Das *et al. Research and Practice in Technology Enhanced Learning* (2021) 16:5
https://doi.org/10.1186/s41039-021-00151-1

Research and Practice in Technology Enhanced Learning

REVIEW Open Access

Automatic question generation and answer assessment: a survey

Bidyut Das[1*], Mukta Majumder[2], Santanu Phadikar[3] and Arif Ahmed Sekh[4]

*Correspondence:
bidyut2002in@gmail.com
[1]Department of Information Technology, Haldia Institute of Technology, Haldia, India
Full list of author information is available at the end of the article

Abstract

Learning through the internet becomes popular that facilitates learners to learn anything, anytime, anywhere from the web resources. Assessment is most important in any learning system. An assessment system can find the self-learning gaps of learners and improve the progress of learning. The manual question generation takes much time and labor. Therefore, automatic question generation from learning resources is the primary task of an automated assessment system. This paper presents a survey of automatic question generation and assessment strategies from textual and pictorial learning resources. The purpose of this survey is to summarize the state-of-the-art techniques for generating questions and evaluating their answers automatically.

Keywords: Question generation, Automatic assessment, Self learning, Self assessment, Educational assessment

Introduction

Online learning facilitates learners to learn through the internet via a computer or other digital device. Online learning is classified into three general categories depends on the learning materials: textual learning, visual learning, and audio-video learning. Online learning needs two things: the learning resources and the assessment of learners from the learning resources. The learning resources are available, and learners can able to learn from many sources on the web. On the other hand, the manual questions from the learning materials are required for the learner's assessment. To the best of our knowledge, no generic assessment system has been proposed in the literature to test the learning gap of learners from the e-reading documents. Therefore, automatic question generation and evaluation strategies can help to automate the assessment system. This article presents several techniques for automatic question generation and their answer assessment. The main contributions of this article are as follows:

- This article at first presents a few survey articles that are available in this research area. Table 1 lists the majority of the existing review articles, which described several approaches for question generation. Table 2 presents the survey articles on learner's answer evaluation techniques.

Table 1 Recent surveys on automatic question generation

Year	Existing Work	Broad Topics
2014	Le et al. 2014	Different approaches of automatic question generation for education.
2017	Divate et al. 2017	A review of automatic question generation and evaluation techniques.
2018	Ch and Saha 2018	A survey of automatic multiple-choice question generation.
2018	Amidei et al. 2018	A survey of evaluation methodologies used in automatic question generation.
2020	Kurdi et al 2020	A review of automatic question generation for educational purposes.

- The second contribution is to summarize the related existing datasets. We also critically analyzed various purposes and limitations of the use of these datasets.
- The third contribution is to discuss and summarize the existing and possible question generation methods with corresponding evaluation techniques used to automate the assessment system.

Thearrangement of the rest of the article is as follows. In the "Question Generation and Learner's Assessment" section, we describe the overview of question generation and assessment techniques. The "Related datasets" section describes the datasets used by researchers for different applications. The "Objective Question Generation" section presents the different types of objective question generation techniques. In the "Subjective Question Generation and Evaluation" section, we illustrate the existing methods of subjective question generation and their answer evaluation. The "Visual Question-Answer Generation" section describes methods of image-based question and answer generation. Finally, we present a few challenges in the "Challenges in question Generation and Answer Assessment" section and conclude the paper in the "Conclusion" section.

Question Generation and Learner's Assessment

Automatic question generation (AQG) performs a significant role in educational assessment. Handmade question creation takes much labor, time and cost, and manual answer assessment is also a time-consuming task. Therefore, to build an automatic system has attracted the attention of researchers in the last two decades for generating questions and evaluating the answers of learners (Divate and Salgaonkar 2017). All question types are broadly divided into two groups: objective question and subjective question. The objective-question asks learners to pick the right answer from two to four alternative options or provides a word/multiword to answer a question or to complete a sentence. Multiple-choice, matching, true-false, and fill-in-the-blank are the most popular assessment items in education (Boyd 1988). On the other side, the subjective question requires an answer in terms of explanation that allows the learners to compose and write a response in their own words. The two well-known examples of the subjective question

Table 2 Recent surveys on automatic answer evaluation

Year	Existing Work	Broad Topics
2010	Rozali et al. 2010	A survey on adaptive qualitative assessment
2013	Shermis et al. 2013	A handbook of automatic essay evaluation
2015	Burrows et al. 2015	A comprehensive review of automatic short answer grading.
2015	Roy et al. 2015	A survey on computer-assisted assessment of short answers.
2016	Hasanah et al. 2016	A review of automatic short-answer grading.
2018	Alruwais et al. 2018	Advantages and challenges of e-assessment.

are short-answer type question and long-answer type question (Clay 2001). The answer to a short question requires a sentence or two to three sentences, and a long-type question needs more than three sentences or paragraphs. However, both subjective and objective questions are necessary for good classroom test (Heaton 1990). Figure 1 shows the overall diagram of different question generation and answer evaluation methods for automatic assessment system. We initially categorized the online learning techniques into three different types, namely text-based, audio and video-based, and image-based. We emphasized mainly text-based approaches and further extended the modality towards assessment methods. We discussed audio-video and image-based learning in this article, but the extensive analysis of such learning methods is out of the scope of this article.

