Progress in Artificial Intelligence (2023) 12:1–32
https://doi.org/10.1007/s13748-023-00295-9

REVIEW

Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications

Nikahat Mulla[1] · **Prachi Gharpure**[2]

Received: 11 June 2022 / Accepted: 2 January 2023 / Published online: 30 January 2023

Abstract

Question generation in natural language has a wide variety of applications. It can be a helpful tool for chatbots for generating interesting questions as also for automating the process of question generation from a piece of text. Most modern-day systems, which are conversational, require question generation ability for identifying the user's needs and serving customers better. Generating questions in natural language is now, a more evolved task, which also includes generating questions for an image or video. In this review, we provide an overview of the research progress in automatic question generation. We also present a comprehensive literature review covering the classification of Question Generation systems by categorizing them into three broad use-cases, namely standalone question generation, visual question generation, and conversational question generation. We next discuss the datasets available for the same for each use-case. We further direct this review towards applications of question generation and discuss the challenges in this field of research.

Keywords Automatic question generation · Natural language generation · Natural language processing

1 Introduction

Automatic question generation (AQG) systems are those in which questions are generated based on a topic or idea or context in natural language from either a paragraph of text or images. Such systems are becoming more popular of late, with their requirement in machine reading comprehension applications, conversational systems, and even educational applications. The eventual goal of AQG systems is the capability to generate questions that are correct syntactically and semantically as well as meaningful in the context of the use-case. For instance, in some cases, the goal is to generate questions on a topic of interest or based on different spans of text in a passage [1]. On the other hand, in conversational systems, say, a question-asking bot, it is imperative to be consistent with the context of the conversation and also at the same time maintain the interest of the user in the conversation [2].

AQG has been widely experimented with, in educational settings. In [3], an attempt is made to generate questions based on the content of English stories. Questions in five different categories of understanding were framed by extracting syntactic and semantic information present in the stories using natural language processing. Their work assisted specifically to the language learning ability of the learners. The authors compared their generated questions to those that were asked in their collection of book problems and also evaluated them for semantic correctness. In [4], the concept of self-questioning in the context of reading comprehension is explored. In their approach, they generate instructions that help the learners to ask questions relevant to the passage. Children's stories were considered as the dataset of passages. Rather than generating questions randomly, the questions related to the characters' mental states involved in the passages are framed to enable to infer connections between key story characters. Ten different categories of modal verbs were used for constructing questions of three different types (what/why/how) making use of question templates. Evaluation of generated questions was done by testing for acceptability. About 71% of generated questions were marked acceptable, but they suffered from parsing and grammatical errors.

Question Generation Systems fall into one of the following domain categories: closed domain or open domain. In closed-domain question generation, questions are generated

✉ Nikahat Mulla
nikahatmulla@gmail.com

[1] Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India

[2] SVKM's NMIMS University, Indore Campus, Indore, India

for a specific domain like medicine [5, 6], educational text [7]. Here, the questions usually rely on some domain-specific knowledge restricted by an ontology. Open-domain question generation does not depend on any domain and allows you to generate questions irrespective of the domain requiring only universal ontologies. The data on the basis of which such systems can generate questions are readily available and in abundance. This type of question generation system does not cater to any specific domain of discourse and can be applied to any domain in general. There are two major approaches for open-domain questions, which have been researched upon. The first approach makes use of the syntactic structure of sentences and other natural language processing operations to produce a question from a specified sentence, and the second approach makes use of an end-to-end method which uses an approach similar to machine language translation using neural networks for generating questions. Significant contributions using both approaches are made over the years including constituent and dependency parsing [8], a representation using lexical functional grammar [9], semantic role labeling in a rule-based set-up [10], and neural network-based approaches including the generation of factoid questions using recurrent neural networks [11], where to focus on for generating questions for reading comprehension [12] or generating questions by recognizing the question type [13].

The main focus of this review is to present the researchers and practitioners with a comprehensive overview of the research carried out in the field of automatic question generation. The significant contributions this paper make are listed below:

1. To provide a detailed overview of automatic question generation methodologies.
2. To provide the list of datasets available for AQG.
3. To provide an overview of the challenges and applications in the field of AQG.

For this review, we carefully selected papers which were published in journals and conferences of repute. We used key phrases like automatic question generation, question generation, visual question generation and the like for searching for relevant research articles in this field. We then carefully categorized these articles on the basis of the use-case they try to model. We then used certain inclusion criteria to cater to the quality of content in our survey. We included only those papers exhibiting sufficient experimental proof with their models experimented on benchmarked datasets, papers which introduced the state-of-the-art methodology used for the purpose of question generation, the articles which compare their proposed models with existing work. We also included the articles which introduced various datasets for benchmarking this problem area. We focused more on articles which used various machine learning and deep learning-based architectures. We also included articles which catered to various application domains in this field for showing the usefulness of the task of AQG. We excluded the remaining articles which had incomplete experiments or less sufficient proof of results and those with no comparison among various methodologies. Also, we excluded the articles that were not written in English.

Several reviews have been published in the past by various research works in the field of automatic question generation. In [14], a review of the automatic question generation from text is presented in the years from 2008 to 2018. However, this field has advanced to a great extent over the years with introduction of recent deep learning architectures. There are a few specific reviews, in particular domain areas of question generation. For example, the work done for question generation in the educational domain is discussed in [15]. The authors have provided a systematic review of contemporary literature with the focus on quality of question structure, sub-domains in the educational field and how the research is largely focused for the assessment purpose. A similar review is presented in [16], where the authors explore the joint task of question generation with answer assessment. Also, a comprehensive survey based on the task of visual question generation is presented in [17]. Our review is different from the earlier ones as it provides a detailed discussion on the methodologies for question generation. We also categorize the question generation techniques broadly based on three different use-cases. We analyze the datasets and metrics used in question generation for each of the use-cases identified.

This paper is organized as follows: We first formally represent the problem of automatic question generation and discuss the various question categories, summarizing technically the idea of such a system. Next, we provide a classification of the AQG methodologies based on three distinct use-cases: standalone question generation, visual question generation, and conversational question generation along with a comparative analysis of these methodologies. We next provide a comprehensive overview of the datasets available for training such automatic question generation systems. We also list all the types of metrics, both automatic and human-based used for rating the performance of question generation models. We evaluate the datasets available for training question generation systems and categorize the datasets which are existing for different use-cases of AQG. We finally identify the various research challenges in AQG systems and briefly discuss applications of such systems as seen in various research works.

2 Automatic question generation overview

Automatic question generation systems are realized using varied approaches. Also, the kind of questions that such a system generates is important when choosing an approach. Before we delve into the approaches, we first formally define the problem of automatic question generation and give a short description of the types of questions that can be generated.

2.1 Problem definition

The problem of automatic question generation can be formally described as discussed in this section. Consider a given input modality, text or image based on which a question has to be generated. Let I represent the input, Q represent the question to be generated, and A be the answer relevant to the question. We define the automatic question generation problem as follows:

Find a function

$$f(I,\ A) = Q^{'}$$

such that Q' is semantically equivalent to Q.

The input I can be represented as a vector of relevant features, either an image or text. The question generation problem is to find a model that approximates the question generated by it, namely Q' to the labelled question Q. To realize this problem, the dataset is first pre-processed as per the requirement to make data available in the desired format. Based on the question type, an appropriate strategy is chosen for question generation. Depending on the type of question generation system, an appropriate data set can be chosen. The data could be in text form or images. The question generator model may be rule-based or neural network-based. For a rule-based AQG system, an appropriate NLP technique is used for generating the questions while for a deep-learning-based strategy, appropriate representation is chosen for training the model.

2.2 Question categories

When we consider the question categories, various taxonomies were proposed. An important contribution towards this direction is Lehnert's classification [18]. As part of development of a computational model for question-answering, Lehnert classified questions based on the idea of conceptual categorization. As per this idea, in order for the question to be interpreted correctly, it must be placed in the right conceptual category, otherwise it will lead to wrong reasoning. In this sense, the emphasis should be on the context in which the question was asked. Accordingly, Lehnert proposed thirteen such conceptual categories, namely causal antecedent, goal orientation, enablement, causal consequent, verification, disjunctive, instrumental, concept completion, expectational, judgmental, quantification, feature specification and request [18].

A similar classification scheme includes [19], where an analysis of questions during tutoring sessions was made. Thus, for any question generation system, it is important to identify the types of questions which can be generated by it. Moreover, there can be several types of questions based on whether they are meant to be asked for expecting to-the-point answers or span several lines or fill-in-the-blank type questions. Questions may also be characterized, on the basis of cognitive levels of the answer expected. Other classifications could determine whether the questions are extractive or abstractive. Extractive questions are based on words extracted from the passage itself while abstractive questions would have as answers meaningful words but different from the passage. Question categorization helps in chalking out the exact use-case that has to be realized. We list below a classification of questions on the basis of various research carried out in the field of question generation.

(1) Factual questions
This category of questions is simple objective questions that start with what, which, when, who, how. Here, the expected answer is a word or a group of words from sentences on a paragraph of text. Most of these questions are asked by choosing a single sentence from a paragraph and expect a known fact as an answer. Complex natural processing is not required for answering factual questions.

(2) Multiple sentences spanning questions
Some questions can require multiple sentences of a paragraph as the answer. The facts are present in several sentences. These questions are again W4H (What/Where/When/Where/How) questions and can be solved using the same approaches that are used for solving factual questions.

(3) Yes/no type questions
These are questions that require a Boolean response, yes/no. These require some level of reasoning. Such questions require a higher level of reasoning in order to reply with a yes or no correctly.

(4) Deep understanding questions
These are inference-oriented questions that require a proper inference mechanism. These questions might require deriving a fact from several related facts in a piece of text. These are complex questions that require different information from varied parts of the text.

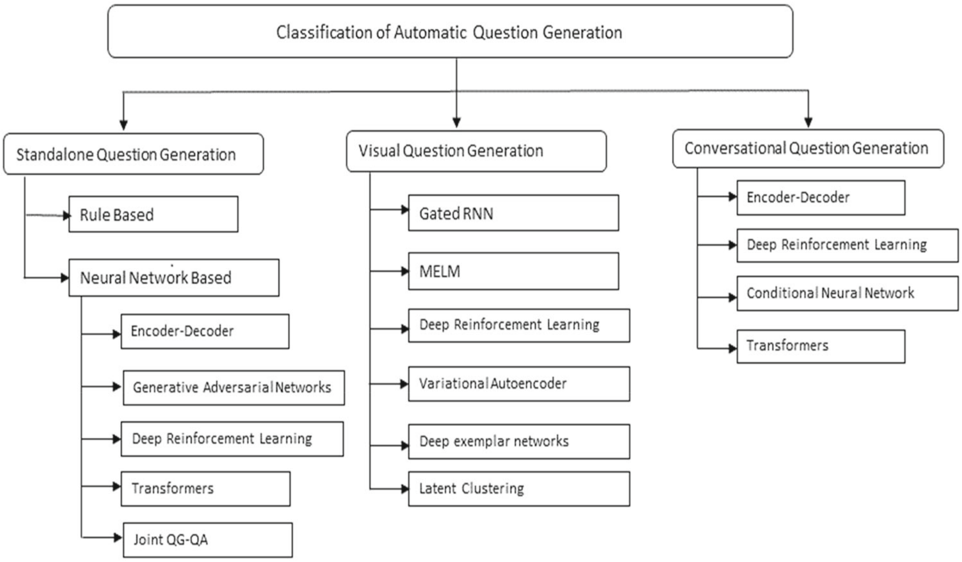

Classification of Automatic Question Generation
Standalone Question Generation
Rule Based
Neural Network Based
Encoder-Decoder
Generative Adversarial Networks
Deep Reinforcement Learning
Transformers
Joint QG-QA
Visual Question Generation
Gated RNN
MELM
Deep Reinforcement Learning
Variational Autoencoder
Deep exemplar networks
Latent Clustering
Conversational Question Generation
Encoder-Decoder
Deep Reinforcement Learning
Conditional Neural Network
Transformers

Fig. 1 Classification of automatic question generation

3 Classification of automatic question generation techniques

In this section, we come up with the different categories of automatic question generation. We make the distinction based on two different aspects. The first aspect is based on the application use-case they try to model. We further provide a categorization using the different classes of methodologies used in each use-case.

Broadly, we identify three types of question generation systems: standalone question generation (SQG), visual question generation (VQG), and conversational question generation (CQG) (Fig. 1).

3.1 Standalone question generation

In this type of question generation, the questions are generated independently of each other. This is typically the idea about machine reading comprehension systems where the only goal is to produce semantically and syntactically correct questions based on a paragraph of text or certain rules for language modeling. However, there is no correlation of the different questions generated.

3.1.1 Rule-based approaches

The authors attempt the problem of question generation using an educational learning resource called OpenLearn which covers a wide range of different discourse types authored by various subject experts in [20]. For their implementation, the authors convert the matter from OpenLearn, which is represented in XML format into plaintext, and further apply NLP processing for forming a syntax tree. The system then uses pattern matching to generate questions on sentences. The patterns are used as a part of rules, which can match sentences from the text for generating questions and the corresponding answers.

In [21], a rule-based approach is employed for producing questions from declarative sentences. The approach first simplifies the sentence and then applies a transformation technique for question generation. The generated questions are then ranked through logistic regression for quality. The ranked questions are then annotated for acceptance. This approach of ranking improved the acceptability of the generated questions by the annotators.

