A Survey of Approaches to Automatic Question Generation: from 2019 to Early 2021

Chao-Yi Lu
Chingshin Academy
Taipei, Taiwan
chaoyilu.zoey@gmail.com

Sin-En Lu
Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
alznn@g.ncu.edu.tw

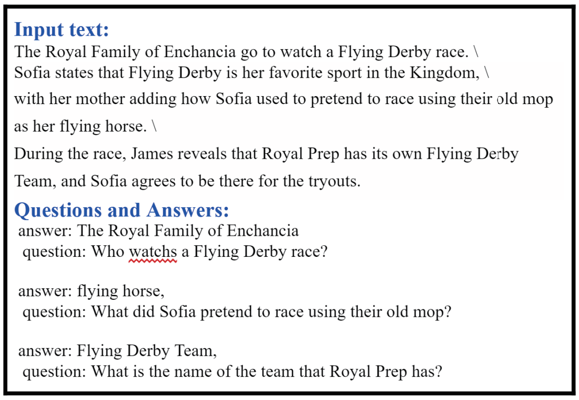
Abstract

To provide analysis of recent researches of automatic question generation from text, we surveyed 15 papers between 2019 to early 2021, retrieved from Paper with Code (PwC). Our research follows the survey reported by Kurdi et al. (2020), in which analysis of 93 papers from 2014 to early 2019 are provided. We analyzed the 15 papers from aspects including: (1) purpose of question generation, (2) generation method, and (3) evaluation. We found that recent approaches tend to rely on semantic information and Transformer-based models are attracting increasing interest since they are more efficient. On the other hand, since there isn't any widely acknowledged automatic evaluation metric designed for question generation, researchers adopt metrics of other natural language processing tasks to compare different systems.

Keywords: Automatic question generation, Survey, Natural language processing

1 Introduction

Questions are crucial tools for assessments and providing assistance throughout the process of learning. The functions of well-designed questions include: (1) providing opportunities to practice retrieving information from memory, (2) giving learners feedback about their misconceptions, (3) focusing learners' attention on the most important material, and (4) reinforcing what learners have acquired through repeating core concepts (Thalheimer, 2003). With the rapid growth of online learning, the demand for questions has increased. However, creating questions by humans is not efficient since the process requires training and cannot produce results immediately.



Input text:
The Royal Family of Enchancia go to watch a Flying Derby race. \
Sofia states that Flying Derby is her favorite sport in the Kingdom, \
with her mother adding how Sofia used to pretend to race using their old mop
as her flying horse. \
During the race, James reveals that Royal Prep has its own Flying Derby
Team, and Sofia agrees to be there for the tryouts.

Questions and Answers:
answer: The Royal Family of Enchancia
question: Who watchs a Flying Derby race?

answer: flying horse,
question: What did Sofia pretend to race using their old mop?

answer: Flying Derby Team,
question: What is the name of the team that Royal Prep has?

Figure 1: An example of AQG using the model by Lopez et al. (2020). Questions and answers are provided originally as generated. Text source: Sofia the First Wiki[1]

Question generation refers to the task of generating questions from various inputs.(Rus et al., 2008). Compared with humans, automatic question generation (AQG) can produce questions in lower cost and higher efficiency. Despite the development of visual question generation (generating questions from images) is undoubtedly essential since it **combines natural language processing and computer vision** (Sarrouti et al., 2020), the focus of this survey is on AQG from texts due to its extensive usage including assessments (Stanescu et al., 2008; Ai et al., 2015), learning activities, and serving as a data augmentation approach for training Question Answering (QA) systems (Lee et al., 2020; Fabbri et al., 2020). For an example of question generation from text, please refer to Figure 1.