The objective question becomes popular as an automated assessment tool in the examination system due to its fast and reliable evaluation policies (Nicol 2007). It involves the binary mode of assessment that has only one correct answer. On the other side, the subjective examination has obtained the attention of the evaluators to evaluate a candidate's deep knowledge and understanding of the traditional education system for centuries (Shaban 2014). Individually, each university has followed different patterns of subjective examination. Due to the rapid growth of e-learning courses, we need to consider such assessments and evaluations done by the automated appraisal system. The computer-based assessment of subjective questions is challenging, and the accuracy of it has not achieved adequate results. Hopefully, the research on automatic evaluation of subjective-questions in examination discovers new tools to help schools and teachers. An automated tool can able to resolve the problem of hand-scoring thousands of written answers in the subjective-examination. Today's computer-assisted examination excludes the subjective-questions by MCQs, which are not able to assess the writing skills and critical reasoning of the students due to its unreliable accuracy of evaluation. Table 3 shows the different types of questions and compares the level of difficulties to generate questions and evaluate the learner's answers.

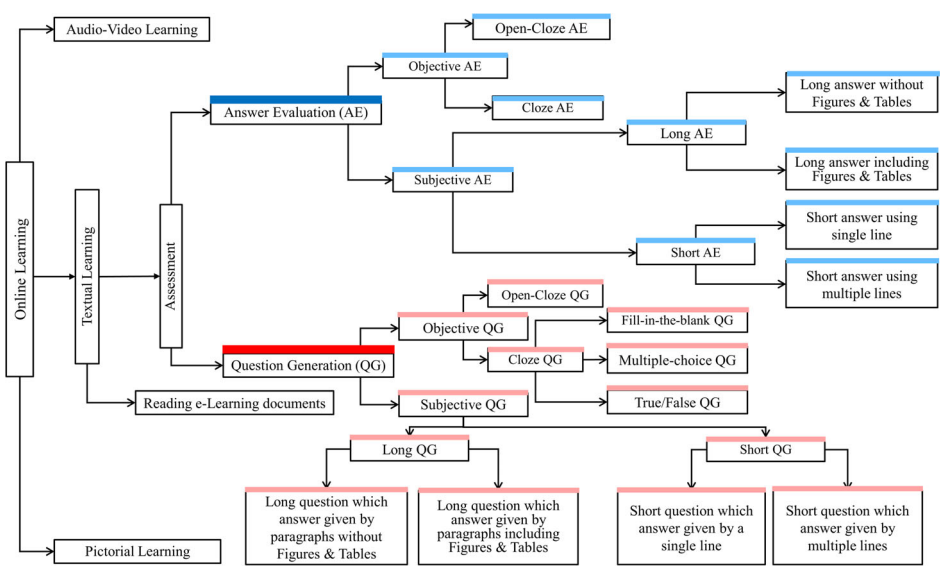

Fig. 1 Different modalities of question generation and assessment methods reported in literature

Table 3 Types and difficulty level of automatic question generation and learner's answer evaluation

Question Types	Source	Question Generation	Learner's Assessment
		Easy	**Easy**
Open-Cloze Question	Pino and Eskenazi (2009), Agarwal (2012), Das and Majumder (2017)	(Omit a word or phrase from an informative sentence)	(Match a word or phrase)
	Coniam (1997); Brown et al. (2005); Chen et al. (2006),	**Moderate**	**Very Easy**
Cloze Question	Hoshino and Nakagawa (2007); Pino et al. (2008), Agarwal and Mannem (2011); Correia et al. (2012), Narendra et al. (2013)	(Main difficulty arise in distractors generation)	(Check option is true or false)
		Moderate	**Very Easy**
Multiple-Choice Question	Mitkov et al. (2006); Aldabe and Maritxalar (2010), Papasalouros et al. (2008); Bhatia et al. (2013), Majumder and Saha (2014) and Majumder and Saha (2015)	(Difficulty arise in distractors generation)	(Check option is true or false)
		Difficult	**Difficult**
Subjective Question	Rozali et al. (2010); Dhokrat et al. (2012); Deena et al. (2020), Leacock and Chodorow (2003); Bin et al. (2008); Kakkonen et al. (2008), Noorbehbahani and Kardan (2011); Dhokrat et al. (2012), Islam and Hoque (2010); Ramachandran et al. (2015), Sakaguchi et al. (2015)	(Difficult in question formation, generally used predefined template or question pattern)	(Match learner's answer with handcraft model answer)
		Very Difficult	**Very Difficult**
Visual Question	Simoncelli and Olshausen (2001); Mora et al. (2016), Mostafazadeh et al. (2016); Zhu et al. (2016); Yu et al. (2015), Ren et al. (2015); Zhang et al. (2017); Johnson et al. (2016); Jain et al. (2017)	(Image tagging and question formation)	(Generate answer from image-based question)

ACL, IEEE, Mendeley, Google Scholar, and Semantic Scholar are searched to collect high-quality journals and conferences for this survey. The search has involved a combination and variation of the keywords such as automatic question generation, multiple-choice questions generation, cloze questions generation, fill-in-the-blank questions generation, visual question generation, subjective answer evaluation, short answer evaluation, and short answer grading. A total of 78 articles are included in this study. Figure 2 shows the statistics of articles for different question generation and learners' answer evaluation that found in the last 10 years in the literature.

Related datasets

In 2010, a question generation system QGSTEC used a dataset that contains overall 1000 questions (generated by both humans and machines). The system generated a few questions for each question type (which, what, who, when, where, and how many). Five fixed criteria were used to measure the correctness of the generated questions—relevance, question type, grammatically correct, and ambiguity. Both the relevancy and the syntactic correctness measures did not score well. The agreement between the two human judges was quite low.

The datasets SQuAD, 30MQA, MS MARCO, RACE, NewsQA, TriviaQA, and NarrativeQA contain question-answer pairs and are mainly developed for machine-reading comprehension or question answering models. These datasets are not designed for direct question-generation from textual documents. The datasets are also not suited for educational assessment due to their limited number of topics or insufficient information for generating questions and further answer the questions.