A mechanism of generating questions from the online text for self-learning is proposed in [22]. The authors focus on what to ask the question about from a given sentence, i.e., the problem of gap selection. For this task, they use articles from Wikipedia and perform key sentence extraction via automatic text summarization [23, 24]. Then, multiple question/answer pairs are generated from a single sentence, which is later on given to the question quality classification model. They use [25] as their text summarization model, semantic and syntactic constraints via constituency parser and semantic role labeler for generating multiple questions from a sentence, and crowdsourcing for rating question quality. The aggregated ratings along with a set of extracted features from the source sentence and the generated question are then given to a classifier that tests question quality using L2-regularized logistic regression [26]. Features used for training the classifier were in different categories like token count, lexical, syntactic, semantic, named entity and Wikipedia link features. After their experimentation, the authors were able to train the classifier, which could largely agree with the human judgments on question quality.

An approach for high-level question generation based on text is discussed in [27]. A combined ontology and crowd-relevance-based technique on the Wikipedia corpus are proposed for this task. The authors first create an ontology of categories and sections. They make use of Freebase for creating categories and for each category, they use sections. For example, if there is an article about Albert Einstein, it falls under the category 'Person' and is further segmented using sections like Early_life, Awards, Political_views, etc. The authors then present this ontologically classified data to crowd workers to generate questions based on a Category-Section part of the articles. With these generated questions, the authors train 2 different models. The first model is for finding the category and section of an unseen article segment. For this, they use logistic regression classifiers for both the categories and the sections individually. The other model is also a classification model which predicts whether a question is relevant for a section. The authors concluded their experimentation by reporting recall and precision scores on an end-to-end task of generating questions on an article-segment pair given by the user.

A technique that employs natural language understanding (NLU) for generating questions is proposed in [28]. The technique improves the acceptability ratio of generated generations. In their approach, the authors first examine the pattern of constituent arrangement for understanding what a sentence is trying to communicate to determine the type of question that should be asked for that sentence as part of the DeconStructure algorithm that they propose. The algorithm works in two phases: the deconstruction phase, in which the sentence is parsed by means of a dependency parser and a Semantic Role Labeling (SRL) parser, and the structure formation phase, in which the output from the parses is combined to recognize the clause components and a label is assigned for function representation of each clause component. After this step, the sentence patterns are classified into relevant categories before proceeding for question generation. The question generation is based on matching approximately 60 templates with the template which has the best match being used for generating the question. Subsequently, a ranking mechanism was employed for deciding whether a question is acceptable or not using the TextRank algorithm [29] for keyword extraction. This helped to identify the most important questions. The authors found that they were able to improve the acceptability of questions by 71% from the top-ranked questions in comparison with state-of-the-art systems.

A system for generating questions from Turkish biology text has been proposed in [30]. For this, a corpus was created which was semantically annotated using SRL. Biology high school textbooks were chosen as the text for the corpus. SRL proceeds with POS tagging for predicate identification, argument identification by following a set of rules, and argument classification utilizing self-training. After the SRL, the system proceeds with automatic question generation using set templates and rules. In this approach, first templates are tried, and if no template is found, then an appropriate rule is used to formulate the question. Turkish sentence structure is used to formulate a question.

Comments on rule-based models Table 1 provides a comparative overview of the models used in rule-based techniques for standalone question generation. As seen from the table, most approaches make use of Wikipedia as their dataset and the evaluation metrics for automated evaluation include f1-score or precision. Evaluation is not strongly made in terms of a well-defined metric though human based ratings have been explored. These algorithms rely often on extracted features and later add a classifier model. They extract syntactic and semantic parts of text and make use of templates to construct the question. Usually, smaller target topics are considered where specific types of questions are required to be generated. General texts will not be converted effectively to questions if rule-based algorithms are used for question generation.

3.1.2 Neural network-based approaches

With the humongous number of datasets available in current years, neural-based approaches have become very popular for automatic question generation. In this section, we discuss the different approaches which have been employed to solve this problem.

Encoder–decoder (sequence-to-sequence) architectures The encoder–decoder architecture was introduced by Google in [31]. This architecture promotes end-to-end learning for tasks that require a sequence of tokens as input and a sequence of tokens as output. This makes it very convenient to use in language processing tasks. Encoder–decoder models are typically used for modeling problems that are based on sequences as inputs and sequences as outputs; hence, they are often called sequence-to-sequence models. The architecture is further divided into two subparts: an encoder, which is used to encode the input sequence by passing it through a series of recurrent neural network layers, and a decoder, also a series of recurrent layers, which attempts to produce the output sequence. Shown in Fig. 2 is a typical encoder–decoder architecture that can be used for SQG. When used for SQG, the encoder–decoder model takes input passage (and answer) as input and attempts to produce a question similar to the labelled question.

The authors have attempted to generate questions based on a paragraph of text for machine-reading comprehension in [32]. The authors have used an attention-based mechanism built upon a sequence-to-sequence model for the same. They have used an RNN-based encoder–decoder mechanism in which they have created two different encoders. The first encoder network is for encoding the sentence-level information, while the second network encodes the combined sentence and paragraph-level information. Both encoders are attention-based bidirectional LSTM networks. The authors

Table 1 Summary of rule-based approaches used for SQG

References	Model	Question type	DataSet	Automatic evaluation	Human evaluation
[20]	Pattern templates using NLP-based Ceist tool	Factoid	OpenLearn	–	–
[21]	Over generate and rank framework	Factoid	Wikipedia, Penn Treebank	–	Rating for acceptance level
[22]	Text-summarization, semantic and syntactic features	Gap-fill questions	Wikipedia	–	Crowdsourced rating for quality
[27]	Ontology-crowd-relevance workflow	Deep questions	Wikipedia	Information Retrieval based evaluation in terms of relevance	–
[28]	Deconstruct algorithm to generate sentence patterns	Factoid	Open source text + Wikipedia	–	Crowdsourcing using Amazon's Mechanical turk service
[30]	Syntactic and semantic approach using SRL	Questions on descriptive sentences	Turkish biology text	Precision, recall, F1 and accuracy for evaluating SRL	–

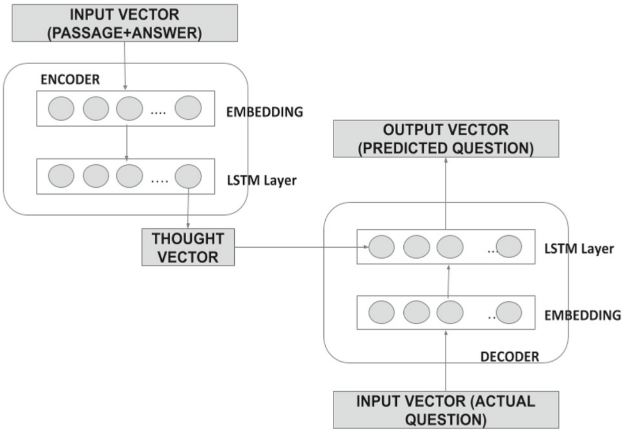

INPUT VECTOR (PASSAGE+ANSWER)
ENCODER
EMBEDDING
LSTM Layer
THOUGHT VECTOR
OUTPUT VECTOR (PREDICTED QUESTION)
LSTM Layer
EMBEDDING
DECODER
INPUT VECTOR (ACTUAL QUESTION)

Fig. 2 Encoder–decoder architecture for automatic question generation

use the SQuAD dataset for training their model with randomly generated training, development, and test sets. They use pre-trained glove embeddings with 300 dimensions for word representation. They used 2 LSTM layers in both the encoder and decoder networks and strained them using Stochastic Gradient Descent. For experimental analysis, the authors considered five different baseline models, namely IR(Information Retrieval) [33], MOSES+ [34], H&S [35], and Seq2Seq vanilla model [31], and performed automatic as well as human evaluation. For automatic evaluation, they used the BLEU, METEOR, and ROUGE metrics and for human evaluation, naturalness and difficulty of question generated were considered as parameters. For human evaluation, a set of 100 randomly sampled question–answer pairs were chosen and evaluated by 4 professional English speakers on a rating from 1 to 5(5-best). The authors observed that both the sentence-based and paragraph-based models that they proposed performed better than the baselines in both the automatic and human-based evaluation. However, the paragraph-level model was not the best for all metrics, so the paragraph-level information can be used more efficiently for implementation to improve performance which provides directions for future work.

The authors propose a novel neural network-based question generation technique that generates a question towards a target aspect from an input piece of text [35]. The idea is inspired by the fact that in a typical conversation, seldom, questions are asked randomly and are always asked

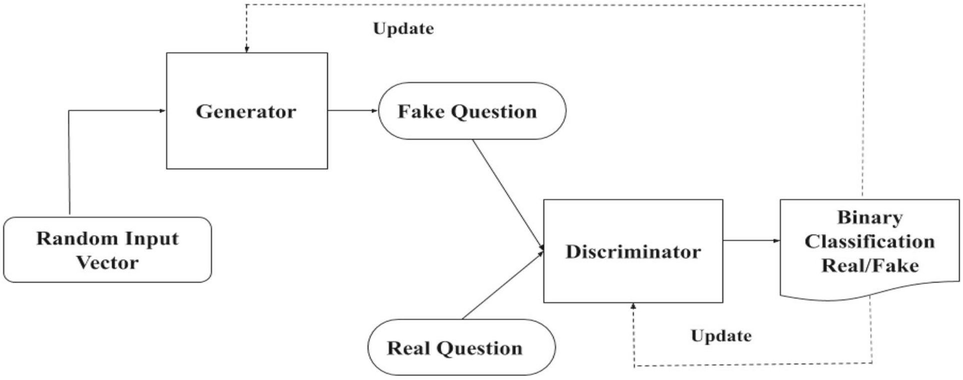


Fig. 3 GAN-based general architecture for automatic question generation

on some relevant aspects. For this purpose, a sequence-to-sequence neural network-based framework is used, which employs a pre-decode mechanism for improving the framework performance. They also employ two techniques to the sequence-to-sequence framework, an Aspect, and a Question Type so that they can improve the question quality based on the question type. Another mechanism that they use is an encoder–decoder framework with separate encoders for aspect, question type, and answer. For generating aspects from a given sentence, the authors make use of the cosine similarity metric to identify the semantically similar words from the sentence based on words from the question. Using a voting mechanism, the candidate words are selected from the sentence as part of the aspect if the average vote is greater than some threshold. After this extraction of aspects from the sentences, the authors perform noise removal by using a pre-decode mechanism and perform stop words removal. For question types, the authors make a distinction among 7 categories of question types, namely yes/no, W4H, and others. They make use of these keywords to identify the question types. For aspect and question type, authors make use of a bidirectional LSTM and for the answer, they use another bidirectional LSTM. For the decoder, the authors use another LSTM network. As part of the pre-decode mechanism to clean the generated aspect, the authors use yet another LSTM which acts as a filter for noise removal. For their experimentation, the authors make use of the Amazon Question/Answer corpus (AQAD) which contains 1.4 million question–answer pairs about products and services from Amazon. For results analysis, they divide the corpus into training, development, and test sets. The authors performed both automatic evaluations using BLEU, METEOR, and ROUGE as well as human-based evaluation and found that their model outperformed the baseline model in [32] that they considered.

In [36], question–answer pairs in natural language are extracted through a knowledge graph making use of the RNN-based model for question generation. In this approach, a set of keywords are first extracted from a knowledge graph. A subset of these keywords is then used for generating questions through a sequence-to-sequence-based RNN model. An encoder–decoder architecture is used in which a bi-directional RNN is used as the encoder with a hidden layer of 1000 neurons and the decoder is also constructed similarly. 1mn questions are extracted from WikiAnswers which forms the dataset for model training. In their approach, the authors create 2 models as part of the framework for generating QA pairs. In the first module, knowledge about the entities is extracted from the knowledge graph, and it is independent of language. In the second module, questions in natural language are generated using question keywords using an RNN. The RNN model gave higher BLEU4 scores over other compared baselines, phrase-based machine translation [34] and template-based method [37].

Generative adversarial network-based approaches The generative adversarial networks (GANs) were introduced in [38]. This class of deep neural networks makes use of two different networks, the generator and discriminator, which compete against each other in an adversarial set-up. The role of the generator is to generate samples of the required problem domain close enough to the labelled data such that the discriminator is not able to identify whether the sample generated is fake or real. A typical architecture for a question generating GAN is shown in Fig. 3. The generator tries to generate fake questions similar to the labelled question for the given answers. The role of the discriminator is to be able to correctly identify that the generated question is fake or not. In the process the generator is able to fool the discriminator by generating fake questions which are close enough to the original labelled questions and that is when training stops.

The problem of fill-in-the-blank (FITB)-type question generation is dealt with in [39]. The creation of important distractors is done using generative adversarial networks for training. A FITB question consists of the sentence, the key (correct answer), and the distractor answers. The authors attempt to generate distractors, given a sentence and the key. In this approach, the generator of the GAN is used to capture real data (key) distribution from a given question sentence and the discriminator tries to estimate whether the key came from the actual data(real) or the generator(fake). The model training is performed on the subject of biology from the Wikipedia corpus. The proposed method performed better than already existing similarity-based methods.

The introduction of variability in the questions generated and prediction of question type is incorporated in a GAN framework discussed in [40]. The GAN model accounts for variability using a latent variable and its discriminator evaluates the genuineness of the question and predicts the type of the question. The generator of the GAN is an encoder–decoder network based on conditional variational autoencoders [41]. The discriminator is modified to act as a classifier for the question type (WHO, WHAT, WHICH, HOW, WHEN and OTHER) along with the task of classifying the real questions from fake questions. Experiments were conducted on the SQuAD dataset [42] and several variants of the models were created and compared against baselines from [43] and [44]. Automatic evaluation on BLEU, METEOR, and ROUGE scores was performed along with human-based evaluation and their proposed model outperforms the baselines considered in both types of evaluation.

