-Unaware question generation for education. In *Findings of the association for computational linguistics: ACL 2022* (pp. 1919–1926)

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., Auli, M. (2019). ELI5: Long form question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3558-3567). Association for Computational Linguistics

FitzPatrick, B., Hawboldt, J., Doyle, D., & Genge, T. (2015). Alignment of learning objectives and assessments in therapeutics courses to foster higher-order thinking. *American Journal of Pharmaceutical Education, 79*(1), 10.

Gao, Y., Bing, L., Chen, W., Lyu, M., King, I. (2019). Difficulty controllable generation of reading comprehension questions. In *Proceedings of the Twenty-Eighth international joint conference on artificial intelligence* (pp. 4968–4974). California: International Joint Conferences on Artificial Intelligence Organization

Gao, Z., Xia, X., Grundy, J., Lo, D., & Li, Y.-F. (2020). Generating question titles for stack overflow from mined code snippets. *ACM Transactions on Software Engineering and Methodology, 29*(4), 1–37.

Graesser, A., Ozuru, Y., Sullins, J. (2010). What is a good question? M.G. McKeown (Ed.), Bringing reading research to life , (pp (vol. 320, pp. 112–141). New York, NY, US: Guilford Press, xvi

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*(1), 104–137.

Guo, D., Sun, Y., Tang, D., Duan, N., Yin, J., Chi, H., Zhou, M. (2018). Question generation from SQL queries improves neural semantic parsing. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1597–1607). Association for Computational Linguistics

Guo, H., Pasunuru, R., Bansal, M. (2018). Soft Layer-Specific Multi-Task summarization with entailment and question generation. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (vol. 1: Long papers) (pp. 687–697). Association for Computational Linguistics.

Heilman, M., & Smith, N.A. (2010). Good question! statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 609–617). Association for Computational Linguistics.

Hofstein, A., Navon, O., Kipnis, M., & Mamlok-Naaman, R. (2005). Developing students' ability to ask more and better questions resulting from inquirytype chemistry laboratories. *Journal of Research in Science Teaching, 42*(7), 791–806.

Horbach, A., Aldabe, I., Bexte, M., de Lacalle, O.L., Maritxalar, M. (2020). Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In *Proceedings of the 12th language resources and evaluation conference* (pp. 1753–1762).

Ilievski, I., & Feng, J. (2017). Multimodal learning and reasoning for visual question answering. In *I. Guyon et al. (Eds.), Advances in neural information processing systems*(vol. 30). Curran Associates, Inc.

Jia, X., Zhou, W., Sun, X., Wu, Y. (2021). EQG-RACE: examination-type question generation. In *AAAI*. aaai.org.

Jouault, C., Seta, K., & Hayashi, Y. (2016). Content-Dependent question generation using LOD for history learning in open learning space. *New Generation Computing, 34*, 367–394.

Kahou, S.E., Michalski, V., Atkinson, A., Kadar, A., Trischler, A., Bengio, Y. (2017). FigureQA: An annotated figure dataset for visual reasoning. arXiv:1710.07300 [cs.CV]

Kane, H., Kocyigit, M.Y., Abdalla, A., Ajanoh, P., Coulibali, M. (2020). NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st workshop on evaluating NLG evaluation* (pp. 28–37). Association for Computational Linguistics

Karpicke, J. D. (2012). Retrieval-based learning: active retrieval promotes meaningful learning. *Current Directions in Psychological Science, 21*(3), 157–163.

Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H. (2017). Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5376–5384)

Kim, Y., Lee, H., Shin, J., & Jung, K. (2019). Improving neural question generation using answer separation. *AAAI, 33*(01), 6602–6609.

Kispal, A. (2008). Effective teaching of inference skills for reading. literature Review. research report DCSF-RR031. ERIC

Krishna, K., & Iyyer, M. (2019). Generating Question-Answer hierarchies. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2321–2334). Association for Computational Linguistics

Krishna, R., Bernstein, M., Fei-Fei, L. (2019). Information maximizing visual question generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2008–2018)

Kulshreshtha, D., Belfer, R., Serban, I.V., Reddy, S. (2021). Back-Training excels Self-Training at unsupervised domain adaptation of question generation and passage retrieval. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7064–7078). Association for Computational Linguistics.