TabMCQ dataset contains large scale crowdsourced MCQs covering the facts in the tables. This dataset is designed for not only the task of question answering but also information extraction, question parsing, answer-type identification, and lexical-semantic modeling. The facts of the tables are not adequate to generate MCQs. The SciQ dataset also consists of a large set of crowdsourced MCQs with distractors and an additional passage that provides the clue for the correct answer. This passage does not contain sufficient information to generate MCQs or distractors. Therefore, both the TabMCQ and SciQ datasets are not applicable for multiple-choice question generation as well as distractors generation.

MCQL dataset is designed for automatic distractors generation. Each MCQ associates with four fields: sentence, answer, distractors, and the number of distractors. We observed that the sentence is not sufficient for generating MCQs for all times. The dataset does

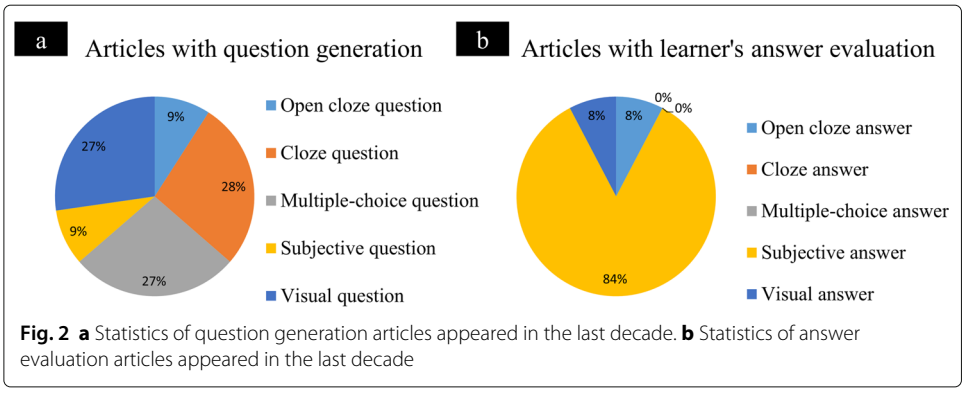

Fig. 2 **a** Statistics of question generation articles appeared in the last decade. **b** Statistics of answer evaluation articles appeared in the last decade

not include the source text from where it collects the MCQs and distractors. Distractors not only depend on the question, sentence, and correct answer but also the source text. Therefore, the MCQL dataset is not applicable when it needs to generate questions, answers, and distractors from the same source text or study materials.

LearningQ dataset covers a wide range of learning subjects as well as the different levels of cognitive complexity and contains a large-set of document-question pairs and multiple source sentences for question generation. The dataset decreases the performance of question generation when the length of source sentences increases. Therefore, the dataset is helpful to forward the research on automatic question generation in education.

Table 4 presents the existing datasets which contain question-answer pairs and related to question-answer generation. Table 5 includes the detail description of each dataset.

Objective Question Generation

The study of literature review shows that most of the researchers paid attention to generate objective-type questions, automatically or semi-automatically. They confined their works to generate multiple-choice or cloze questions. A limited number of approaches are found in the literature that shows interest in open-cloze question generation.

Pino and Eskenazi (2009) provided the hint in an open-cloze question. They noted the first few letters of a missing word gave a clue about the missing word. Their goal was to vary the number of letters in hint to change the difficulty level of questions that facilitate the students to learn vocabulary. Agarwal (2012) developed an automated open-cloze question generation method. Their approach composed of two steps—selected relevant and informative sentences and identified keywords from the selected sentences. His proposed system had taken cricket-news articles as input and generated factual open-cloze questions as output. Das and Majumder (2017) described a system for open-cloze question generation to evaluate the factual knowledge of learners. They computed the

Table 4 Related existing datasets

Source	Dataset	Purpose
Rus et al., 2012	QGSTEC	Automatic Question Generation
Jauhar et al., 2015	TabMCQ	Question Answering, Information Extraction, Question Parsing, Answer-type Identification, and Lexical Semantic Modeling
Rajpurkar et al., 2016	SQuAD	Reading Comprehension: Answer a question posed by humans from a corresponding passage
Serban et al., 2016	30MQA	Question Answering: Generate Question Answer Pairs from Knowledge Bases
Nguyen et al., 2016	MS MARCO	Machine Reading Comprehension and Question-Answering
Lai et al., 2017	RACE	Machine Comprehension and Question Answering: Evaluating the reading comprehension ability of students
Trischler et al., 2017	NewsQA	Machine Comprehension
Joshi et al., 2017	TriviaQA	Reading Comprehension, Question Answering over structured Knowledge Bases and joint modeling of Knowledge Bases and Text
Welbl et al., 2017	SciQ	Question Generation and Question Answering
Liang et al., 2018	MCQL	Automatic Distractor Generation
Kocisk'y et al., 2018	NarrativeQA	Reading Comprehension
Chen et al., 2018	LearningQ	Automatic Educational Question Generation

Table 5 Dataset description

Dataset	Description
QGSTEC	A corpus of over 1000 questions. The questions are generated from individual sentences or a paragraph.
TabMCQ	The dataset contains a large set of crowd-sourced MCQs covering the facts in the 65 hand-crafted tables.
SQuAD	The dataset consists of 100K+ samples collecting from Wikipedia articles. Each sample consists of question-answer pairs with a passage. The answer is a part of the text from the passage.
30MQA	The corpus consists of 30M question-answer pairs created by humans and their corresponding Freebase fact which represents by a triple. A triple consists of a subject, a relationship, and an object which is converted into a question with this subject and object where the object is the correct answer.
MS MARCO	The dataset covers 1,010,916 questions from the query log of Bing's search with human-generated answers.
RACE	The dataset consists of a large set of questions (nearly 100K), answers and associated passages generated by human experts.
NewsQA	A large-scale dataset contains over 100K human-generated question-answer pairs based on a set of over 10K news articles.
TriviaQA	The dataset contains over 650K question-answer-evidence documents triples. The documents are collected from web search and Wikipedia pages.
SciQ	The dataset consists of 13.7K crowdsourced multiple-choice science questions. Every MCQ has one correct answer with three distractors, and one additional passage to support the evidence of the correct answer. Most instances get from the passages used to generate the question.
MCQL	The dataset has crawled from the Web and contains 7.1K MCQs. Each MCQ associates with four fields - sentence, answer, distractors, and the number of distractors.
NarrativeQA	The dataset contains a large number of question-answer pairs from a smaller collection of large documents. The dataset has designed for answering the questions correctly that require much understanding of the underlying narrative rather than just pattern matching.
LearningQ	The dataset contains 230K+ document-question pairs created by instructors and learners.