In [45], the authors have addressed the problem of generating questions on a specific domain with the absence of labeled data. For this, they have proposed a new model which uses doubly adversarial networks. These networks use the data with ground-truth labels from one domain and unlabeled data from the goal domain for training. They used the SQuAD[46] dataset for unlabeled data and NewsQA[47] dataset as labeled data. Their experimentation proved that their model gave better results than existing methods.

In [48], an attempt is made to generate clarification questions on text to extract useful information that captures the context of the text. In this GAN-based approach, the generator is a sequence-to-sequence model which first generates the question on the basis of a context and then generates a hypothetical answer to that question. The question, answer, context triplet is then given to the discriminator, which uses a utility-based function to compute the usefulness of the question. The evaluation was made on two datasets: the first one was a combined dataset which consisted of the Amazon question answering dataset [49] and the Amazon reviews dataset [50], and the second was the Stack Exchange dataset [51] curated from stackexchange.com. The baseline model was an information retrieval-based model Lucene based on [49], which was compared against several variants of the proposed model, namely GAN-Utility, MaxUtility, and MLE. The models were evaluated using both automatic evaluation metrics BLEU, METEOR, and Diversity [52] and human evaluation based on the criteria of relevance, grammar, seeking new information, usefulness, and specificity. It was observed that the adversarial training given to the model produced good results than both MLE and the reference models.

Deep reinforcement learning architectures In [53], authors have attempted to create a framework for generating intelligent questions in the context of conversational systems. As most of the work in automatic question generation is utilizing neural-based systems, authors have extended this approach and have created a model based on deep reinforcement learning for question generation. The authors have used an end-to-end model which uses a generator and an evaluator. The generator model is based on the question's semantics and structure. It uses a pointer network mechanism to identify the target answers and a copy mechanism to retain the contextually important keywords. It also uses a coverage mechanism for removing redundancy in the sentences. The evaluator mechanism employed in this paper uses direct optimization based on the structure of sentences using BLEU, GLEU scores, etc. It also matches the generated questions against an appropriate set of ground-truth sentences. The authors also introduce two new reward functions for evaluating the quality of the generated questions, namely Question Sentence Overlap Score (QSS) and Predicted and Encoded Answer Overlap Score (ANSS). The authors conducted their experimentation on the publicly available SQuAD dataset and compared their model's results with two state-of-the-art QG models as baselines, namely L2A [32] and AutoQG [54]. They compared the baselines with eight variants of their model and used standard automatic evaluation techniques like BLEU, ROUGE-L, and METEOR as well as human evaluation techniques to further analyze the quality of their questions for syntactic correctness, semantic correctness, and relevance. For human evaluation, the authors randomly selected a subset of 100 sentences and presented the 100 sentence-question pairs to 3 different judges for getting a binary response on each quality parameter (syntax, semantics, and relevance), and the responses from all judges for every parameter were averaged for each model. After comparison, the authors observed that their model variant which included ROUGE, QSS, and ANSS outperformed the two state-of-the-art baselines on automatic evaluation using BLEU, METEOR, and ROUGE. In human evaluation, their model variant which used DAS, QSS, and ANSS was the best model on syntactic and semantic correctness and the one which used BLEU, QSS and ANSS gave the best relevance among all models. So, the authors conclude that the QG-specific reward metrics that they proposed, namely QSS and ANSS, improved the model significantly and outperformed the state-of-the-art methods.

In [55], the authors propose a refinement of ill-formed questions generated to well-formed questions by using a reward-based mechanism using reinforcement learning for training a deep learning model. The rewards used utilize the wording of questions as a short time reward and the correlation of question and the answer as the long-term reward. They also make use of character embedding and BERT-based embedding for enriching the representation for question generation. The authors conclude that their approach could produce comparatively readable questions.

A graph-to-sequence-based architecture guided by deep reinforcement learning is proposed in [56]. In their proposed architecture, the authors make use of a gated bidirectional neural network architecture and use a hybrid cross-entropy and reinforcement learning-based loss function to train the network. They also add answer information in the model training process. They use several state-of-the-art models to generate questions and compare them with 2 different variants of their model, the first one using syntactic information from the passage to construct a static graph and the other using a semantics-based dynamic graph. They evaluated the models on the SQuAD dataset and found that it outperformed the earlier state-of-the-art models on automatic evaluation metrics as well as human evaluation methods by a substantial margin.

Joint question answering-question generation approaches Some approaches make use of a joint question answering and question generation approach for automatic question generation.

For instance, in [57], the approach used is one where the question is asked as well as answered. The proposed model was trained on the joint task of question-answering and there was a substantial improvement in the model performance for the SQuAD dataset. An attention-mechanism-based sequence-to-sequence model was used for this task, which used a binary signal to set the learning mode as answer generation and question generation, respectively. The model was then compared to the QA-only model and it was observed that the joint model had better results than the QA-only model.

In another approach for joint QA-QG, the correlation between the task of QA and QG is exploited to improve model performance. A sequence-to-sequence model is used for QG and a recurrent neural network is used for QA and the results of the model are compared on 3 different datasets: MARCO, SQuAD, and WikiQA. The QA model is implemented using a bi-directional RNN which uses word embeddings for representing the inputs—the question+list of candidate answers and predicts the best answer from the candidates. The input to the QG model is an answer and its goal is to generate a relevant question. The QG model uses an encoder–decoder approach in which the answer representation is first done using an encoder, and later, the decoder generates a question based on the answer representation. The architecture jointly trained both models to improve the overall performance of both QA and QG on the datasets used [58].

Transformer-based approaches Several approaches based on transformers [59] have been experimented with recently.

In [60], the task of question generation from passages was attempted on the SQuAD dataset using transformers. Word error rate (WER) was used as the metric for comparing the generated questions with the target questions. The authors observed that the generated questions were correct syntactically and were of relevance to the passage. WER was low for the shorter questions, while it increased for longer questions.

In [61], the transformer model is improved further to generate questions on the SQuAD dataset. The ELMO (Embeddings from Language Models) representation [62] is employed to denote the tokens. Placeholding strategy for named entities is used as also a copying mechanism is employed in the different variants of the models experimented with. The models were evaluated on automatic metrics (BLEU, ROUGE) as well as human evaluation for correctness, fluency, soundness, answerability, and relevance and found that the model employing the ELMO+placeholding+copy mechanism gave better results on SQuAD.

A single pre-trained transformer-based model is used for generating questions from text in [63]. In particular, they use the smallest variant of the GPT-2 model [64], a pre-trained model which was fine-tuned further for their models. The model is purely dependent on the context, so answer labeling is not required. The model was evaluated on automatic metrics of BLEU, METEOR, and ROUGE and found to give average results. However, the simplicity of the model accounts for this observation. With more processing resources and bigger GPT-2 models and also other parameters of consideration, the model may give better evaluation scores.

In [65], the authors use a combination of the transformer-based decoder of the GPT-2 [66] model with the transformer encoder of BERT [67]. The authors train their model on the SQuAD dataset and use a joint question answering-generation-based approach for training. Each network (encoder and decoder) is trained individually for answering and generating questions, respectively. The authors evaluate their model on quantitative and qualitative metrics. They also propose a metric BLEU QA as a surrogate metric for assessing question quality. Their model produces good quality questions with maximum semantic similarity to ground-truth answers using the semi-supervised approach proposed.

A recurrent BERT-based model is explored in [68]. In their approach, the authors use a BERT model as an encoder and another BERT model as the decoder to generate questions using the SQuAD dataset. In comparison with other models using standard evaluation metrics, their model gave better results on both sentence-level and paragraph-level question generation.

In [69], an attempt is made to work on multiple question types using a single architecture based on pre-trained transformers, namely T5 (text-to-text transfer transformer) and BART (bidirectional and auto-regressive transformers). Among the question types, the authors chose extractive, abstractive, MCQ, yes-no and also abstractive questions combining various datasets comprising of such question types. The authors fine-tune the T5 and BART models on their combined dataset containing passages of text from 9 existing datasets. Evaluation of their unified model was done on both automatic metrics and qualitative parameters, which gave state-of-the art results.

Comments on neural-based models Table 2 gives a comparative overview of the models used in neural network-based question generation. As seen from the table, the features of the models are listed and the automatic evaluation scores using the standard metrics of BLEU, METEOR and ROUGE and others are compared. The various models that performed better have their automatic scores highlighted in bold. When we add more features, for example, RNN+ knowledge Graphs better than a purely RNN-based approach. In a few cases, reinforcement learning, which employs a reward-based training mechanism, also shows promising results. Also, the advanced models, namely transformers like GPT-2 and BERT, gave best results on the most frequently used dataset (SQuAD).

3.2 Visual question generation

Such systems are useful as an alternative to the solution of image captioning. In image captioning, the goal is to generate an account of the objects seen in an image. On the other hand, visual question generation (VQG) tries to accomplish the same goal by generating questions based on the objects in the image.

3.2.1 Methodologies used for VQG

Most of the techniques used for VQG include the use of neural network architectures employed in different perspectives. The task of visual question generation is introduced in [72]. The purpose of VQG is to generate questions that are natural and engaging for the user to answer. Three different datasets which range from object to event-centric images are also created for this purpose by the authors. The authors form 2 datasets, one with 5000 images from the MS COCO dataset [73] and the other with 5000 images from Flickr [74]. A third dataset was curated from the Bing search engine, which was queried with 1200 event-centric terms. The three datasets together comprise a wide range of visual concepts and events. The authors further present different retrieval and generative model architectures to accomplish the task of VQG. Among the generative models, maximum entropy language model (MELM) [31, 75, 76], machine translation (MT) sequence-to-sequence model, and gated recurrent neural networks (GRNN), derived from [77, 78] were constructed and evaluated. Among the retrieval-based models, different variants of the K-nearest neighbor model (KNN) were created and evaluated. The evaluation was performed using BLEU and METEOR for automatic evaluation and performed human evaluation by crowdsourcing three crowd workers for rating the semantic quality of generated questions on a scale of 1 to 3. This was the first paper that discussed the task of VQG and released 3 public datasets for the research community to solve this problem using different models. Also, various architectures were discussed, which could be used in training such models.

In [79], the problem of generating goal-centric questions on images is addressed using a deep reinforcement learning approach. A game GuessWhat?! with a goal-oriented flavor is used for applying the proposed model. In their approach, the authors prompt the agent to ask multiple questions with informative answers till the goal is achieved. For this, three different reward functions are proposed which compute intermediate rewards. The first reward is the goal-achieved reward for achieving the final goal, the second is the progressive reward, which ensures that every new question asked by the agent leads it more towards the goal and the third reward is called ‘informativeness’ which checks whether the agent is not asking useless questions. For evaluating the performance of their model, the authors use the GuessWhat?! dataset and create several variants of their model using different combinations of the reward functions and compare with [80]and Soler [81]. They find that their model variant with all three rewards surpasses all the compared models. They also conducted a human evaluation to compare all their models with other variants and their model outperformed in the case of human evaluation as well.

The generation of good informative questions is tackled by a model in which the maximization of mutual information between the question generated by the model, the image sample and the label answer is performed [82]. In their model, the authors used a latent space formed by embedding the target answer and the image example, and a variational autoencoder [83] is used to reconstruct them. A second latent space is set up which uses only the image and the answer category for encoding. Thus, the need for having an answer is eliminated. They make use of VQA [84] as their dataset and compare their model against several baselines. It was observed that their model outperformed all the other approaches considered for comparison.

In [85], the authors make use of an exemplar module in the existing deep learning framework for the task of VQA and VQG. For exemplars, two variants, namely attention-based and fused exemplars, are used for classification. The VQA

Table 2 Summary of neural network-based approaches for SQG

Methodology	References	Machine learning technique	DataSet	Automatic evaluation			Human evaluation
				BLEU 4	METEOR	ROUGE/OTHER	
Encoder–Decoder	[32]	RNN encoder–decoder framework with Attention, LSTM Sentence level and paragraph-level information	SQuAD	12.28	16.62	39.75	Naturalness: (1 to 5): 3.36 Difficulty: 3.03
	[35]	Encoder–decoder framework, bidirectional LSTM, Aspect and Question Type	AQAD	13.09	17.50	42.48	naturalness: 68
	[36]	Knowledge graphs, RNN	WikiAnswers	**50.14**	–	–	Ratings (1-worst:4-best) 60.13% accuracy
Generative Adversarial Networks	[39]	Distractor generation with conditional GAN	Wikipedia Corpus	–	–	–	Rating by experts: Good-40%, Fair-11.7%, Bad-48.3%
	[40]	Sequence-to-sequence model with conditional variational autoencoders as generator of GAN	SQuAD	**13.36**	17.7	40.42	–
	[45]	Doubly adversarial network	SQuAD, NewsQA	5.58	–	–	–
	[48]	Utility-based GAN, seq-2-seq model for generator	AQAD, Amazon Reviews (A) Stack Exchange (SE)	A: 15.20 SE: 4.26	A: 12.82 SE: 8.99	A: 0.1296 SE: 0.2256 (diversity)	Relevance:.0.94 Grammar: 0.96 New info: 0.87 Usefulness: 0.96 Specificity:3.52
Deep Reinforcement Learning	[70]	Generator Evaluator Framework with reward metrics using Deep Reinforcement Learning	SQuAD	16.48	20.21	44.11	Syntax: 84 Semantics:81.3 Relevance: 78.33

	[55]	QRefine-PPO, Reinforcement Learning for S2S model	Yahoo(Y)
			CSU(C)
				40.19(Y)
				40.33(C)
					35.37(Y)
					71.54(C)
						67.41(Y)
						82.74(C)
							–

Table 2 (continued)

Methodology	References	Machine learning technique	DataSet	Automatic evaluation			Human evaluation
				BLEU 4	METEOR	ROUGE/OTHER	
	[56]	Graph-to-Sequence BERT embedding, RL loss, Deep Alignment Network	SQuAD	17.94	21.76	**46.02**	Syntax: 4.41 Semantcis:4.31 Relevance: 3.79 (Scale: 1–5)

Joint QA-QG	[57]	Bi-directional RNN for QA, Encoder–Decoder for QG, and joint training	SQuAD (S), MARCO (M), WikiQA (W)	9.31(M)
				5.03 (S)
				3.15(W)
					–	–	–

	[58]	Seq-Seq + Attention, pointer softmax decoder, answer words sequence	SQuAD	10.2	–	–	–
Transformer	[60]	Transformers	SQuAD	WER: low for short, high for long context			–
	[61]	Transformer + Placeholding + Copying + ELMO	SQuAD	13.23	–	40.22	Correctness:4.5 Fluency: 4.12 Soundness: 3.78 Answerability: 2.87 Relevance:3.59
	[71]	GPT-2(small) + attention	SQuAD	8.27	21.2	44.38	–
	[65]	Transformer-based decoder of GPT-2 + Transformer encoder of BERT	SQuAD	7.84	–	34.51	–
	[68]	BERT Highlight Sequential Question Generation (Sentence level (S), paragraph level (P) context)	SQuAD	21.20 (S) 22.17 (P)	24.02 (S) 24.80 (P)	48.68 (S) 49.68 (P)	–
	[69]	T5 + BART	SQuAD, NarrativeQA	**25.42** **33.91**	**45.75** **58.34**	**51.86** **60.15**	Fluency, relevance, reasonable acceptance rate: 68.4%

Bold values depict best models

and VQA2 datasets were used for testing their models for the task of VQA, while the VQA and VQG-COCO [73] datasets were used for the question generation task. They observed through various variants of their models that they gave an enhanced performance on state-of-the-art methods on VQA and VQG on standard automatic metrics.