Hoping to compare existing AQG systems in our future works, we search the literature re-

[1] `https://sofia.fandom.com/wiki/Princess_Sofia`

viewed in this paper from Papers with Code[2] (PwC). As a survey paper, our project is concerned with reading and analyzing previous literature on AQG from text. We refer to the survey reported by Kurdi et al. (2020), which contains analysis of 93 papers from 2015 to early 2019, focusing on education. The objectives of Kurdi et al. (2020)'s review are (1) providing an overview of the AQG community and its activities, (2) summarising current QG approaches, (3) identifying the gold-standard performance in AQG, (4) Tracking the evolution of AQG since the review by Alsubait et al. (2016), which includes 81 papers published up to the end of 2014. We focus on the second objective and the evaluation of AQG systems, on the other hand, we discuss RNN-based and Transformer-based methods, both of which are classified as "statistical methods" during the procedure of transforming declarative sentences into inquisitive ones in the review proposed by Kurdi et al. (2020). Since we aim to continue their work and track the evolution of the AQG task, the papers investigated in this review range from 2019 to early 2021.

2 Background

2.1 Summary of Kurdi et al.s' Review

The work of Kurdi et al. (2020) groups the 93 papers they included together if they have at least one shared author and use the same type of AQG approach. There are a total of 72 groups, and evaluations have been made based on these groups. Not only did they provide information on AQG studies about (1) rate of publication, (2) types of papers and publication venues, and (3) research groups, but they also analyzed AQG studies based on multiple dimensions. The most crucial ones are presented in Table 1.

The results of Kurdi et al. (2020)'s evaluation on different dimensions are summarized in Table 1. Regarding "Domain", "Question format", and "Response format", the statistics are similar to the ones purposed by Alsubait et al. (2016), which implies that these aspects of AQG haven't changed much throughout the past decades. Generating domain-specific questions are more common than generating generic ones, and language learning received the most attention; wh- questions and gap-fill questions remain the most popular; multiple choice and free response are two of the most prevalent response formats. As for the development of the AQG field, Kurdi et al. (2020) found an rising tendency of publications per year and research groups, which indicates that AQG is attracting increasing interest and the community is expanding.

2.2 Data Sources

We search PwC for papers from different conferences on the question generation task and only keep papers published from 2019 to early 2021. The search queries used and results are provided in Table 2.

Using the data collecting method mentioned in the previous paragraph, we select 15 papers from conferences and journals including ACL, ICLR, EMNLP, and IJCNLP. Among the 15 papers, 3 were published in 2019 (Alberti et al., 2019; Zhang and Bansal, 2019; Cho et al., 2019a), 8 were published in 2020 (Lee et al., 2020; Pan et al., 2020b; Dhole and Manning, 2020; Fabbri et al., 2020; Chen et al., 2019; Wang et al., 2020; Qi et al., 2020; Su et al., 2020), and 4 in 2021 (Majumder et al., 2021; Pan et al., 2020a; Roemmele et al., 2021; Cho et al., 2019b).

3 Dimensions of AQG

The phrase "dimension" in our paper refers to different aspects of an AQG system. We will be providing analysis regarding (1) purpose of question generation, which is the usage of the systems purposed in review literature, (2) generation method, which stands for the approaches of understanding the input and transforming declarative sentences into inquisitive ones, and (3) evaluation, which includes the metrics and datasets the researchers used.

[2] https://paperswithcode.com/

[3] Categories that occurred three times or less are classified as "Others".

[4] Studies that do not specify any targeted field is classified as "Generic"

[5] Gap-fill questions or distract or generation are considered not having a transformation method since they only remove or select a word or phrase of the input.

[6] In their review, verbalization is defined as "**Any process carried out to improve the surface structure of questions (grammaticality and fluency) or to provide variations of questions (i.e. paraphrasing).**"