evaluation score using a formula that depends on the number of hints used by the learners to give the right answers. The multiword answer to the open-cloze question makes the system more attractive.

Coniam (1997) proposed one of the oldest techniques of cloze test item generation. He applied word frequencies to analyze the corpus in various phases of development, such as obtain the keys for test items, generate test item alternatives, construct cloze test items, and identify good and bad test items. He matched word frequency and parts-of-speech of each test item key with a similar word class and word frequency to construct test items. Brown et al. (2005) revealed an automated system to generate vocabulary questions. They applied WordNet (Miller 1995) for obtaining the synonym, antonym, and hyponym to develop the question key and the distractors. Chen et al. (2006) developed a semi-automated method using NLP techniques to generate grammatical test items. Their approach implied handcraft patterns to find authentic sentences and distractors from the web that transform into grammar-based test items. Their experimental results showed that the method had generated 77% meaningful questions. Hoshino and Nakagawa (2007) introduced a semi-automated system to create cloze test items from online news articles to help teachers. Their test items removed one or more words from a passage, and learners were asked to fill those omitted words. Their system generated two types of distractors: grammatical distractors and vocabulary distractors. The human-based evaluation revealed that their system produced 80% worthy cloze test items. Pino

et al. (2008) employed four selection criteria: well-defined context, complexity, grammaticality, and length to give a weighted score for each sentence. They selected a sentence as informative if the score was higher than a threshold for generating a cloze question. Agarwal and Mannem (2011) presented a method to create gap-fill-questions from a biological-textbook. The authors adopted several features to generate the questions: sentence length, the sentence position in a document, is it the first sentence, is the sentence contains token that appears in the title, the number of nouns and pronouns in the sentence, is it holds abbreviation or superlatives. They did not report the optimum value of these features or any relative weight among features or how the features combined. Correia et al. (2012) applied supervised machine learning to select stem for cloze questions. They employed several features to run the classifier of SVM: the length of sentence, the position of the word in a sentence, the chunk of the sentence, verb, parts-of-speech, named-entity, known-word, unknown-word, acronym, etc. Narendra et al. (2013) directly employed a summarizer (MEAD)[1] to select the informative sentences for automatic CQs generation. Flanagan et al. (2013) described an automatic-method for generating multiple-choice and fill-in-the-blanks e-Learning quizzes.

Mitkov et al. (2006) proposed a semi-automated system for generating MCQs from a linguistic-textbook. They employed several NLP approaches for question generation—shallow parsing, key term extraction, semantic distance, sentence transformation, and ontology such as WordNet. Aldabe et al. 2010 presented a system to generate MCQ in the Basque language. They suggested different methods to find semantic similarities between the right answer and its distractors. A corpus-based strategy was applied to measure the similarities. Papasalouros et al. (2008) revealed a method to generate MCQs from domain ontologies. Their experiment used five different domain ontologies for multiple-choice question generation. Bhatia et al. (2013) developed a system for automatic MCQ generation from Wikipedia. They proposed a potential sentence selection approach using the pattern of existing questions on the web. They also suggested a technique for generating distractors using the named entity. Majumder and Saha (2014) applied named entity recognition and syntactic structure similarity to select sentences for MCQ generation. Majumder and Saha (2015) alternately used topic modeling and parse tree structure similarity to choose informative sentences for question formation. They picked the keywords using topic-word and named-entity and applied a gazetteer list-based approach to select distractors.

Subjective Question Generation and Evaluation

Limited research works found in the literature that focused on subjective question generation. Rozali et al. (2010) presented a survey of dynamic question generation and qualitative evaluation and a description of related methods found in the literature. Dhokrat et al. (2012) proposed an automatic system for subjective online examination using a taxonomy that coded earlier into the system. Deena et al. (2020) suggested a question generation method using NLP and bloom's taxonomy that generated subjective questions dynamically and reduced the occupation of memory.

Proper scoring is the main challenge of subjective assessment. Therefore, automatic subjective-answer evaluation is a current trend of research in the education system

[1] http://www.summarization.com/mead/

(Burrows et al. 2015). It reduces the assessment time and effort in the education system. Objective-type answer evaluation is easy and requires a binary mode of assessment (true/false) to test the correct option. But, the subjective answer evaluation does not achieve adequate results due to its complex nature. The next paragraph discusses some related works of subjective-answer evaluation and grading techniques.