Visual question generation in the presence of visual question answering as a dual-task is experimented with, in [86]. The framework makes use of inverted MUTAN (Multimodal Tucker Fusion for Visual Question Answering) and attention in its design. The experiments are performed on CLEVR and VQA2 datasets and give better performance than the existing methods using the dual training framework.

In [87], an attempt is made to guide the VQG system in generating questions based on objects and categories. They employ three distinct model architectures, explicit, implicit, and variational implicit. The VQA dataset was used in the experiments. In the explicit model, the image is first labeled with objects using an object detection model and an image captioning model to generate captions. This is then given to an actor which then chooses random samples from among the candidates according to the category, which in combination are given to the text encoder. Image encoder is used to encode image features. The image and text encoder outputs are combined into the decoder to generate questions. In implicit guiding, they use only image as input and try to predict the category and objects using a classifier network. Further, a variational encoder-based implicit guiding is also experimented with where a generative encoder and variational encoder together produce a discrete vector which is then fed into the decoder. The experiments result in an improvement over several metrics of VQG.

Other visual question generation approaches include using a human in the loop where VQG is used for asking questions to users and collecting their responses to build a dataset for visual question answering [88], use of reinforcement learning along with bi-discriminators using generative adversarial networks [89] and category-wise question generation using latent clustering [90].

Comments on models used for VQG Table 3 gives a comparative overview of the models used in visual question generation. As seen from the table, the features of the models are listed and the automatic evaluation scores using the standard metrics of BLEU 4, METEOR, CIDEr and others are compared. The various models that performed better with respect to various metrics have their automatic scores highlighted in bold. For extracting image-related features, ResNet variants have been explored in most techniques. Also, attention-based models are used giving good results. We observe that reinforcement learning using bi-discriminators provide the best scores on the VQA dataset.

3.3 Conversational question generation

The primary objective of a conversational question generation model lies in generating questions that are rich in the context of the conversation. The main idea is to generate a series of questions for maintaining the conversation. In these systems, care must be taken that the conversation does not get stuck in a loop or gets too boring. The primary use-case of such a system is conversational chatbots.

3.3.1 Methodologies used for CQG

Significant research has been carried out for implementing conversational question generation systems. In [91], a neural-network-based approach is proposed which makes use of coreference alignment along with maintaining a conversational flow. This ensures that the questions generated in consequent turns are related to each other based on the conversation history. A multi-source encoder along with a decoder based on attention and copy mechanism is employed for this task. The experiments were performed on the CoQA dataset [92] and compared over various baselines models, and several ablations of existing models [93, 94] have been used in their proposed model.

An encoder–decoder-based architecture employing a dynamic reasoning technique using reinforcement learning is explored in [95]. The authors attempt to generate the next question based on the previous few questions on a given passage from the CoQA dataset. They also test their trained model on the SQuAD dataset for multi-turn question–answer-based conversations. The experimental analysis proves that the model gives better results than several compared baselines using automatic and human-based evaluation metrics.

Answer-unaware conversational question generation is explored in [96]. The proposed framework of the authors comprises three parts. The first part is the question focus estimation which decides which context to focus on to generate the next question. The second part is the identification of the question pattern which is done using either question generation or classification. These two parts are given as input to the encoder of the proposed model and the third part which is question decoding is the role of the decoder. The experiments were performed on the CoQA dataset and evaluated using BLEU scores although the authors suggest the development of new metrics for conversational question generation due to weakness of the existing automatic metrics for evaluation.

An approach using question classification based on Lehnert's classification [97] is used to tag questions and later on used in a conditional neural network-based model in [98]. The authors introduce a new task called SQUASH (Specificity-controlled Question Answer Hierarchies) which

Table 3 Summary of approaches for VQG

References	Machine learning technique	Model parameters	DataSet	Automatic evaluation			Human evaluation
				BLEU 4	METEOR	CIDEr/OTHER	

[72]	Gated RNN(G) (Generative),
	KNN(K) (Retrieval)
		VGGNet for deep convolutional image features	Bing (B)
			MSCOCO(C)
			Flickr (F)
				12.3(B-G)
				19.2(C-K)
				11.7(F-K)
					16.2(B-G)
					19.7(C-K)
					14.9(F-G)
						11.6(B-G)
						16.29(C-K)
						9.8 (F-K)
						(ΔBLEU)
							1.76(B-G)
							1.96(C-K)
							1.57 (F-G)
							(3 point semantic scale-3 raters average)

[79]	Deep reinforcement learning	Reward functions: goal achieved, progressive informativeness on objects(O)/images(I)	GuessWhat
			*Metric = correct target object located accuracy
			*Inference = sampling,greedy,beam-search
				63.2(O)
				59.8(I)
				(sampling)
					63.6(O)
					60.7(I)
					(greedy)
						63.9(O)
						60.8(I)
						(beam-search)
							76
							(human in the loop on proposed model)

[82]	Variational autoencoder	mutual information of the image, the generated question,the expected answer, ResNet18 as image encoder	VQA	14.49	18.35	85.99	98 Relevance of question with an image (3 raters)

[85]	Deep exemplar network	Attention,fused exemplars Multimodal differential network	VQA (A), VQG COCO (G)	BLEU1:
				65.1(A)
				36(G)
					22.7(A)
					23.4(G)
						A:
						52.0 (ROUGE)
						42.6
						G:
						41.8 (ROUGE)
						50.7
							–

[86]	Invertible Question Answering Network	MUTAN, Attention, ResNet152 for visual features	VQA2 (V), CLEVR (C) *Top-1 accuracy-VQA2 CIDEr-CLEVR	Top-1 accuracy:55.1	–	76.30	–
[87]	Transformer-based text and image encoder and text decoder VQG	Category features, image features	VQA v2.0	24.4	25.2	214	Grammar-93.5% Relevance(Generated Ques to Image)-77.6% Relevance (Objects to Generated Ques)-74.1%
[88]	VQA + VQG with humans and attention	Resnet for image features, LSTM for QG	Visual genome	Response rate: 26–31%	–	–	–
[89]	Reinforcement learning	Bi-discriminators: natural and human-written	MSCOCO of VQG, VQA	**26.265**	**25.634**	**57.679(ROUGE)** **63.388**	1.86(3 evaluators avg.)
[90]	Variational autoencoders	Category consistent cyclic VQG	VQA	10.04	13.60	42.34 (ROUGE) 46.87	Relevance Image:97.80% Category:60.50% (Crowdsourced)

Bold values depict best models

converts the text into a hierarchy consisting of question–answer sets which start at broader “high-level” questions and keep proceeding with more refined questions down the hierarchy. The authors test their proposed pipeline on the 3 datasets (SQuAD, QuAC, CoQA) and get promising results.

The problem of generating informative questions is dwelled upon in [99]. For generating information seeking questions in the context of a conversation, the authors use reinforcement learning to optimize the information seeking content. The architecture used in this work consists of two fragments: the automatic question generation model and the informativeness and specificity measurement model for the generated question. The experimental analysis is modelled in the form of a teacher–student communication game. A sequence-to-sequence model is used with the encoder containing the representation of the topic of interest shared between the student and teacher and the decoder is used to generate the conversational question. The informativeness is measured using what additional information a given answer by the student provides which was not present the history of the conversation so far. Apart from the informativeness metric, the authors also propose a specificity reward which is obtained by training a classifier to distinguish positive with negative samples (in terms of how a question would divert from the current topic). The experiments are performed on the conversational QuAC dataset and the combined metrics for informativeness and specificity help to direct the conversation towards rationally relevant questions.

An architecture which employs flow-propagation-based learning for generating conversational questions is discussed in [2]. In this work, a question is generated based on a given passage, a target answer and the history of dialogue which has occurred before the current turn in a multi-turn dialog set-up. For encoding the answer and for question generation, the GPT-2 model [66] is used. The authors introduce a flow-propagation based training mechanism which considers the losses accumulated after *n* turns in a dialog sequence, thus improving the flow of conversation. The model outperforms several baselines considered, including T5 model and BART-large.

Comments on models for CQG Table 4 gives a comparative overview of the models used in conversational question generation. As seen from the table, the features of the models are listed and the automatic evaluation scores using the standard metrics of BLEU 4, METEOR, ROUGE and others are compared. In general, it is observed that encoder–decoder give decent results if using other model parameters like coreference alignment, multiple encoders for representing the text involved or usage of classifiers to improve QG. On the other hand, reinforcement learning boosts the performance when goal based QG is targeted. However, the use of transformer-based architecture like GPT-2 gives promising results. This is because transformers perform much better at remembering sequences than encoder–decoder architectures based on recurrent nets. The various models which performed better have their automatic scores highlighted in bold. We observe that GPT-2-based models gave the best results on CoQA, the most used dataset.

3.4 Summary of approaches for AQG

We summarize the use-case-based question generation classification in terms of the preprocessing techniques and the methodologies in Fig. 4.

In Table 5, we list the different models used for each use-case of question generation. We also list the strengths and weaknesses of each model and identify the research gaps in them.

4 Evaluation techniques for question quality

The questions generated by a model must be evaluated for question quality so that the questions generated make sense for the purpose for which they were generated. For this purpose, there are broadly two types of evaluation techniques: automatic evaluation and human-based evaluation. We describe them briefly in this section.

4.1 Automatic evaluation

Several metrics are available for evaluating the output produced by language production systems. These metrics can be used to check the closeness of the machine-produced questions which the actual questions. For automatic evaluation, there are two scores, namely precision and recall. Precision is a measure of specificity, while recall is a measure of sensitivity. The popular metrics for automatic evaluation make use of precision and recall. A few of the popular metrics for evaluation are discussed in the sections below.