Kumar, V., Joshi, M., Ramakrishnan, G., Li, Y.-F. (2020). Vocabulary matters: A simple yet effective approach to paragraph-level question generation. In *Proceedings of the 1st conference of the Asia-*

Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing (pp. 781–785). Association for Computational Linguistics

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121–204.

Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 785–794). Association for Computational Linguistics.

Laneuville, O., & Sikora, D. (2015). Quantitative analysis of the usage of a pedagogical tool combining questions listed as learning objectives and answers provided as online videos. *Future Internet, 7*(2), 140–151.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kiela, D. (2020). Retrieval-Augmented generation for Knowledge-Intensive NLP tasks. In *Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F. & Lin H. (Eds.), Advances in neural information processing systems* (vol. 33, pp. 9459–9474). Curran Associates, Inc.

Liu, T., Fang, Q., Ding, W., Li, H., Wu, Z., Liu, Z. (2021). Mathematical word problem generation from commonsense knowledge graph and equations. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 4225–4240). Association for Computational Linguistics.

Lopez, L.E., Cruz, D.K., Cruz, J.C.B., Cheng, C. (2021). Simplifying Paragraph-Level question generation via transformer language models. In *PRICAI 2021: trends in artificial intelligence* (pp. 323–334). Springer International Publishing.

Masry, A., Long, D., Tan, J.Q., Joty, S., Hoque, E. (2022). ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022* (pp. 2263–2279). Association for Computational Linguistics

Mathur, N., Baldwin, T., Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4984–4997). Association for Computational Linguistics

Methani, N., Ganguly, P., Khapra, M.M., Kumar, P. (2019). PlotQA: reasoning over scientific plots. arXiv:1909.00997 [cs.CV]

Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM, 38*(11), 39–41.

Mishra, S.K., Goel, P., Sharma, A., Jagannatha, A., Jacobs, D., Daumé, H., III. (2020). Towards automatic generation of questions from long answers. arXiv:2004.05109 [cs.CL]

Mitkov, R., & Le An, H. (2003). Computer-Aided generation of Multiple-Choice tests. In *Proceedings of the HLT-NAACL 03 workshop on building educational applications using natural language processing* (pp. 17–22)

Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L. (2016). Generating natural questions about an image. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (vol. 1: Long papers) (pp. 1802–1813). Association for Computational Linguistics

Mostow, J., & Chen, W. (2009). Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 2009 conference on artificial intelligence in education: Building learning systems that care: From knowledge representation to affective modelling*(pp. 465–472). IOS Press

Murakhovs'ka, L., Wu, C.-S., Niu, T., Liu, W., Xiong, C. (2021). MixQG: Neural question generation with mixed answer types. arXiv:2110.08175 [cs.CL]

Nappi, J. S. (2017). The importance of questioning in developing critical thinking skills. *Delta Kappa Gamma Bulletin, 84*(1), 30.

Osueke, B., Mekonnen, B., Stanton, J.D. (2018). How undergraduate science students use learning objectives to study. *Journal of Microbiology and Biology Education 19*(2)

Pan, L., Xie, Y., Feng, Y., Chua, T.-S., Kan, M.-Y. (2020). Semantic graphs for generating deep questions. *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1463–1475). Association for Computational Linguistics.

Pandraju, S., & Mahalingam, S.G. (2021). Answer-Aware question generation from tabular and textual data using T5. *International Journal of Emerging Technologies in Learning 16*(18)

Qi, P., Zhang, Y., Manning, C.D. (2020). Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. arXiv:2004.14530 [cs.CL]

Qu, F., Jia, X.,Wu, Y. (2021). Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 2583–2593). Association for Computational Linguistics.

Reddy, S., Raghu, D., Khapra, M.M., Joshi, S. (2017). Generating natural language Question-Answer pairs from a knowledge graph using a RNN based question generation model. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 1, long papers* (pp. 376–385). Association for Computational Linguistics.