Leacock and Chodorow (2003) proposed an answer grading system *C-rater* that deals with semantic information of the text. They adopted a method to recognize paraphrase to grad the answers. Their approach achieved 84% accuracy with the manual evaluation of human graders. Bin et al. (2008) employed the K-nearest neighbor (KNN) classifier for automated essay scoring using the text categorization model. The Vector Space Model was used to express each essay. They used words, phrases, and arguments as essay features and represented each vector using the TF-IDF weight. The cosine similarity was applied to calculate the score of essays and achieved 76% average accuracy using the different methods of feature selection, such as term frequency (TF), term frequency-inverse document frequency (TF-IDF), and information gain (IG). Kakkonen et al. (2008) recommended an automatic essay grading system that compares learning materials with the teacher graded essays using three methods: Latent Semantic Analysis (LSA), Probabilistic LSA (PLSA), and Latent Dirichlet Allocation (LDA). Their system performed better than the k-NN based grading system. Noorbehbahani and Kardan (2011) introduced a method for judging free text answers of students using a modified Bilingual Evaluation Understudy (M-BLEU) algorithm. The M-BLEU recognized the most similar reference answer to a student answer and estimated a score to judge the answers. Their method achieved higher accuracy than the other evaluation methods, like latent semantic analysis and n-gram co-occurrence. Dhokrat et al. (2012) proposed an appraisal system for evaluating the student's answer. The system used a centralized file that includes the model answer with the reference material for each question. Their system found overall 70% accuracy. Islam and Hoque (2010) presented an automatic essay grading system using the generalized latent semantic analysis (GLSA). The GLSA based system used word-ordering in the sentences by including the word n-gram for grading essays. The GLSA based system performs better than the LSA-based grading system and overcomes the limitations of the LSA based system, where the LSA does not consider word-order of sentences in a document. Ramachandran et al. (2015) described a unique technique for scoring short answers. They introduced word ordering graphs to recognize the useful patterns from handcraft rubric texts and the best responses of students. The method also employed semantic metrics to manage related-words for alternative answer options. Sakaguchi et al. (2015) used different sources of information for scoring content-based short answers. Their approach extracted features from the responses (word and character n-grams). Their reference-based method found the similarity between the response features with the information from the scoring guidelines. Their model outperformed when the training data is limited.

Recent progress in deep learning-based NLP has also shown a promising future in answer assessment. Sentiment-based assessment techniques Nassif et al. 2020; Abdi et al2019 used in many cases because of the generalized representation of sentiment in NLP. The success of recurrent neural networks (RNN) such as Long short-term memory (LSTM) becomes popular in sequence analysis and applied in various answer assessment (Du et al. 2017; Klein and Nabi 2019).

Visual Question-Answer Generation

Recently, question generation has been included in the field of computer vision to generate image-based questions (Gordon et al. 2018; Suhr et al. 2019; Santoro et al. 2018). The most recent approaches use human-annotated question-answer pairs to train machine learning algorithms for generating multiple questions per image, which were labor-intensive and time-consuming (Antol et al. 2015; Gao et al. 2015). One of the recent examples, Zhu et al. 2016 manually created seven wh-type questions such as when, where, and what. People also investigated automatic visual question generation by using rules. Yu et al. (2015) proposed the question generation as a task of removing a content word (answer) from an image caption and reforms the caption sentence as a question. Similarly, Ren et al. 2015 suggested a rule to reformulate image captions into limited types of questions. Some considered model-based methods to overwhelm the diversity issue of question types. Simoncelli and Olshausen (2001) trained a model using a dataset of image captions and respective visual questions. But, their model could not generate multiple questions per image. Mora et al. (2016) proposed an AI model to generate image-based questions with respective answers simultaneously. Mostafazadeh et al. (2016) collected the first visual question generation dataset, where their model generated several questions per image. Zhang et al. (2017) proposed an automatic model for generating several visually grounded questions from a single image. Johnson et al. (2016) suggested a framework named *Densecap* for generating region captions, which are the additional information to supervise the question generation. Jain et al. (2017) combined the variational auto-encoders and LSTM networks to generate numerous types of questions from a given image. The majority of these image-based question-answers were related to image understanding and reasoning in real-world images.

Visual Question-Answer Dataset

Figure 3a shows a few examples where various pattern identification and reasoning tests used synthetic images. Johnson et al. (2017) proposed a diagnostic dataset CLEVR,

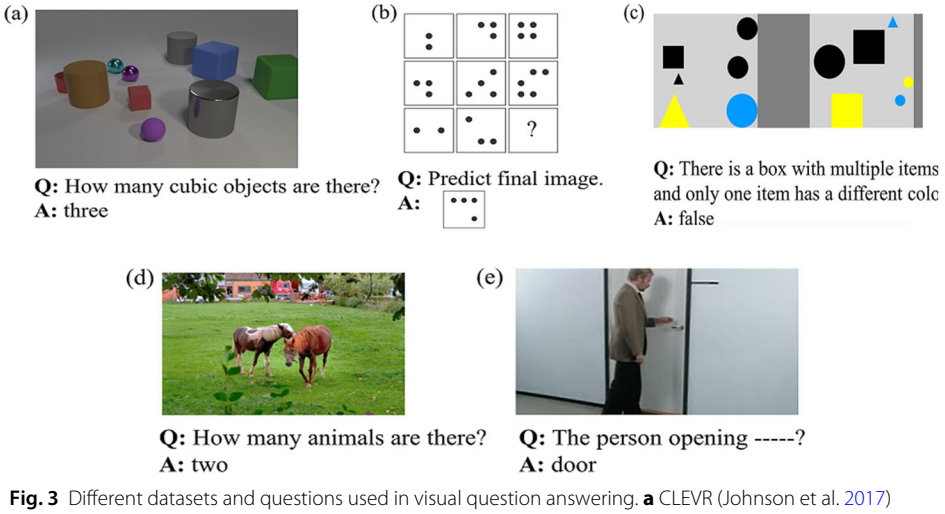

Fig. 3 Different datasets and questions used in visual question answering. **a** CLEVR (Johnson et al. 2017) dataset, **b** abstract reasoning dataset (Santoro et al. 2018), **c** NLVR (Suhr et al. 2019) dataset, **d** VQA (Antol et al. 2015) dataset, and **e** IQA (Gordon et al. 2018) dataset

which has a collection of 3D shapes and used to test the skill of visual reasoning. The dataset is used for question-answering about shapes, positions, and colors. Figure 3b presents Raven progressive matrices based visual-reasoning that is used to test shape, count, and relational visual reasoning from an image sequence (Bilker et al. 2012). Figure 3c is an example of NLVR dataset. The dataset used the concepts of 2D shapes and color to test visual reasoning. The dataset is used to generate questions related to the knowledge of shape, size, and color. Figure 3d is an example of visual question answering dataset (VQA). The dataset consists of a large volume of real-world images and is used to generate questions and corresponding answers related to objects, color, and counting. Figure 3e is also a similar dataset related to event and actions. All these datasets are used to generate image-specific questions and also used in various assessments.