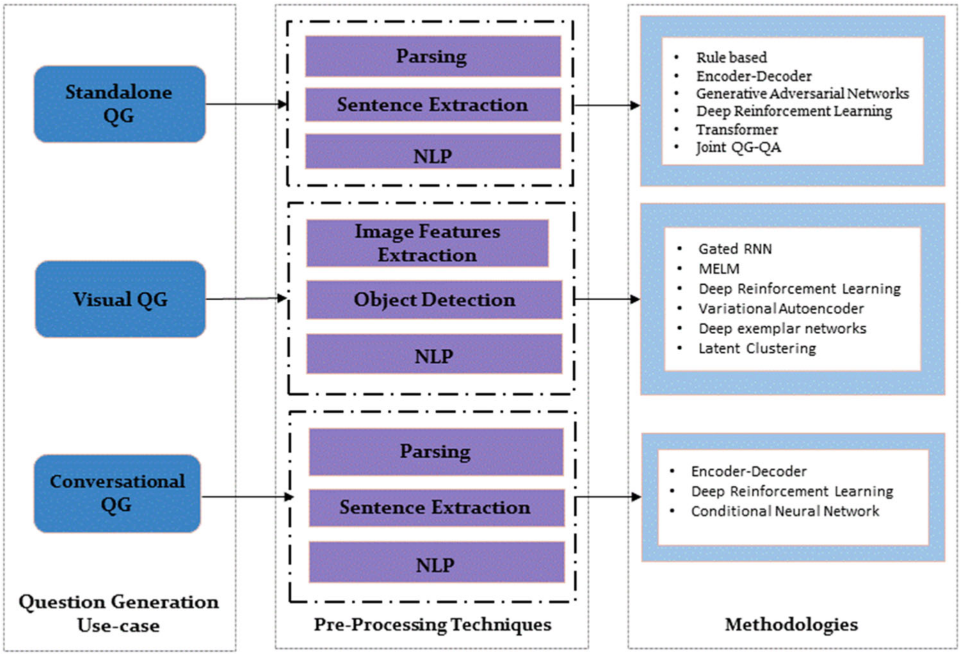
- *BLEU (BiLingual Evaluation Understudy Score)*

In this metric, modified n-gram precision and the best match are used for computing precision and recall. n-gram precision is the fraction of n-grams in the given text found in one or more of the ground truth (reference) texts available. BLEU modifies this quantity as finding words that are present only those many times as they are existing in any of the reference texts. The best match length is used to compute the sensitivity of a candidate to general reference texts. For this, sentences with shorter lengths are penalized by a multiplicative factor [100]

Table 4 Summary of approaches for CQG

References	Machine learning technique	Model parameters	DataSet	Automatic evaluation			Human evaluation
				BLEU 4	METEOR	ROUGE/OTHER	(Scale: 1 to 3)
[91]	Encoder–decoder	Coreference alignment Multi-source encoder: Passage + Conversation,Decoder:attention + Copy	CoQA	BLEU3: 18.44	–	47.45	Grammar: 2.89 Answerability: 2.74 Interconnectedness: 2.67
[95]	Encoder–decoder	Answer unaware encoder–decoder, question pattern identification, classification and generation	CoQA	7.10	–	–	–
[96]	Deep reinforcement learning	Reinforced Dynamic Reasoning	CoQA	**19.69**	–	34.05	Naturalness: 1.92 Relevance: 2.02 Coherence: 1.94 Richness:2.30 Answerability:1.86
[98]	Conditional NN	SQUASH (Specificity-controlled Question Answer Hierarchies), QG based on conditional encoder–decoder with paragraph, answer span, conceptual class as inputs	SQuAD QuAC CoQA	**–**	–	–	Well-formedness: 85.8% Relevance:78.7% Span containment:74.1%
[99]	Sequence-sequence with reinforcement learning	Answer unaware question generation with informativeness and specificity metrics	QuAC	INFO: 0.752	SPEC:0.835	24.9	–
[2]	GPT-2 small and medium	Flow-propagation based training using combined losses for n turns in the dialog	CoQA	15.78	**40.15**	**50.98**	Consistency: 0.817 Fluency:0.548

Fig. 4 Summary of automatic question generation techniques



- *METEOR (Metric for Evaluation of Translation with Explicit ORdering)*

METEOR is an alternate metric for evaluating machine translation-based texts. It was modeled to have a better correlation with the human way of judgment. It tries to remove the drawback of BLEU which impacts the scores of individual sentences as in BLEU average lengths are calculated spanning the whole corpus. For this, METEOR uses a weighted F-score based on mapping unigrams along with a function that imposes a penalty for the wrong order of words [101].

- *ROUGE (Recall-Oriented Understudy for Gisting Evaluation)*

ROUGE is a metric for automatic evaluation which is based on recall alone. It is commonly used for the evaluation of text summaries. There are different variants of ROUGE which are created based on the feature type used for computing recall. They are ROUGE-N (based on n-grams), ROUGE-L (based on longest common subsequence statistics), ROUGE-W (based on weighted longest common subsequence statistics), and ROUGE-S (skip-bigram co-occurrence) [102].

- *CIDEr (Consensus-based Image Description Evaluation*

CIDEr (Consensus-based Image Description Evaluation) is an automatic metric proposed for evaluating the quality of image description. In this metric, the model-generated sentence is compared with a set of human written sentences of ground truth. A novel metric for automatic consensus on description quality is proposed by making use of 2 datasets, PASCAL-50S and ABSTRACT-50S. The consensus makes use of up to 50 reference sentences rather than 5 in the available datasets. This metric measures how similar a model generated sentence is to the consensus of the ground truth sentences for that image. Consensus is calculated in terms of how often the majority of the sentences used to describe an image are similar. Further, CIDEr claims that the aspects of grammar, importance, accuracy and saliency are innately captured by our metric [103].

Among the automatic metrics, BLEU, METEOR and ROUGE are the more favored metrics for Standalone QG and Conversational QG. BLEU is even used often for Visual QG but the other metrics have been replaced by CIDEr for Visual QG.

4.2 Human-based evaluation

It is observed that most of the techniques used for the evaluation of the performance of AQG systems are not an effective measure of the quality of the question generated. Hence, evaluation is also performed by human evaluation techniques.

An approach in which three crowd-workers are used for rating questions based on a scale of 1 to 5 (5 being good) on two parameters fluency and relevance is employed in [12, 94]. Other approaches make use of naturalness [32, 35] and difficulty [35].

In [104], human evaluators were asked to rate the quality of the generated questions based on three factors, syntactic correctness, semantic correctness, and relevance.

Table 5 Overview of techniques used in AQG

Use-case of AQG	Techniques	Strengths	Weaknesses	References
Standalone question generation	Rule-based	1. Rules making use of templates are simple to implement 2. Simple WH questions can be framed without any difficulty	1. The no. of templates is fixed, cannot be generalized to forming any type of question 2. Complex questions cannot be generated using templates	[20–22, 25, 26, 30]
	Encoder–decoder	1. Generalized questions can be generated without requirement of templates 2. Use of bidirectional LSTM, attention, and graph-based models can boost the performance thus yielding questions of relevance to an extent	1. A few questions may be syntactically wrong or lack relevance 2. Longer sentences may not be remembered by the network well enough to be able to produce questions	[32, 35, 36] [54, 55]
	Generative adversarial networks	1. Use of variational encoders helps increase variability of questions 2. Can be used for goal-based question generation like clarification questions, distractor generation, etc.	1. Questions may be lacking syntax or may produce wrong distractors as lot depends on the training stability 2. Difficult to train and stabilize	[39, 40, 45, 48]
	Deep reinforcement learning	1. Using Reinforcement learning, the deep learning models are guiding faster towards the goal 2. Goal based QG is better suited to reinforcement learning	1. Reinforcement learning requires that the optimization 2. Can get stuck which finding optimal parameters 3. Datasets must be very large, don't give good results on smaller datasets	[70] [53, 55]
	Transformer	1. Transformer models can capture context 2. If pre-trained models are used, faster in fine-tuning	1. Computationally intensive to be loaded into memory 2. Required to limit the length of paragraphs to some tokens 3. Modification in architecture requires training from scratch which would be computationally extensive	[57, 58, 62, 68, 68, 69]
Visual question generation	Encoder–decoder	1. Improved training speed if GRU is used due to less redundancy 2. Adding exemplars improves performance	1. Problem with longer sequences 2. Multiple sources should be considered to improve training	[72, 86] [85, 88]

Table 5 (continued)

Use-case of AQG	Techniques	Strengths	Weaknesses	References
	Deep reinforcement learning	1. Goal based VQG can be targeted using suitable reward functions	1. Difficult to set the goals with complex images	[79, 89]
	Variational autoencoder	1. Joint learning of text, image and question is possible as a complex function to be modeled, enables much better results 2. Addition of categorical information helps in generating better questions	2. Some images during internal learning are blurry	[82, 90]
	Latent clustering	1. Categorization of questions is easier with latent clustering 2. Well-suited to category-based question generation	1. Deciding the number of clusters is a challenging task	[90]
	Transformer	1. Use of multiple encoders for representing objects and text makes it easy to train	1. Pre-training of transformer is required to give better results	[87]
Conversational question generation	Encoder–decoder	1. Use of encoder–decoder helps in generalizing the sequence of questions based on history of conversation 2. Addition of mechanism for tracking the context like coreference alignment improves the flow of questions	1. The amount of context is limited to a few turns in the past conversation history in RNN based encoder–decoder models 2. Use of coreferences throughout the conversation may make it difficult for the model to identify without any explicit mechanism for the same	[89, 92]
	Deep reinforcement learning	1. Use of question pattern recognition, focus estimation improves question generated 2. Reward functions can be included in the form of loss functions for making the training directed towards the goal	1. The goal to be optimized should be correctly identified to use reinforcement learning for directing the deep learning model training	[93, 96]
	Conditional NN	1. Use of specificity label allows to create questions which are of a specific category	1. There is a dependency on templates for question categorization which limits the type of questions to be generated	[95]

Table 5 (continued)

Use-case of AQG	Techniques	Strengths	Weaknesses	References
	Transformer	1. Transformer architecture helps to remember context to a larger extent instead of the usual sequence-to-sequence RNN based models 2. Pre-training on earlier data helps in shorter fine-tuning time	1. Some work needs to be done in order to include context into the loop in the form of rewards or reinforcement learning	[2]

4.3 Other evaluation techniques

Over the years, different techniques have been proposed by various researchers for the evaluation of the generated questions. One notable contribution is made in [105], where the authors first use the answerability of a question through human evaluation to modify the existing automatic evaluation metrics to include the influence of relevant words, question types, function words, and named entities. They then use the weighted average of precision and recall of these scores to get the final metric proposed by them. They further proved that their proposed metrics had a better correlation with the human evaluation scores.

Figures 5a and b represent the automatic and human-based evaluation metrics used in SQG, VQG and CQG in the research reviewed in this survey.

5 Datasets

As a result of constant efforts in this direction, many open datasets have been created for supporting research in question generation. The type of dataset chosen depends on various factors like whether the type of question generation is closed domain or open domain, whether the questions to be generated are independent, that is, one at a time or related like in a conversational system.

We categorize the datasets based on the use-cases. For SQG, there are several benchmarking datasets in both open-domain and closed-domain QG. Based on the cognitive level of the question, we can choose shallow datasets like SQuAD, NewsQA, [46, 47] or deep cognitive level datasets like LearningQ and NarrativeQA [106, 107].

For CQG, conversational datasets are used, in which question generation takes place in the form of a conversation of questions and answers. Depending on the application and the type of questions to be generated, appropriate datasets can be chosen.

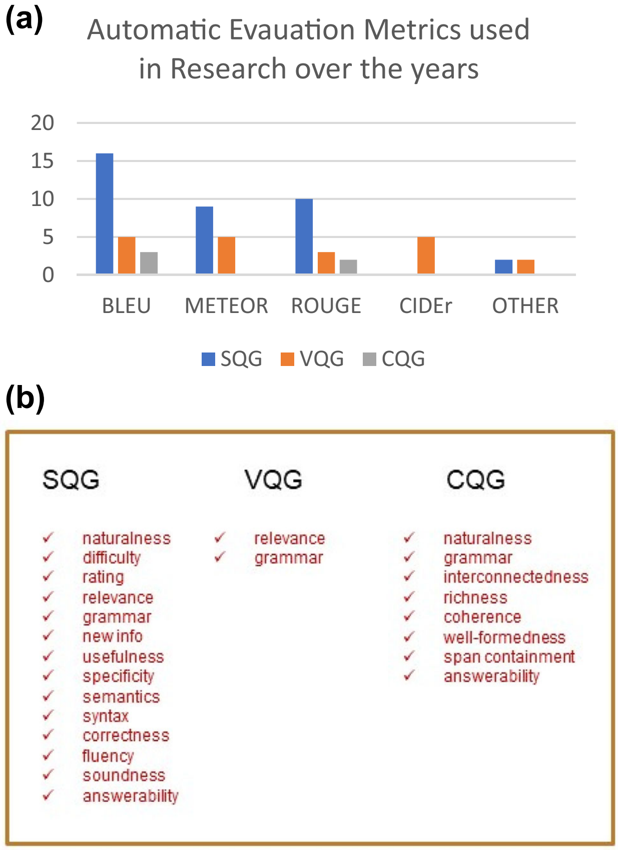


Fig. 5 **a** Automatic evaluation metrics for question generation. **b** Summary of human-based evaluation metrics for question generation

VQG Datasets require images and the questions asked are based on the objects/scenes depicted in the images. Majority of the datasets are focused on recognition of objects and images and the questions range from MCQs to yes/no and short answer questions.

Most datasets are curated by crowdsourcing and the number of training examples are sufficiently large in datasets (SQuAD, NewsQA). Some datasets are open domain like general English passages curated from Wikipedia, while other closed-domain datasets make use of news(NewsQA,

Table 6 Summary of SQG datasets for AQG systems

Dataset name	Source	Statistics			Question Type
		#Questions	#Answers	#Documents	
SQuAD (Stanford Question Answering Dataset)[46]	Wikipedia	100 k	100 k	23,215	Factoid spanning one or more sentences
CMU Q/A Dataset[108]	Wikipedia	4000	4000	150	Factoid with difficulty ratings
News QA [47]	CNN news articles	120 K	120 k	12 k	Factoid spanning one or more sentences
DeepMind Q&A Dataset [109]	CNN + Daily Mail	1 M	1 M	93 K-CNN, 220 k-Daily Mail	Cloze type
WIKIQA [110]	Bing Query logs	3047	1473	29,258 sentences	Factoid
MSMARCO [111]	Bing Query logs	1,010,916	1,026,758	8,841,823	Factoid
RACE [112]	English exams	100 k	100 k	28,000	Multiple Choice Questions spanning one or more sentences
LearningQ[106]	TED-Ed and Khan Academy	230 K	–	11 K	Spanning multiple sentences covering all Bloom's levels
NarrativeQA[107]	Stories and human-generated summaries	46,765	46,765	1572	Questions involving deep reasoning on summaries
Natural Questions [113]	Wikipedia and aggregated queries through Google Search engine	323,045	323,045	323,045	Long answer and short answer type questions

DeepMind), educational content (RACE, NarrativeQA) and some are from search queries(WIKIQA, MSMARCO).

The details of datasets used for question generation are summarized in Tables 6, 7, and 8.

We select the most commonly used benchmark dataset for each use-case among those that we surveyed and list the various models used along with the work done in Table 9.

6 Challenges and future directions

Although the application of deep learning techniques and combining Natural Language Processing with them has made a tremendous improvement in question generation systems, there are still a few challenges to consider. We discuss the various challenges and possible future directions in this field in the following section.