Dimension	Categories	studies	Percentage
Purpose	Assessment	40	55.56%
	Education(unspecified)	10	13.89%
	Support Learning	10	13.89%
	Self learning, self-study or self-assessment	9	12.50%
	Generate practice questions	8	11.11%
	Tutoring	5	6.94%
	Others[3]	5	6.94%
Input	Text	43	59.72%
	QuestionStem/QuestionKey	10	13.89%
	Ontology	8	8.33%
	RDFKB	5	6.94%
	Others[3]	10	13.89%
Domain	Generic[4]	33	45.83%
	Language	21	29.17%
	Math	4	5.56%
	Others[3]	13	18.06%
Generation method-Level of understanding	Semantic	60	83.33%
	Syntactic only	10	13.89%
Generation method-Procedure of transformation	Template	27	37.50%
	Rule	16	22.22%
	Statistical methods	9	12.50%
	Not having one[5]	20	27.78%
Question Format	wh-questions	22	30.56%
	Gap-fill Questions	20	27.78%
	Word Problem	4	5.56%
	Others[3]	37	51.39%
Response Format	Multiple Choice	38	52.78%
	Free Response	36	50.00%
	True/false	2	2.78%
	Sound	1	1.39%
Difficulty Controlling	Yes	14	19.44%
	No	58	80.56%
Feedback Generation	Yes	1	1.39%
	No	71	98.61%
Verbalization[6]	Yes	10	13.89%
	No	61	84.72%
	Not Clear	1	1.39%
Evaluation	Expert Review	22	30.56%
	Compare with Human-authored questions	15	20.83%
	Mock Exam	14	19.44%
	Automatic Evaluation	12	16.67%
	Student Review	10	13.89%
	Review(not clear by who)/Author Review	10	13.89%
	Crowd-sourcing	9	12.50%
	Compare with Another Generator	8	11.11%

Table 1: Results of Kurdi et al.s’ review. A study may include multiple purposes and question formats

Database	Conference	Filter by Task	No. of Search Results	No. of Studies Included
PwC	ACL 2019	Question Generation	5	1
	ACL 2020	Question Generation	5	4
	NeurlPS 2019	Question Generation	1	0
	NAACL 2019	Question Generation	1	0
	NAACL 2021	Question Generation	2	2
	ICLR 2020	Question Generation	1	1
	ICLR 2021	Question Generation	2	0
	EMNLP 2020	Question Generation	4	1
	IJCNLP 2019	Question Generation	3	2
	EACL 2021	Question Generation	3	2
	Findings of the Association for Computational Linguistics 2020	Question Generation	3	2
			Total: 28	Total: 15

Table 2: Search queries and results. **No. of Search Results** shows the total papers involved with question generation. **No. of Studies Included** refer the papers are under the category we are discussing.

3.1 Purpose of Question Generation

We found out that six of our reviewed papers apply AQG for data augmentation of question answering (QA), three aim to generate clarification questions, questions that identify important and missing information in the given text, one for boosting reading comprehension, and eight papers do not have clearly-stated purpose. The result is different from that of the review reported by Kurdi et al. (2020). (Table 1). As for domain, every paper falls into the "generic" category. Despite not included, we find the cross-lingual training method proposed by Kumar et al. (2019) useful for rare languages.

3.2 Generation Method

In this section, we will discuss several approaches commonly used in AQG. In Kurdi et al. (2020)'s review, Generation methods are classified based on the level of understanding and the procedure of transformation. Regarding the level of understanding, the two categories are (1) syntactic approach, which is defined as leveraging syntactic features of the input (i.e. part of speech), and (2) semantic approach, which requires deeper understanding than lexical and syntactic information, such as contextual similarity and named entities recognition. For example, obtaining information through semantic role labeling (Màrquez et al., 2008), which means identifying the semantic relations held among a predicate and its associated properties, are considered using a semantic approach.

As for the procedure of transformation, AQG has been mainly tackled by rule-based approach, defined as template-based in this survey along with the one reported by Kurdi et al. (2020), and neural QG approach (Du et al., 2017), classified as a "statistical method" in our paper and Kurdi et al. (2020)s' work. Following the categories purposed by Kurdi et al. (2020), we adopt a more detailed classification, adding rule-based into the categories. The three categories are as following: (1) template-based, which refers to structures consisting of fixed texts and spaces that will be substituted by values, (2) rule-based, which annotates the input to navigate the selection of a suitable question type and the manipulation of the input to construct questions, and (3) statistical methods, referring to learning the transformation to inquisitive sentences from training data.