Challenges in Question Generation and Answer Assessment

Informative-simple-sentence extraction

Questions mainly depend on informative sentences. An informative-sentence generates a quality question to assess learners. We found that text-summarization, sentence-simplification, and some rule-based techniques in the literature exacted the informative-sentences from an input text. Most of the previous articles did not focus adequately on the step of informative-sentence selection. But it is a useful-step for generating quality questions. Generate simple-sentences from complex and compound sentences are also complex. A simple-sentence eliminates the ambiguity between multiple answers to a particular question. Therefore, a generic technique is needed to extract the informative-simple-sentences from the text for generating questions (Das et al. 2019). The popular NLP packages like NLTK, spaCy, PyNLPl, and CoreNLP did not include any technique for extracting informative-sentences from a textual document. It is a future direction of research to incorporate it into the NLP packages.

Question generation from multiple sentences

Different question generation techniques generate different questions that assess the knowledge of learners in different ways. An automated system generates questions from study material or learning content based on informative keywords or sentences and multiple sentences or a passage. Generate questions from multiple sentences or a paragraph is difficult and consider a new research direction for automatic question generation. It requires the inner relation between sentences using natural language understanding concepts.

Short and long-type answer assessment

We found many works in the last decade for automatic grading short answers or free-text answers. But the unreliable results of previous research indicates that it is not practically useful in real life. Therefore, most of the exams conduct using MCQs and exclude the short type and long type answers. We found only one research that evaluates long-answers in the literature. Therefore, future research expects a reliable and real-life system for short answer grading as well as long type answer evaluation that fully automate the education system.

Answer assessment standard

Question generation and assessment depend on many factors such as learning domain, type of questions for assessments, difficulty level, question optimization, scoring techniques, and overall scoring. Several authors proposed different evaluation techniques depend on their application, and the scoring scale is also different. Therefore, an answer assessment standard is required in the future to evaluate and compare the learner's knowledge and compare the research results.

Question generation and assessment from video lectures

We found that the majority of question generation and assessment systems focus on generating questions from the textual document to automate the education system. We found a limited number of works in the literature that generate questions from the visual content for the learner's assessment. Assessment from video lectures by generating questions from video content is a future research direction. Audio-video content improves the learning process (Carmichael et al. 2018). Automated assessments from video content can help learners to learn quickly in a new area.

Question generation and assessment using machine learning

Due to the many advantages of the machine learning method, recent works focus on it to generate questions and evaluate answers. Most of the textual question generation used natural language processing (NLP) techniques. The advancement of NLP is natural language understanding (NLU) and natural language generation (NLG) that used a deep learning neural network (Du et al. 2017; Klein and Nabi 2019). The visual question generation method mainly used machine learning to generate image captions. Image caption translates into a question using NLP techniques. VQG is a combined application of computer vision and NLP. In some articles used sequence-to-sequence modeling for generating questions. Limited works found in the literature that assess the learners using a machine learning approach. More research works need to focus on this area in the future.

Conclusion

Due to the advances in online learning, automatic question generation and assessment are becoming popular in the intelligent education system. The article first includes a collection of review articles in the last decade. Next, it discusses the state-of-the-art methods of various automatic question generation as well as different assessment techniques that summarizes the progress of research. It also presents a summary of related existing datasets found in the literature. This article critically analyzed the methods of objective question generation, subjective question generation with the learner's response evaluation, and a summarizing of visual question generation methods.

Abbreviations

AI: Artificial intelligence; AQG: Automatic question generation; BLEU: Bilingual evaluation understudy; GLSA: Generalized latent semantic analysis; KNN: K-nearest neighbor; LDA: Latent Dirichlet allocation; LSA: Latent semantic analysis; LSTM: Long short term memory; MCQ: Multiple choice question; NLP: Natural language processing; PLSA: Probabilistic latent semantic analysis; TF-IDF: Term frequency-inverse document frequency; VQA: Visual question answering; VQG: Visual question generation;

Acknowledgements

This research was supported/partially supported by Indian Center for Advancement of Research and Education (ICARE), Haldia

Authors' contributions
All authors equally contributed and approved the final manuscript.

Funding
This study is not funded from anywhere.

Availability of data and materials
Not applicable

Informed consent
Informed consent was obtained from all individual participants included in the study.

Declarations

Ethics approval and consent to participate
This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of interest
The authors declare that there is no conflict of interest regarding the publication of this paper.

Author details
[1]Department of Information Technology, Haldia Institute of Technology, Haldia, India. [2]Department of Computer Science and Application, University of North Bengal, Darjeeling, India. [3]Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, West Bengal, India. [4]Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø, Norway.