6.1 Challenges in AQG

We identify and discuss the various challenges in AQG in this section.

6.1.1 Quality of Questions

Generating questions with proper syntax have been accomplished with the help of employing a language model in the loop and using other similar language-related features. However, the questions generated lack to a certain extent in terms of semantics and relevance as reported in most studies through human evaluation [32, 35, 36, 39, 56, 61]. Also, generating meaningful questions is a challenge as most existing techniques focus more on the syntactical aspects rather than information extracting questions. Syntax plays an important part but generating questions which make sense in extraction of meaningful information [99] is an important requirement in many applications and must be explored extensively.

6.1.2 Types of questions

The type of questions to be generated range from the typical short answer span questions to multiple spans for question generation from the text [42, 47, 106, 107, 112]. If we consider the case of images, most systems try to identify objects placed in various scenes or images, this task has been solved to quite an extent, although the kind of models which are able to extract meaningful information from an image are limited

Table 7 Summary of CQG datasets for AQG systems

Dataset name	Source	Statistics		Question type
		#Question–answer pairs	#Conversations	
CoQA [92]	Conversation between two crowd workers to chat about a passage in the form of questions and answers	127 k	8 k	seven diverse domains
QuAC (Question Answering inContext) [114]	Conversations created through crowd worker student–teacher pairs where the student poses a question about a hidden Wikipedia text and the teacher who answers the question using short spans from the text	100 K	14 k	Short-span answers
ShARC [115]	Crowdsourced task instances based on real-world rules	32 k	6.6 k	Varied types of reasoning with the answer not present in the given text

Table 8 Summary of VQG datasets for AQG Systems

Dataset name	Source	Statistics		Domain type	Question type
		#Images	#Question–Answer Pairs		
DAQUAR [116]	Synthetic and human-based QA pairs built on NYU-Depth V2 [117] dataset	1449	12 k	Colors, objects	One word answers
VQA [84]	MS-COCO [118] + abstract scenes	204,721-images 50,000-abstract scenes	614 k questions and 7984 k answers—images 150 K questions and 195 k answers-abstract scenes	Objects, animals	One word answers, Yes/No questions
VQA2 [119]	Crowdsourcing to generate complementary images from VQA	443 K train, 214 K validation, and 453 K test (question, image) pairs		Objects, animals	
Visual7W [120]	Crowdsourcing to generate QA pairs on images	47,300	327,939	Objects	7 W multiple-choice
VizWiz [121]	Images of photos clicked by blind people and ask spoken questions	31 k	31 k Questions, 248 k answers	Objects	Short answers, yes/no, unanswerable
Visual Genome (pairs) [122]	Crowdsourcing	108 K	1.7 M	Objects, attributes, relationships	Freeform, region-based, one word answers
CLEVR [123]	Crowdsourcing	100 k	850 k	Objects, attributes, relationships	Short answers

Table 9 Overview of benchmark datasets: models and work done

Dataset name	References	Models used	Work done
SQuAD (Stanford question answering dataset) [46]	[30]	RNN encoder–decoder framework with attention, LSTM sentence level and paragraph-level information	Paragraph-level information and attention for factoid questions
	[38]	Sequence-to-sequence model with conditional variational autoencoders as generator of GAN	Question type information to categorize questions
	[43]	Doubly adversarial network	Domain independent questions
	[67]	Generator evaluator framework with reward metrics using deep reinforcement learning	Factoid questions using optimization of standard metrics and reward functions
	[53]	Graph-to-Sequence BERT embedding, RL loss, Deep Alignment Network	Use of syntactic and semantic graphs to generate questions
	[54]	Bi-directional RNN for QA, encoder–decoder for QG	Joint training task of question-answering and question generation
	[55]	Seq-Seq + Attention, pointer softmax decoder, answer words sequence	Joint training task of question-answering and question generation
	[57]	Transformers	Shorter questions syntactically correct
	[58]	Transformer + placeholding + copying + ELMO	Use of ELMO for representation of tokens, placeholding for named entities
	[68]	GPT-2(small) + attention	Pre-trained model fine-tuning
	[62]	Transformer-based decoder of GPT-2 + Transformer encoder of BERT	Combination of different encoder–decoder transformer architectures
	[65]	BERT highlight question generation	Dual BERT as encoder and decoder both for sentence and paragraph level QG
	[69]	T5 + BART	Multiple question types using single hybrid architecture
	[82]	Variational autoencoder	Mutual information of the image, the generated question, the expected answer
VQA [84] VQA2 [119]	[82]	Deep exemplar network	Attention-fused exemplars for QG
	[84]	Transformer-based text and image encoder and text decoder VQG	Use of category and image features for encoding in transformer
	[86]	Reinforcement learning	Bi-discriminators: natural and human-written as rewards
	[90]	Variational autoencoder	Use of category to generate questions
CoQA [92]	[89]	Encoder–decoder	Coreference alignment, Multi-source encoder
	[92]	Answer unaware encoder–decoder	Question patterns for classification and generation
	[93]	Deep reinforcement learning	Reinforced dynamic reasoning
	[95]	Conditional NN	Conditional encoder–decoder with conceptual class mapping of questions
	[2]	GPT-2 small and medium	Training based on combined losses for n turns in the dialog

to trivial use-cases (refer Table 7). Another direction in which research could progress is generating relevant questions for the text given a topic. Although there are a few approaches where topic-based questions [99] have been looked at, there is no existing approach that completely solves the problem.

6.1.3 Datasets challenge

Most datasets that are currently available for training question generation systems are crowd-sourced [42, 47, 84, 92, 114, 115, 119, 120, 122, 123], which largely impacts the quality of generated questions. Also, many of the datasets are very generic rather than contoured to specific domains (refer Tables 6, 7, 8). Domain-specific datasets must be generated keeping in mind the quality of content while curating the dataset. It would also be important to note that for conversational question generation, only a few datasets are available. As a result, not much work is done in this use-case. This is a potential way of adding to the research community so that the models to be tested are provided with high-quality data for specific purposes.

6.1.4 Metrics challenge

Another area of working towards this field is building some metrics for a thorough evaluation of the generated questions. Although standard metrics for evaluating text generated like BLEU, METEOR, ROUGE, CIDEr can be used for automatic evaluation for generated questions, a more relevant metric which includes other factors like naturalness in the language used, weightage given to the syntax and grammar of the question generated can be experimented with. Although some research work uses such metrics in the form of human evaluation [2, 28, 31, 32, 35, 43, 57, 65, 67, 75, 84–86, 91–93], we need to devise efficient metrics which automatically give an estimate of the quality of the questions generated, thus eliminating the need for human evaluation.

6.2 Future research directions

6.2.1 Transfer learning

What happens in case of training data is that most of the datasets available are curated from open domain data like Wikipedia, reddit, social networking platforms and the like. Some works have recently focused on the transfer of training of one domain to another. For instance, in [124], transfer learning is performed by training on non-educational datasets like SQuAD and NQA (Natural Questions) and the evaluation is performed on an author curated dataset called TQA-A with questions based on educational text and tagged with answers. Several pre-trained BERT-based models were explored and, answer selection was investigated. This study found that there was a significant difference in which the answers were selected in educational and non-educational question generation. Transfer learning helps in cases where the data for training is less, and we have pre-trained models which exist on other similar datasets. With several such pre-trained architectures available, it is very useful to employ transfer learning for question generation.

6.2.2 Creating corpora of high quality

As most of the datasets for question generation are either crowd-sourced or borrowed from open-sourced communities like Wikipedia, Reddit and others. These are not suitable to domain-based question generation. Specialized domains like education and medical domains have requirements other than only focusing on purely generation of relevant questions. For instance, educational domain would require generating questions at a specific cognitive level or mapping to a a specific category. On the other hand, medical domain would require questions aimed at correct diagnosis. One such effort worthy of mention is [125]. Here, the authors have curated a dataset from discharge summaries with the help of 10 experts in the medical domain to construct 2000+ questions related to the diagnosis. After analysis of the type of questions, they used pre-trained transformer models to train on the dataset and achieved promising results. Hence, domain-based corpora of high quality need to be created and current approaches could be further improved by working in this direction.

6.2.3 Multimodality for QG

Visual question generation directs the research in question generation towards the aspect of multimodality. Multimodal question generation involves using different input modalities including text, images, videos for generation of questions. Using multiple modalities, more real-world applications can be targeted, including generating questions based on pictures and diagrams in educational domain [126] and helping the visually challenged identify objects around them [127]. Video QG is an upcoming area, where some works have been introduced lately. In [128], using a newly constructed architecture of a generator, which generates a question given a video clip and an answer, and a pre-tester, which tries to answer the generated question, has been investigated for joint QA-QG from videos. They make use of Video encoder for extracting video features of 20 frames from each video and later encode them using faster R-CNN and Resnet101. Overall, they obtained promising results on the TVQA [129] and ActivityNetQA [130] datasets.

6.2.4 Working on QG centric metrics

Metrics for QG are essential to gauge the quality of generated questions in terms of how meaningful they are. Although the standard metrics used to evaluate the generated questions are the ones used generally for text generation. QG metrics should include grammatical correctness, question well-formedness and domain centric metrics depending on the application being used in. Human-based evaluation is currently in use for the above aspects, although some efforts in this direction are being made recently. In [131], an evaluation metric called QAScore is proposed which makes use of pre-trained language model RoBERTa (Robustly Optimized BERT Pre-training Approach). The metric is reference-free and correlates well with human judgments as observed in the experimentation performed by the authors. Similar studies would be helpful for strengthening the metrics for QG evaluation.

Table 10 Summary of applications of AQG systems

Type of AQG	References	Application domain	Work done	Description of research	Question type	DataSet
Standalone question generation	[138]	Deep reasoning on text	Inquisitive question generation	High level text comprehension question generation for deeper understanding of the news articles by GPT-2 model	Why type reasoning questions	INQUISITIVE (Introduced)
	[139]	Education	Question Generation for Adaptive Education	Pre-trained language models combined with sequence-to-sequence models for adaptive learning on language translation tasks using a target difficulty level	Question generation for language translation	Duolingo [140]
	[141]	News quiz	Quiz-style question generation for news stories	Question answer generation from paragraph summaries using modified PEGASUS [142] transformer model with minimum reference loss and mid training as additional parameters	Multiple Choice Quiz	NewsQuizQA (Introduced)
	[143]	Social media	Generation of poll questions for posts in social media	Poll questions based on social media posts using a dual decoder based sequence to sequence architecture	Factoid	Sina Weibo Chinese Microblogging
	[144]	Indonesian language	Question generation in indonesian language	Factual questions generated in Indonesian language based on translated SQUAD dataset using neural models	Factoid	SQuAD 2.0 translated to Indonesian
Conversational question generation	[145]	Medicine-COVID-19	Summarizing a corpus using questions	Question generation for summarizing multiple documents using T5 base transformer model	Factoid	CORD-19 [146]

Table 10 (continued)

Type of AQG	References	Application domain	Work done	Description of research	Question type	DataSet
	[147]	Multi-document question generation	Contrastive multi-document question generation	Multi-source coordinated question generation using a combined approach of supervised and reinforcement learning for GPT-2. Use of positive as well as negative documents during the training	Factoid	MS-MARCO
Visual question generation	[148]	Visual dialog generation	Large-scale answerer in questioner’s mind	Context based question generation for task-oriented visual dialogs by maximizing information gain in an RNN based model	Factoid	Guess Which [149]
	[150]	Visual dialog generation	Goal-oriented visual dialog Generation	Question generation for images in GuessWhat dataset using a visual state estimator which is based on the answer for establishing a goal-oriented dialog to guess the correct answer	Yes–No	Guess What
	[151]	Multimedia based learning	Use of NLP based question generation for assisting multimedia-based learning	Use of text-based and image-based pre-questions for facilitating effective learning. Evaluation by testing performance of students in answering of post-questions based on the video after watching the video	Factoid-WH	Documentary videos and transcripts

7 Applications of automatic question generation

There are several applications of question generation systems. A very important use case is generating questions for passages. This can be useful in an educational setup where the input will be passages of text and will result in saving the time and effort required for setting question papers. In [132], a desiderata for generating cloze type and WH-questions has been discussed. In [133], a system for generating questions is discussed, by generating simple factoid questions using syntactic rules by question types. Classification schemes for questions are presented in [134] while asking students to generate questions themselves for improving meta-cognitive abilities has been discussed in [135].

Closed-domain question generation can be applied to healthcare bots where the bots can interview patients for specific symptoms for the preliminary investigation of diseases. A conversational bot with the ability to generate relevant questions in natural language can thus aid in speeding up the process of diagnosing a patient. Visual question generation, for instance, has been used for generating meaningful questions on radiology images in [136].

Open-domain generation systems can be used for working across use-cases that are not restricted to a limited collection of scenarios. Such systems can be used where it is difficult to comprehend and chalk out an exact flow of events like for example, open-ended conversational agents for their question generation capability.

Table 10 shows an overview of various application-based research work carried out in this field. If we consider the different domains, standalone question generation has been applied to education [123], news [124] and social media [126]. Another interesting application of standalone QG is design of reference-less metrics for summaries. For this purpose, a recent work, which was proposed in [137], attempts to use a reference less metric for text-to-text evaluation tasks. In their approach, a question generation model is first trained on SQuAD and then synthetic questions are generated using the trained QG model on a dataset which includes structured-input and a textual description, which is multimodal in nature. This synthetic dataset is then used to train on QA-QG models. The resulting model metric is directly used to compare generated summaries with source text. Conversational question generation largely has seen applications in the generation of questions from multiple documents field as also the medical domain. Visual question generation has been deployed in visual dialogue generation in a few works [131, 133, 134]. For a few of these applications, some datasets were used like SQuAD, MS-MARCO and the like. Some works curated their own dataset owing to the distinctive application involved [122, 124]. Also, most works used factoid-based questions but Yes/No [133], MCQ type [124] and reasoning type questions [122] were also generated for some applications. Most works used RNN-based architectures with additional features and a few used transformer models.