3.2.1 Level of understanding

Level of understanding discusses the extend AQG systems comprehend the input text. According to Dhole and Manning (2020), whose system takes semantic roles as the heuristic information, relying on syntactic information alone is unlikely to obtain sufficient understanding for answering complicated questions that contain multiple ”wh” words. Nine studies (Lee et al., 2020; Dhole and Manning, 2020; Wang et al., 2020; Zhang and Bansal, 2019; Pan et al., 2020b; Chen et al., 2019; Fabbri et al., 2020 Cho et al., 2019b; Su et al., 2020) take advantage of both semantic and syntactic information, three systems (Alberti et al., 2019; Cho et al., 2019a; Qi et al., 2020) exploit only semantic features, and three of the included studies (Majumder et al., 2021; Pan et al., 2020a; Roemmele et al., 2021)only rely on syntactic features.

As shown in Table 1, Kurdi et al. (2020) suggests that most of the AQG studies from 2014 to early 2019 take semantic features into consideration, and we observe that the trend of performing AQG through semantic approach has become more and more prevalent among systems purposed between 2019 and early 2021.

3.2.2 Procedure of Transformation

We take the survey reported by Kurdi et al. (2020) as reference of the categories. As presented in Table 3, various statistical methods are the most popular, while the use of rules and templates each reported by one study. The results are different from that of the review by Kurdi et al. (2020)(see Table 1). Compared with rule-based and template-based techniques, which demands human effort including expert knowledge to construct guidelines and the variety of questions generated are limited, statistical approaches require far less labor and enable better language flexibility (Pan et al., 2020b; Tuan et al., 2019). We will succinctly introduce RNN-based (recurrent neural networks) and Transformer in the following section.

RNN-Based RNN-based QG models use encoder-decoder architecture to transform one sequence into another. The major drawback of RNN-based approaches is that they can only function sequentially, which makes them slow and suboptimal for longer sequences (Vaswani et al., 2017). Since Serban et al. (2016) and Du et al. (2017) applied neural-based approaches for AQG, many improvements of RNN-based models have been proposed. For instance, Du et al. (2017) adopt an attention mechanism to make the models focus on certain elements of the input.

Method	Approach	Studies
Statistical methods	RNN-based	8
	Transformer	4
	Graph to sequence	1
Template	-	1
Rule	-	1

Table 3: Procedure of Transformation. **Statistical methods** refers to the approaches in which systems are trained upon massive amount of data. In our study, three approaches are reported: RNN models, Transformer, and Graph to sequence. As for **Rule-based** and **Template-based** methods, the former defines the law of the question formation, the models have to generate the whole sequence; the latter has prewritten templates, the models only need to fill in the blanks.

Transformer Transformer was proposed by Vaswani et al. (2017). Like Seq2Seq, Transformer converts one sequence to another one with encoder and decoder. However, instead of recurrent networks, Transformer uses self-attention mechanism instead, which can be seen as the most important feature of Transformer. In self-attention, a word is operated with every other word, including those that appear later. Furthermore, since self-attention computation has no notion of the order of the inputs, parallelization is allowed and boosts the efficiency. Since word order is an important information as it may change the meaning of the input sentences, the relative positions of the words are added to the embedded representation (n-dimensional vector) of each word.

3.3 Paper Study

After discussing the generation methods, we will move on to the overview of the AQG studies from 2019 to early 2021. In the 15 papers we reviewed, 10 papers take various approaches including reinforcement learning, encoder-decoder, knowledge graph along with RNN, semantic graph, and rule-based method to tackle QG directly; 5 researches implement QG as a method of generating datasets or gather question-answer pairs for QA training. We will mainly describe those papers focusing on QG succinctly in the following paragraphs. Zhang and Bansal (2019) apply POS and NER to deep contextualized word vectors to enrich input information, along with self-attention mechanism and reinforcement learning implemented to solve the "semantic drift" problem in QG. Two semantics-enhanced rewards, QPP and QAP were proposed, the former refers to the probability of the generated question and the ground-truth question being paraphrased, and the latter stands for the probability of the generated question being correctly answered by the given answer. The proposed mechanism were obtained from downstream question paraphrasing and question answering tasks, aiming to improve the quality of questions generated by regularizing the QG model to produce semantically valid questions.