Received: 10 July 2020 Accepted: 24 February 2021
Published online: 18 March 2021

References

Abdi, A., Shamsuddin, S.M., Hasan, S., Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing & Management*, *56*(4), 1245–1259.
Agarwal, M. (2012). Cloze and open cloze question generation systems and their evaluation guidelines. Master's thesis. *International Institute of Information Technology, (IIIT),* Hyderabad, India.
Agarwal, M., & Mannem, P. (2011). Automatic gap-fill question generation from text books, In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 56–64). Portland: Association for Computational Linguistics.
Aldabe, I., & Maritxalar, M. (2010). Automatic distractor generation for domain specific texts, In *Proceedings of the 7th International Conference on Advances in Natural Language Processing* (pp. 27–38). Berlin: Springer-Verlag.
Alruwais, N., Wills, G., Wald, M. (2018). Advantages and challenges of using e-assessment. *International Journal of Information and Education Technology*, *8*(1), 34–37.
Amidei, J., Piwek, P., Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018, In *Proceedings of The 11th International Natural Language Generation Conference* (pp. 307–317). Tilburg University: Association for Computational Linguistics.
Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D. (2015). Vqa: Visual question answering, In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2425–2433).
Bhatia, A.S., Kirti, M., Saha, S.K. (2013). Automatic generation of multiple choice questions using wikipedia, In *Proceedings of the Pattern Recognition and Machine Intelligence* (pp. 733–738). Berlin: Springer-Verlag.
Bilker, W.B., Hansen, J.A., Brensinger, C.M., Richard, J., Gur, R.E., Gur, R.C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354–369.
Bin, L., Jun, L., Jian-Min, Y., Qiao-Ming, Z. (2008). Automated essay scoring using the KNN algorithm, In *2008 International Conference on Computer Science and Software Engineering*, 1 (pp. 735–738). Washington, DC: IEEE.
Boyd, R.T. (1988). Improving your test-taking skills. *Practical Assessment, Research & Evaluation*, *1*(2), 3.
Brown, J.C., Frishkoff, G.A., Eskenazi, M. (2005). Automatic question generation for vocabulary assessment, In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 819–826). Vancouver: Association for Computational Linguistics.
Burrows, S., Gurevych, I., Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117.
Carmichael, M., Reid, A., Karpicke, J.D. (2018). *Assessing the impact of educational video on student engagement, critical thinking and learning: The Current State of Play*, (pp. 1–21): A SAGE Whitepaper, Sage Publishing.
Ch, D.R., & Saha, S.K. (2018). Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, *13*(1), 14–25. https://doi.org/10.1109/TLT.2018.2889100.
Chen, C.-Y., Liou, H.-C., Chang, J.S. (2006). Fast–an automatic generation system for grammar tests, In *Proceedings of the COLING/ACL on Interactive Presentation Sessions* (pp. 1–4). Sydney: Association for Computational Linguistics.
Chen, G., Yang, J., Hauff, C., Houben, G.-J. (2018). Learningq: A large-scale dataset for educational question generation, In *Twelfth International AAAI Conference on Web and Social Media* (pp. 481–490).
Clay, B. (2001). A short guide to writing effective test questions. *Lawrence: Kansas Curriculum Center, University of Kansas*. https://www.k-state.edu/ksde/alp/resources/Handout-Module6.pdf.
Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *Calico Journal*, *14*(2-4), 15–33.

Correia, R., Baptista, J., Eskenazi, M., Mamede, N. (2012). Automatic generation of cloze question stems, In *Computational Processing of the Portuguese Language* (pp. 168–178). Berlin: Springer-Verlag.
Das, B., & Majumder, M. (2017). Factual open cloze question generation for assessment of learner's knowledge. *International Journal of Educational Technology in Higher Education*, *14*(1), 1–12.
Das, B., Majumder, M., Phadikar, S., Sekh, A.A. (2019). Automatic generation of fill-in-the-blank question with corpus-based distractors for e-assessment to enhance learning. *Computer Applications in Engineering Education*, *27*(6), 1485–1495.
Deena, G., Raja, K., PK, N.B., Kannan, K. (2020). Developing the assessment questions automatically to determine the cognitive level of the E-learner using NLP techniques. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, *11*(2), 95–110.
Dhokrat, A., Gite, H., Mahender, C.N. (2012). Assessment of answers: Online subjective examination, In *Proceedings of the Workshop on Question Answering for Complex Domains* (pp. 47–56).
Divate, M., & Salgaonkar, A. (2017). Automatic question generation approaches and evaluation techniques. *Current Science*, *113*(9), 1683–1691.
Du, X., Shao, J., Cardie, C. (2017). Learning to Ask: Neural Question Generation for Reading Comprehension, In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1342–1352). Vancouver: Association for Computational Linguistics.
Flanagan, B., Yin, C., Hirokawa, S., Hashimoto, K., Tabata, Y. (2013). An automated method to generate e-learning quizzes from online language learner writing. *International Journal of Distance Education Technologies (IJDET)*, *11*(4), 63–80.
Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W. (2015). Are you talking to a machine? Dataset and methods for multilingual image question, In *Advances in Neural Information Processing Systems* (pp. 2296–2304).
Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., Farhadi, A. (2018). Iqa: Visual question answering in interactive environments, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4089–4098).
Hasanah, U., Permanasari, A.E., Kusumawardani, S.S., Pribadi, F.S. (2016). A review of an information extraction technique approach for automatic short answer grading, In *2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 192–196). Yogyakarta: IEEE.
Heaton, J.B. (1990). *Classroom testing*.
Hoshino, A., & Nakagawa, H. (2007). Assisting cloze test making with a web application, In *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), Waynesville, NC USA (pp. 2807–2814).
Islam, M.M., & Hoque, A.L. (2010). Automated essay scoring using generalized latent semantic analysis, In *2010 13th International Conference on Computer and Information Technology (ICCIT)* (pp. 358–363). Dhaka: IEEE.
Jain, U., Zhang, Z., Schwing, A.G. (2017). Creativity: Generating diverse questions using variational autoencoders, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6485–6494).
Jauhar, S.K., Turney, P., Hovy, E. (2015). TabMCQ: A Dataset of General Knowledge Tables and Multiple-choice Questions. https://www.microsoft.com/en-us/research/publication/tabmcq-a-dataset-of-general-knowledge-tables-and-multiple-choice-questions/.
Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2901–2910).
Johnson, J., Karpathy, A., Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4565–4574).
Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L. (2017). TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1601–1611). Vancouver: Association for Computational Linguistics.
Kakkonen, T., Myller, N., Sutinen, E., Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Journal of Educational Technology & Society*, *11*(3), 275–288.
Klein, T., & Nabi, M. (2019). Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds. *ArXiv*, *abs/1911.02365*.
Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, *6*, 317–328.
Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204.
Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations, In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 785–794). Copenhagen: Association for Computational Linguistics.
Le, N.-T., Kojiri, T., Pinkwart, N. (2014). Automatic question generation for educational applications–the state of art, In *Advanced Computational Methods for Knowledge Engineering* (pp. 325–338).
Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405.
Liang, C., Yang, X., Dave, N., Wham, D., Pursel, B., Giles, C.L. (2018). Distractor generation for multiple choice questions using learning to rank, In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 284–290).
Majumder, M., & Saha, S.K. (2014). Automatic selection of informative sentences: The sentences that can generate multiple choice questions. *Knowledge Management and E-Learning: An International Journal*, *6*(4), 377–391.
Majumder, M., & Saha, S.K. (2015). A system for generating multiple choice questions: With a novel approach for sentence selection, In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 64–72). Beijing: Association for Computational Linguistics.
Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.
Mitkov, R., LE An, H., Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, *12*(2), 177–194.
Mora, I.M., de la Puente, S.P., Nieto, X.G. (2016). Towards automatic generation of question answer pairs from images, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1–2).

Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L. (2016). Generating Natural Questions About an Image, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, (Volume 1: Long Papers), Berlin, Germany* (pp. 1802–1813).
Narendra, A., Agarwal, M., Shah, R. (2013). Automatic cloze-questions generation, In *Proceedings of Recent Advances in Natural Language Processing* (pp. 511–515). Hissar: INCOMA Ltd. Shoumen, BULGARIA (ACL 2013).
Nassif, A.B., Elnagar, A., Shahin, I., Henno, S. (2020). Deep learning for arabic subjective sentiment analysis: Challenges and research opportunities. *Applied Soft Computing*, 106836.
Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., Wang, T. (2016). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv*, arXiv:1611.09268. https://ui.adsabs.harvard.edu/abs/2016arXiv161109268B.
Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and higher Education*, *31*(1), 53–64.
Noorbehbahani, F., & Kardan, A.A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education*, *56*(2), 337–345.
Papasalouros, A., Kanaris, K., Kotis, K. (2008). Automatic generation of multiple choice questions from domain ontologies, In *Proceedings of the e-Learning* (pp. 427–434).
Pino, J., & Eskenazi, M. (2009). Measuring hint level in open cloze questions, In *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference(FLAIRS)* (pp. 460–465). Florida: The AAAI Press.
Pino, J., Heilman, M., Eskenazi, M. (2008). A selection strategy to improve cloze question quality, In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains, 9th International Conference on Intelligent Tutoring Systems* (pp. 22–34). Montreal: Springer.
Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P. (2016). 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). Austin: Association for Computational Linguistics.
Ramachandran, L., Cheng, J., Foltz, P. (2015). Identifying patterns for short answer scoring using graph-based lexico-semantic text matching, In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 97–106).
Ren, M., Kiros, R., Zemel, R. (2015). Exploring models and data for image question answering, In *Advances in Neural Information Processing Systems* (pp. 2953–2961).
Roy, S., Narahari, Y., Deshmukh, O.D. (2015). A perspective on computer assisted assessment techniques for short free-text answers, In *International Computer Assisted Assessment Conference* (pp. 96–109). Zeist: Springer.
Rozali, D.S., Hassan, M.F., Zamin, N. (2010). A survey on adaptive qualitative assessment and dynamic questions generation approaches, In *2010 International Symposium on Information Technology*, 3 (pp. 1479–1484). Kuala Lumpur: IEEE.
Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., Moldovan, C. (2012). A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse*, *3*(2), 177–204.
Sakaguchi, K., Heilman, M., Madnani, N. (2015). Effective feature integration for automated short answer scoring, In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1049–1054).
Santoro, A., Hill, F., Barrett, D., Morcos, A., Lillicrap, T. (2018). Measuring abstract reasoning in neural networks, In *International Conference on Machine Learning* (pp. 4477–4486).
Serban, I.V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., Bengio, Y. (2016). Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus, In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 588–598). Berlin: Association for Computational Linguistics.
Shaban, A.-M.S. (2014). A comparison between objective and subjective tests. *Journal of the College of Languages*, *30*, 44–52.
Shermis, M.D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*.
Simoncelli, E.P., & Olshausen, B.A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.
Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y. (2019). A Corpus for Reasoning about Natural Language Grounded in Photographs, In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6418–6428). Florence: Association for Computational Linguistics.
Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., Suleman, K. (2017). NewsQA: A Machine Comprehension Dataset, In *Proceedings of the 2nd Workshop on Representation Learning for NLP* (pp. 191–200). Vancouver: Association for Computational Linguistics.
Welbl, J., Liu, N.F., Gardner, M. (2017). Crowdsourcing multiple choice science questions, In *Proceedings of the 3rd Workshop on Noisy User-generated Text* (pp. 94–106).
Yu, L., Park, E., Berg, A.C., Berg, T.L. (2015). Visual madlibs: Fill in the blank description generation and question answering, In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2461–2469).
Zhang, S., Qu, L., You, S., Yang, Z., Zhang, J. (2017). Automatic generation of grounded visual questions, In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 4235–4243). Melbourne: The AAAI Press.
Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L. (2016). Visual7w: Grounded question answering in images, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4995–5004).

Publisher's Note