8 Conclusion

In this survey, we presented an overview of the literature for the generation of automatic questions. We classified the methodologies for question generation based on three broad use-cases: standalone question generation, visual question generation and conversational question generation. We also reviewed the different datasets being used for the task. Several challenges and applications of such systems are discussed and summarized. As presented in the survey, most question generation systems today have worked on generating questions from the text. There are a few aspects that are yet to be addressed. For example, questions generated lack naturalness and sometimes are meaningless in the sense of information extraction. Some improvements can be made in generating semantically relevant and information-seeking questions. Also, justifiable metrics for evaluating the quality of questions is still a work in progress. Multiple input modalities are being considered of late and the impact of incorporating them is being studied. There is a need to develop models which are an amalgamation of several techniques considering each aspect and at the same time be relevant to the application being addressed.

References

1. Kumar, V., Chaki, R., Talluri, S.T., Ramakrishnan, G., Li, Y.-F., Haffari, G.: Question generation from paragraphs: a tale of two hierarchical models. (2019)
2. Gu, J., Mirshekari, M., Yu, Z., Sisto, A.: ChainCQG: Flow-aware conversational question generation. In: EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference. pp. 2061–2070 (2021)
3. Kunichika, H., Katayama, T., Hirashima, T., Takeuchi, A.: Automated question generation methods for intelligent english learning systems and its evaluation. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. **62**, 1209–1215 (2005)
4. Mostow, J., Chen, W.: Generating instruction automatically for the reading strategy of self-questioning. Front. Artif. Intell. Appl. **200**, 465–472 (2009). https://doi.org/10.3233/978-1-60750-028-5-465
5. Wang, W., Hao, T., Liu, W.: Automatic question generation for learning evaluation in medicine. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 4823 LNCS, pp. 242–251 (2008). https://doi.org/10.1007/978-3-540-78139-4_22
6. Shen, S., Li, Y., Du, N., Wu, X., Xie, Y., Ge, S., Yang, T., Wang, K., Liang, X., Fan, W.: On the generation of medical question-answer pairs. In: AAAI 2020 - 34th AAAI Conference on Artificial

Intelligence. pp. 8822–8829 (2020). https://doi.org/10.1609/aaai.v34i05.6410
7. Araki, J., Rajagopal, D., Sankaranarayanan, S., Holm, S., Yamakawa, Y., Mitamura, T.: Generating questions and multiple-choice answers using semantic analysis of texts. In: COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers. pp. 1125–1136 (2016)
8. Varga, A., Ha, L.: Wlv: A question generation system for the qgstec 2010 task b. Proceedings of the Third Workshop on Question Generation. pp. 80–83 (2010)
9. Huang, Y., He, L.: Automatic generation of short answer questions for reading comprehension assessment. Nat. Lang. Eng. **22**, 457–489 (2016). https://doi.org/10.1017/S1351324915000455
10. Flor, M., Riordan, B.: A Semantic Role-based Approach to Open-Domain Automatic Question Generation. pp. 254–263 (2018). https://doi.org/10.18653/v1/w18-0530
11. Serban, I.V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., Bengio, Y.: Generating factoid questionswith recurrent neural networks: the 30M factoid question-answer corpus. In: 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers. 1, 588–598 (2016). https://doi.org/10.18653/v1/p16-1056
12. Du, X., Cardie, C.: Identifying where to focus in reading comprehension for neural question generation. In: EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings. pp. 2067–2073 (2017). https://doi.org/10.18653/v1/d17-1219
13. Zhou, W., Zhang, M., Wu, Y.: Question-type driven question generation. In: EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. pp. 6032–6037 (2020). https://doi.org/10.18653/v1/d19-1622
14. Pan, L., Lei, W., Chua, T.-S., Kan, M.-Y.: Recent Advances in Neural Question Generation. (2019)
15. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**, 121–204 (2020). https://doi.org/10.1007/s40593-019-00186-y
16. Das, B., Majumder, M., Phadikar, S., Sekh, A.A.: Automatic question generation and answer assessment: a survey. Res. Pract. Technol. Enhanc. Learn. (2021). https://doi.org/10.1186/s41039-021-00151-1
17. Patil, C., Patwardhan, M.: Visual question generation: the state of the art. ACM Comput. Surv. (2020). https://doi.org/10.1145/3383465
18. Lehnert, W.: A Conceptual Theory of Question Answering. Ijcai-77. (1977)
19. Graesser, A.C., Person, N.K.: Question asking during tutoring. Am. Educ. Res. J. **31**, 104–137 (1994). https://doi.org/10.3102/00028312031001104
20. Wyse, B., Piwek, P.: Generating questions from OpenLearn study units. pp. 66–73 (2009)
21. Heilman, M., Smith, N.A.: Good question! Statistical ranking for question generation. In: NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference. pp. 609–617 (2010)
22. Becker, L., Basu, S., Vanderwende, L.: Mind the gap: Learning to choose gaps for question generation. In: NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference. pp. 742–751 (2012)
23. Lloret, E., Palomar, M.: Text summarisation in progress: a literature review. Artif. Intell. Rev. **37**, 1–41 (2012). https://doi.org/10.1007/s10462-011-9216-z
24. Saggion, H.: Automatic summarization: an overview. Revue Francaise de Linguistique Appliquee. **13**, 63–81 (2008). https://doi.org/10.3917/rfla.131.0063
25. Nenkova, A., Vanderwende, L., McKeqwn, K.: A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In: Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2006, pp. 573–580 (2006)
26. Bishop, C.: Pattern Recognition and Machine Learning. (2006)
27. Labutov, I., Basu, S., Vanderwende, L.: Deep questions without deep understanding. In: ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference. vol. 1, pp. 889–898 (2015). https://doi.org/10.3115/v1/p15-1086
28. Mazidi, K., Tarau, P.: Infusing NLU into automatic question generation. In: INLG 2016 - 9th International Natural Language Generation Conference, Proceedings of the Conference. pp. 51–60 (2016). https://doi.org/10.18653/v1/w16-6609
29. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Comparative Biochemistry and Physiology—Part B: Biochemistry and (1973)
30. Soleymanzadeh, K.: Domain specific automatic question generation from text. In: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop. pp. 82–88 (2017). https://doi.org/10.18653/v1/P17-3014
31. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Adv. Neural Inf. Process. Syst. **4**, 3104–3112 (2014)
32. Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. In: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). vol. 1, pp. 1342–1352 (2017). https://doi.org/10.18653/v1/P17-1123
33. Rush, A.M., Chopra, S., Weston, J.: A Neural Attention Model for Abstractive Sentence Summarization. In: Proceedings of EMNLP 2015. pp. 1–11 (2015)
34. Koehn, P., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07. p. 177 (2007)
35. Hu, W., Liu, B., Ma, J., Zhao, D., Yan, R.: Aspect-based question generation. In: 6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings. pp. 1–10 (2018)
36. Indurthi, S., Raghu, D., Khapra, M.M., Joshi, S.: Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference. vol. 1, pp. 376–385 (2017). https://doi.org/10.18653/v1/e17-1036
37. Zhao, S., Wang, H., Li, C., Liu, T., Guan, Y.: Automatically generating questions from queries for community-based question answering. In: Proceedings of 5th International Joint Conference on Natural Language Processing. pp. 929–937 (2011)
38. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative

adversarial networks. In: Neural Information Processing Systems (2014)

39. Liang, C., Yang, X., Wham, D., Pursel, B., Passonneau, R., Giles, C.L.: Distractor generation with generative adversarial nets for automatically creating fill-in-the-blank questions. In: Proceedings of the Knowledge Capture Conference, K-CAP 2017. 4–7 (2017). https://doi.org/10.1145/3148011.3154463
40. Yao, K., Zhang, L., Luo, T., Tao, L., Wu, Y.J.: Teaching machines to ask questions. In: IJCAI International Joint Conference on Artificial Intelligence. 2018-July, pp. 4546–4552 (2018). https://doi.org/10.24963/ijcai.2018/632
41. Zhao, T., Zhao, R., Eskenazi, M.: Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). pp. 654–664 (2017)
42. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Stroudsburg, PA, USA (2016)
43. Heilman, M. (Carnegie M.U.: Automatic Factual Question Generation from Text, https://lti.cs.cmu.edu/sites/default/files/research/thesis/2011/michael_heilman_automatic_factual_question_generation_for_reading_assessment.pdf, (2011)
44. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015)
45. Bao, J., Gong, Y., Duan, N., Zhou, M., Zhao, T.: Question generation with doubly adversarial nets. IEEE/ACM Trans. Audio Speech Lang. Process. **26**, 2230–2239 (2018). https://doi.org/10.1109/TASLP.2018.2859777
46. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD : 100,000+ Questions for Machine Comprehension of Text. (2015)
47. Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., Suleman, K.: NewsQA: A Machine Comprehension Dataset. pp. 191–200 (2017). https://doi.org/10.18653/v1/w17-2623
48. Rao, S., Daumé, H.: Answer-based adversarial training for generating clarification questions. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. vol. 1, pp. 143–155 (2019)
49. McAuley, J., Yang, A.: Addressing complex and subjective product-related queries with customer reviews. In: 25th International World Wide Web Conference, WWW 2016. pp. 625–635 (2016)
50. McAuley, J., Targett, C., Shi, Q., Van Den Hengel, A.: Image-based recommendations on styles and substitutes. In: SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 43–52 (2015)
51. Rao, S., Daumé, H.: Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). pp. 2737–2746 (2018)
52. Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., Jurafsky, D.: Deep reinforcement learning for dialogue generation. In: EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings. pp. 1192–1202 (2016)
53. Kumar, V., Ramakrishnan, G., Li, Y.-F.: Putting the Horse Before the Cart:A Generator-Evaluator Framework for Question Generation from Text. arXiv:1808.04961 (2018)
54. Kumar, V., Boorla, K., Meena, Y., Ramakrishnan, G., Li, Y.F.: Automating reading comprehension by generating question and answer pairs. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 335–348 (2018)
55. Liu, Y., Zhang, C., Yan, X., Chang, Y., Yu, P.S.: Generative question refinement with deep reinforcement learning in retrieval-based QA system. In: International Conference on Information and Knowledge Management, Proceedings. pp. 1643–1652 (2019)
56. Chen, Y., Wu, L., Zaki, M.J.: Reinforcement Learning Based Graph-To-Sequence Model for Natural Question Generation. In: Iclr 2020. pp. 498–515 (2020)
57. Wang, T., Yuan, X., Trischler, A.: A Joint Model for Question Answering and Question Generation. Presented at the (2017)
58. Tang, D., Duan, N., Qin, T., Yan, Z., Zhou, M.: Question Answering and Question Generation as Dual Tasks. Presented at the (2017)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł, Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **2017**, 5999–6009 (2017)
60. Kriangchaivech, K., Wangperawong, A.: Question Generation by Transformers. (2019)
61. Scialom, T., Piwowarski, B., Staiano, J.: Self-attention architectures for answer-agnostic neural question generation. In: ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. 6027–6032 (2020). https://doi.org/10.18653/v1/p19-1604
62. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. pp. 2227–2237 (2018)
63. Lopez, L.E., Cruz, D.K., Cruz, J.C.B., Cheng, C.: Simplifying Paragraph-level Question Generation via Transformer Language Models. (2020)
64. Wolf, T., Debut, L., Sanh, V., Chaumond, J.: Huggingface's transformers: State-of-the-art natural language processing. In: arXiv preprint arXiv:1910.03771 (2019)
65. Klein, T., Nabi, M.: Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds. (2019)
66. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. Presented at the (2018)
67. Devlin, J., Chang, M.-W., Lee, K., Google, K.T., Language, A.I.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Naacl-Hlt 2019 (2018)
68. Chan, Y.-H., Fan, Y.-C.: A Recurrent BERT-based Model for Question Generation. pp. 154–162 (2019). Doi: https://doi.org/10.18653/v1/d19-5821
69. Murakhovs'ka, L., Wu, C.S., Laban, P., Niu, T., Liu, W., Xiong, C.: MixQG: Neural Question Generation with Mixed Answer Types. Findings of the Association for Computational Linguistics: NAACL 2022 - Findings. 1486–1497 (2022). https://doi.org/10.18653/v1/2022.findings-naacl.111
70. Kumar, V., Ramakrishnan, G., Li, Y.-F.: A framework for automatic question generation from text using deep reinforcement learning. In: arXiv. 2019 IJCAI Workshop SCAI: The 4th International Workshop on Search-Oriented Conversational AI (2018)
71. Lopez, L.E., Cruz, D.K., Cruz, J.C.B., Cheng, C.: Simplifying Paragraph-level Question Generation via Transformer Language Models. Presented at the (2020)
72. Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L.: Generating natural questions about an image. In:

54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers. 3, 1802–1813 (2016). https://doi.org/10.18653/v1/p16-1170

73. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 740–755 (2014)
74. Huang, T.H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C.L., Parikh, D., Vanderwende, L., Galley, M., Mitchell, M.: Visual storytelling. In: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference. 1233–1239 (2016). https://doi.org/10.18653/v1/n16-1147
75. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G.: From captions to visual concepts and back. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1473–1482 (2015)
76. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. pp. 1724–1734 (2014)
77. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. In: ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference. pp. 100–105 (2015)
78. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2015)
79. Zhang, J., Wu, Q., Shen, C., Zhang, J., Lu, J., van den Hengel, A.: Goal-Oriented Visual Question Generation via Intermediate Rewards. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 11209 LNCS, pp. 189–204 (2018). https://doi.org/10.1007/978-3-030-01228-1_12
80. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: GuessWhat?! Visual object discovery through multi-modal dialogue. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. pp. 4466–4475 (2017)
81. Strub, F., De Vries, H., Mary, J., Piot, B., Courvile, A., Pietquin, O.: End-to-end optimization of goal-driven and visually grounded dialogue systems. In: IJCAI International Joint Conference on Artificial Intelligence. pp. 2765–2771 (2017)
82. Krishna, R., Bernstein, M., Fei-Fei, L.: Information maximizing visual question generation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019-June, 2008–2018 (2019). https://doi.org/10.1109/CVPR.2019.00211
83. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. (2014)
84. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. 2015 Inter, pp. 2425–2433 (2015). https://doi.org/10.1109/ICCV.2015.279
85. Patro, B.N., Namboodiri, V.P.: Deep Exemplar Networks for VQA and VQG. pp. 1–26 (2019)
86. Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X., Zhou, M.: Visual Question Generation as Dual Task of Visual Question Answering. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 6116–6124 (2018). https://doi.org/10.1109/CVPR.2018.00640
87. Vedd, N., Wang, Z., Rei, M., Miao, Y., Specia, L.: Guiding Visual Question Generation. In: NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference. pp. 1640–1654 (2022). https://doi.org/10.18653/v1/2022.naacl-main.118
88. Lee, J., Arora, S.: A Free Lunch in Generating Datasets: Building a VQG and VQA System with Attention and Humans in the Loop. (2019)
89. Fan, Z., Wei, Z., Wang, S., Liu, Y., Huang, X.: A reinforcement learning framework for natural question generation using bi-discriminators. Coling. pp. 1763–1774 (2018)
90. Uppal, S., Madan, A., Bhagat, S., Yu, Y., Shah, R.R.: C3VQG: Category consistent cyclic visual question generation. In: Proceedings of the 2nd ACM International Conference on Multimedia in Asia, MMAsia 2020. (2021). https://doi.org/10.1145/3444685.3446302
91. Gao, Y., Li, P., King, I., Lyu, M.R.: Interconnected question generation with coreference alignment and conversation flow modeling. In: ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. pp. 4853–4862 (2020). https://doi.org/10.18653/v1/p19-1480
92. Reddy, S., Chen, D., Manning, C.D.: CoQA: A Conversational Question Answering Challenge. In: Transactions of the Association for Computational Linguistics. pp. 249–266 (2019)
93. See, A., Liu, P.J., Manning, C.D.: Get To The Point: Summarization with Pointer-Generator Networks. In: 55th Annual Meeting of the Association for Computational Linguistics. pp. 1073–1083 (2017)
94. Du, X., Cardie, C.: Harvesting paragraph-level question-answer pairs from wikipedia. In: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). 1, 1907–1917 (2018). https://doi.org/10.18653/v1/p18-1177
95. Pan, B., Li, H., Yao, Z., Cai, D., Sun, H.: Reinforced dynamic reasoning for conversational question generation. In: ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. 2114–2124 (2020). https://doi.org/10.18653/v1/p19-1203
96. Nakanishi, M., Kobayashi, T., Hayashi, Y.: Towards Answer-unaware Conversational Question Generation. pp. 63–71 (2019). Doi: https://doi.org/10.18653/v1/d19-5809
97. Colclough, M., Lehnert, W.G.: The Process of Question Answering -- A Computer Simulation of Cognition. (1979)
98. Krishna, K., Iyyer, M.: Generating question-answer hierarchies. In: ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. pp. 2321–2334 (2020). https://doi.org/10.18653/v1/p19-1224
99. Qi, P., Zhang, Y., Manning, C.D.: Stay Hungry, Stay Focused: Generating Informative and Specific Questions in Information-Seeking Conversations. pp. 25–40 (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.3
100. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2001)
101. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In:

Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
102. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Proceedings of the workshop on text summarization branches out (WAS 2004). pp. 25–26 (2004)
103. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 07–12-June, pp. 4566–4575 (2015). https://doi.org/10.1109/CVPR.2015.7299087
104. Liu, Z., Huang, K., Huang, D., Zhao, J.: Semantics-reinforced networks for question generation. Front. Artif. Intell. Appl. **325**, 2078–2084 (2020). https://doi.org/10.3233/FAIA200330
105. Nema, P., Khapra, M.M.: Towards a better metric for evaluating question generation systems. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. pp. 3950–3959 (2020). https://doi.org/10.18653/v1/d18-1429
106. Chen, G., Yang, J., Hauff, C., Houben, G.J.: LearningQ: A large-scale dataset for educational question generation. In: 12th International AAAI Conference on Web and Social Media, ICWSM 2018. pp. 481–490 (2018)
107. Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E.: The NarrativeQA reading comprehension challenge. Trans. Assoc. Comput. Linguist. **6**, 317–328 (2018). https://doi.org/10.1162/tacl_a_00023
108. Smith, N.A., Heilman, M., Hwa, R.: Question generation as a competitive undergraduate course project. In: In Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge. pp. 4–6 (2008)
109. Hermann, K.M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. Adv. Neural Inf. Process.. Syst. **2015**, 1693–1701 (2015)
110. Meek, W.Y.C.: W IKI QA : A Challenge Dataset for Open-Domain Question Answering. pp. 2013–2018 (2018)
111. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: a human generated MAchine reading COmprehension dataset. CEUR Workshop Proc. **1773**, 1–11 (2016)
112. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: Large-scale ReAding comprehension dataset from examinations. In: EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings. pp. 785–794 (2017). https://doi.org/10.18653/v1/d17-1082
113. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguist. **7**, 453–466 (2019). https://doi.org/10.1162/tacl_a_00276
114. Choi, E., He, H., Yatskar, M., Yih, W., Choi, Y., Liang, P., Zettlemoyer, L.: QuAC : Question answering in context. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2174–2184 (2018)
115. Saeidi, M., Bartolo, M., Lewis, P., Singh, S., Rocktäschel, T., Sheldon, M., Bouchard, G., Riedel, S.: Interpretation of natural language rules in conversational machine reading. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. 1, 2087–2097 (2020). https://doi.org/10.18653/v1/d18-1233
116. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. Adv. Neural Inf. Process Syst. **2**, 1682–1690 (2014)
117. Indoor Segmentation and Support Inference from RGBD Images.
118. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 8693 LNCS, pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48
119. Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA Matter: elevating the role of image understanding in visual question answering. Int. J. Comput. Vis. **127**, 398–414 (2019). https://doi.org/10.1007/s11263-018-1116-0
120. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7W: Grounded question answering in images. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016-Decem, pp. 4995–5004 (2016). https://doi.org/10.1109/CVPR.2016.540
121. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: VizWiz grand challenge: answering visual questions from blind people. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3608–3617 (2018). Doi: https://doi.org/10.1109/CVPR.2018.00380
122. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**, 32–73 (2017). https://doi.org/10.1007/s11263-016-0981-7
123. Johnson, J., Fei-Fei, L., Hariharan, B., Zitnick, C.L., van der Maaten, L., Girshick, R.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. pp. 1988–1997 (2017). Doi: https://doi.org/10.1109/CVPR.2017.215
124. Steuer, T., Filighera, A., Tregel, T.: Investigating educational and noneducational answer selection for educational question generation. IEEE Access. **10**, 63522–63531 (2022). https://doi.org/10.1109/ACCESS.2022.3180838
125. Lehman, E., Lialin, V., Legaspi, K.Y., Sy, A.J.R., Pile, P.T.S., Alberto, N.R.I., Ragasa, R.R.R., Puyat, C.V.M., Alberto, I.R.I., Alfonso, P.G.I., Taliño, M., Moukheiber, D., Wallace, B.C., Rumshisky, A., Liang, J.J., Raghavan, P., Celi, L.A., Szolovits, P.: Learning to Ask Like a Physician. In: ClinicalNLP 2022 - 4th Workshop on Clinical Natural Language Processing, Proceedings. pp. 74–86 (2022). https://doi.org/10.18653/v1/2022.clinicalnlp-1.8
126. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. pp. 5376–5384 (2017)
127. Patel, A., Bindal, A., Kotek, H., Klein, C., Williams, J.: Generating natural questions from images for multimodal assistants. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 2270–2274 (2021)
128. Su, H.T., Chang, C.H., Shen, P.W., Wang, Y.S., Chang, Y.L., Chang, Y.C., Cheng, P.J., Hsu, W.H.: End-to-end video question-answer generation with generator-pretester network. IEEE Trans. Circuits Syst. Video Technol. **31**, 4497–4507 (2021). https://doi.org/10.1109/TCSVT.2021.3051277
129. Lei, J., Yu, L., Bansal, M., Berg, T.L.: TVQA: Localized, compositional video question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. pp. 1369–1379 (2018). https://doi.org/10.18653/v1/d18-1167
130. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: ActivityNet-QA: A dataset for understanding complex web videos

via question answering. In: 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. pp. 9127–9134 (2019). Doi: https://doi.org/10.1609/aaai.v33i01.33019127

131. Ji, T., Lyu, C., Jones, G., Zhou, L., Graham, Y.: QAScore—an unsupervised unreferenced metric for the question generation evaluation. Entropy **24**, 1514 (2022). https://doi.org/10.3390/e24111514
132. Corbett, A., Mostow, J.: Automating comprehension questions : lessons from a reading tutor. In: Workshop on the Question Generation Shared Task and Evaluation Challenge (2008)
133. Varga, A., Ha, L.: Wlv: A question generation system for the qgstec 2010 task b. In: Proceedings of the Third Workshop on Question Generation. pp. 80–83 (2010)
134. Graesser, A.C., Rus, V., Cai, Z.: Question classification schemes. In: The Workshop on Question Generation. pp. 8–9 (2008)
135. Yu, F.Y., Liu, Y.H., Chan, T.W.: A web-based learning system for question-posing and peer assessment. Innov. Educ. Teach. Int. **42**, 337–348 (2005). https://doi.org/10.1080/14703290500062557
136. Sarrouti, M., ben Abacha, A., Demner-Fushman, D.: Visual question generation from radiology images. In: Proceedings of the First Workshop on Advances in Language and Vision Research. pp. 12–18. Association for Computational Linguistics, Stroudsburg, PA, USA (2020)
137. Rebuffel, C., Scialom, T., Soulier, L., Piwowarski, B., Lamprier, S., Staiano, J., Scoutheeten, G., Gallinari, P.: Data-QuestEval: A reference-less metric for data-to-text semantic evaluation. In: EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings. 8029–8036 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.633
138. Ko, W.-J., Chen, T., Huang, Y., Durrett, G., Li, J.J.: Inquisitive Question Generation for High Level Text Comprehension. pp. 6544–6555 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.530
139. Srivastava, M., Goodman, N.: Question Generation for Adaptive Education. pp. 692–701 (2021). https://doi.org/10.18653/v1/2021.acl-short.88
140. Settles, B., Brust, C., Gustafson, E., Hagiwara, M., Madnani, N.: Second Language Acquisition Modeling. Presented at the (2018)
141. Lelkes, A.D., Tran, V.Q., Yu, C.: Quiz-style question generation for news stories. Association for Computing Machinery (2021)
142. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization. In: 37th International Conference on Machine Learning, ICML 2020. PartF16814, pp. 11265–11276 (2020)
143. Lu, Z., Ding, K., Zhang, Y., Li, J., Peng, B., Liu, L.: Engage the Public: Poll Question Generation for Social Media Posts. pp. 29–40 (2021). https://doi.org/10.18653/v1/2021.acl-long.3
144. Muis, F.J., Purwarianti, A.: Sequence-to-sequence learning for indonesian automatic question generator. In: 2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020. (2020). https://doi.org/10.1109/ICAICTA49861.2020.9429032
145. Surita, G., Nogueira, R., Lotufo, R.: Can questions summarize a corpus? Using question generation for characterizing COVID-19 research. pp. 1–11 (2020)
146. Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D.,Wang, K.,Wilhelm, C., Xie, B., Raymond, D.,Weld, D.S., Etzioni, O., Kohlmeier, S.: CORD-19: The Covid-19 Open Research Dataset. arXiv:2004.10706 (2020)
147. Cho, W.S., Zhang, Y., Rao, S., Celikyilmaz, A., Xiong, C., Gao, J., Wang, M., Dolan, B.: Contrastive multi-document question generation. In: EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference. pp. 12–30 (2021)
148. Lee, S.W., Gao, T., Yang, S., Yoo, J., Ha, J.W.: Large-scale answerer in questioner's mind for visual dialog question generation. In: 7th International Conference on Learning Representations, ICLR 2019. pp. 1–16 (2019)
149. Das, A., Kottur, S., Moura, J.M.F., Lee, S., Batra, D.: Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In: Proceedings of the IEEE International Conference on Computer Vision. 2017-Octob, 2970–2979 (2017). https://doi.org/10.1109/ICCV.2017.321
150. Xu, Z., Feng, F., Wang, X., Yang, Y., Jiang, H., Wang, Z.: Answer-Driven Visual State Estimator for Goal-Oriented Visual Dialogue. Association for Computing Machinery (2020)
151. Skalban, Y., Ha, L.A., Specia, L., Mitkov, R.: Automatic question generation in multimedia-based learning. Coling **1**, 1151–1160 (2012)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.