Renaud, R., & Murray, H. (2003). The effect of higher-order questions on critical thinking skills. In *Annual meeting of the american educational research association*

Rice University (1999). OpenStax. https://openstax.org/. (Accessed 1 June 2022)

Rodrigues, H., Nyberg, E., & Coheur, L. (2021). Towards the benchmarking of question generation: introducing the monserrate corpus. *Language Resources and Evaluation, 56*(2), 573–591.

Sekulić, I., Aliannejadi, M., Crestani, F. (2021). Towards Facet-Driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval* (pp. 167–175). Association for Computing Machinery.

Serban, I.V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., Bengio, Y. (2016). Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. *Proceedings of the 54th annual meeting of the association for computational linguistics* (volume 1: Long papers) (pp. 588–598). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1056

Shi, B., Li, S., Yang, J., Kazdagli, M.E., He, Q. (2020). Learning to ask screening questions for job postings. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 549–558). Association for Computing Machinery.

Shimmei, M., & Matsuda, N. (2021). Learning association between learning objectives and key concepts to generate pedagogically valuable questions. In *Artificial intelligence in education*(pp. 320–324). Springer International Publishing.

Singh, H., Nasery, A., Mehta, D., Agarwal, A., Lamba, J., Srinivasan, B.V. (2021). MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies*(pp. 5317–5332). Association for Computational Linguistics.

Speer, R., Chin, J., Havasi, C. (2016). ConceptNet 5.5: An open multilingual graph of general knowledge. arXiv:1612.03975 [cs.CL]

Srivastava, M., & Goodman, N. (2021). Question generation for adaptive education. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (volume 2: Short papers) (pp. 692–701). Association for Computational Linguistics.

Stasaski, K., Rathod, M., Tu, T., Xiao, Y., Hearst, M.A. (2021). Automatically generating Cause-and-Effect questions from passages. In *Proceedings of the 16th workshop on innovative use of NLP for building educational applications*(pp. 158–170)

Steuer, T., Filighera, A., Meuser, T., Rensing, C. (2021, October). I do not understand what I cannot define: Automatic question generation with Pedagogically-Driven content selection. arXiv:2110.04123 [cs.CL]

Steuer, T., Filighera, A., Rensing, C. (2020). Remember the facts? investigating Answer-Aware neural question generation for text comprehension. In *Artificial intelligence in education* (pp. 512–523). Springer International Publishing.

Stewart, I., & Mihalcea, R. (2021). How well do you know your audience? reader-aware question generation. arXiv:2110.08445 [cs.CL]

Su, M.-H., Wu, C.-H., Huang, K.-Y., Hong, Q.-B., Huang, H.-H. (2018). Follow-up question generation using pattern-based seq2seq with a small corpus for interview coaching. In *INTERSPEECH* (pp. 1006–1010). isca-speech.org.

Sultan, M.A., Chandel, S., Astudillo, R.F., Castelli, V. (2020). On the importance of diversity in question generation for QA. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5651–5656)

Sun, X., Liu, J., Lyu, Y., He, W., Ma, Y., Wang, S. (2018). Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3930–3939)

Sundbye, N. (1987). Text explicitness and inferential questioning: Effects on story understanding and recall. *Reading Research Quarterly, 22*(1), 82–98.

Syed, R., Collins-Thompson, K., Bennett, P.N., Teng, M., Williams, S., Tay, D.W.W., Iqbal, S. (2020). Improving learning outcomes with gaze tracking and automatic question generation. In *Proceedings of the web conference 2020* (pp. 1693–1703). New York, NY, USA: Association for Computing Machinery

Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Berant, J. (2021). Multimodalqa: complex question answering over text, tables and images. International conference on learning representations

Tofade, T., Elsner, J., & Haines, S. T. (2013). Best practice strategies for effective use of questions as a teaching tool. *American Journal of Pharmaceutical Education, 77*(7), 155.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., Suleman, K. (2016). NewsQA: A machine comprehension dataset. arXiv:1611.09830 [cs.CL]

Tuan, L.A., Shah, D., Barzilay, R. (2020). Capturing greater context for question generation. In *Proceedings of the aaai conference on artificial intelligence* (vol. 34, pp. 9065–9072).

Wang, A., Cho, K., Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th annual meeting of the association for computational linguistics*(pp. 5008–5020). Association for Computational Linguistics.

Wang, S., Wei, Z., Fan, Z., Huang, Z., Sun, W., Zhang, Q., Huang, X. (2020). PathQG: Neural question generation from facts. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 9066–9075). Association for Computational Linguistics.

Wang, Y., Liu, C., Huang, M., Nie, L. (2018). Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (volume 1: Long papers) (pp. 2193–2203). Association for Computational Linguistics.

Wang, Z., Lan, A., Baraniuk, R. (2021). Math word problem generation with mathematical consistency and problem context constraints. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 5986–5999). Association for Computational Linguistics.

Wang, Z., Lan, A.S., Nie, W., Waters, A.E., Grimaldi, P.J., Baraniuk, R.G. (2018). QG-net: a data-driven question generation model for educational content. In *Proceedings of the fifth annual ACM conference on learning at scale* (p. 7). ACM.

Welbl, J., Liu, N.F., Gardner, M. (2017). Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd workshop on noisy user-generated text* (pp. 94–106). Association for Computational Linguistics.

White, J., Poesia, G., Hawkins, R., Sadigh, D., Goodman, N. (2021). Opendomain clarification question generation without question examples. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 563–570). Association for Computational Linguistics.

Willis, A., Davis, G., Ruan, S., Manoharan, L., Landay, J., Brunskill, E. (2019). Key phrase extraction for generating educational Question-Answer pairs. In *Proceedings of the sixth (2019) ACM conference on learning @ scale* (pp. 1–10). Association for Computing Machinery.

Xiao, D., Zhang, H., Li, Y., Sun, Y., Tian, H., Wu, H., Wang, H. (2020). ERNIE-GEN: An enhanced multi-flow pre-training and finetuning framework for natural language generation. In *Proceedings of the Twenty-Ninth international joint conference on artificial intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Xin, J., Hao, W., Dawei, Y., Yunfang, W. (2021). Enhancing question generation with commonsense knowledge. In *Proceedings of the 20th chinese national conference on computational linguistics* (pp. 976–987). Chinese Information Processing Society of China.

Yang, Z., Hu, J., Salakhutdinov, R., Cohen, W. (2017). Semi-Supervised QA with generative Domain-Adaptive nets. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (vol. 1: Long papers) (pp. 1040–1050). Vancouver, Canada: Association for Computational Linguistics.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2369–2380). Association for Computational Linguistics.

Yao, B.,Wang, D.,Wu, T., Hoang, T., Sun, B., Li, T.J.-J., Xu, Y. (2022). It is AI's turn to ask humans a question: Question-Answer pair generation for children's story books. In *Proceedings of the 60th*

annual meeting of the association for computational linguistics (vol. 1: Long papers) (pp. 731–744). Association for Computational Linguistics.

Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2022). *A survey of Knowledge-Enhanced text generation*. Surv: ACM Comput.

Yu, X., & Jiang, A. (2021). Expanding, retrieving and infilling: Diversifying Cross-Domain question generation with flexible templates. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 3202–3212). Association for Computational Linguistics.

Yuan, W., Yin, H., He, T., Chen, T., Wang, Q., Cui, L. (2022). Unified question generation with continual lifelong learning. arXiv:2201.09696 [cs.CL]

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Ahmed, A. (2020, July). Big bird: Transformers for longer sequences. arXiv:2007.14062 [cs.LG]

Zhang, R., Guo, J., Chen, L., Fan, Y., & Cheng, X. (2021). A review on question generation from natural language text. *ACM Transactions on Information and System Security, 40*(1), 1–43.

Zhang, Z., & Zhu, K. (2021). Diverse and specific clarification question generation with keywords. In *Proceedings of the web conference 2021* (pp. 3501–3511). Association for Computing Machinery.

Zhao, Y., Ni, X., Ding, Y., Ke, Q. (2018). Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3901–3910).

Zhou, Q., & Huang, D. (2019). Towards generating math word problems from equations and topics. In *Proceedings of the 12th international conference on natural language generation* (pp. 494–503).

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., Zhou, M. (2017). Neural question generation from text: A preliminary study. In *Natural language processing and chinese computing* (pp. 662–671). Springer International Publishing.