Being aware of the fact that ignoring structure information hidden in text or excessively relying on cross-entropy loss can lead to problems such as exposure bias, inconsistency between training and test measurements, and inability to fully exploit the answer information, Chen et al. (2019) propose a reinforcement learning based graph-to-sequence model for QG. Their model includes a Graph2Seq (Xu et al., 2018) generator with an encoder based on a Bidirectional Gated Graph Neural Network, which is introduced to learn the graph embeddings from the constructed text graph effectively. Authors also proposed a hybrid evaluator with objective that combines cross-entropy and RL losses to ensure syntactic and semantical validness. The paper further introduces an effective Deep Alignment Network for incorporating the answer information into the passage at both the word and contextual levels.

The semantically one-to-many relationships between source and target sentences in QG often leads to poor performance when trying to use standard Encoder-decoder model to generate a diverse and fluent output. Cho et al. (2019a) present a method for diverse generation that separates diversification and generation stages. The diversification stage takes advantage of content selection to map the source to multiple sequences, also known as "one-to-many mapping". The generation stage uses a standard encoder-decoder model to perform one-to-one mapping by generating a target sequence given each selected content from the source. In diversification stage, a new module named SELECTOR is proposed to identify key contents to focus on during generation.

Since failing to model fact information may cause QG systems to generate irrelevant and uninformative questions, Wang et al. (2020) defines a new task of question generation in which the system is given a query in the knowledge graph of the input content. The authors further divide the task into two steps, query representation learning and query-based question generation. First, the model learns a query representation which stands for the fact information that will be mentioned in the query path, then a RNN-based generator is employed to produce corresponding questions based on these facts.The two module were trained together in an end-to-end fashion, and the interaction between these two modules is enforced in a various framework.

Pan et al. (2020b) focus on Deep Question Generation (DQG) task, which aims to generate complex questions that require reasoning over multiple pieces of input information. Authors present an innovative structure consisting of three parts: semantic graph construction, semantic-enriched document representation, and joint-task question generation. The proposed model becomes the first research to construct a semantic-level graph of the input document and encode the semantic graph by introducing an attention-based GGNN (Li et al., 2015) in QG area. After that, the document-level and graph-level representations are fused to conduct joint training on content selection and question decoding. Their method allows models to capture the global structure of the document and facilitate reasoning, which greatly reduces semantic errors, increasing the quality of generated question, and improves performance on HotpotQA (Yang et al., 2018).

Multi-hop Question Generation also requires assembling and summarizing information from multiple relevant documents. (Gupta et al., 2020). Proposed by Su et al. (2020), Multi-Hop Encoding Fusion Network for Question Generation (MulQG), features context encoding in multiple hops with Graph

Convolutional Network and encoding fusion via an Encoder Reasoning Gate. The authors claim to be the first to tackle multi-hop reasoning over paragraphs without sentence-level information. Pan et al. (2020a) propose MQA-QG, an unsupervised framework for generating human-like multi-hop QA training data. MQA-QG generates questions by first selecting relevant information from each data source and then integrating the multiple information to form a multi-hop question. Using solely the generated training data, the authors successfully train a competent multi-hop QA system.

Roemmele et al. (2021) present a system that integrates QA and QG in order to produce QA pairs that convey the content of multi-paragraph documents. They explore the impact of different training data by having one system trained on SQUAD and NEWSQA, one on the production of rule-based QG systems, and one on both kinds of data; the latter is the most outstanding. Since their model performs extractive QA, in which answers to questions are extracted directly from the given text, the evaluation focus on whether questions are answerable and relative to the input text.

Dhole and Manning (2020) consider QG as a generally simple syntactic transformation influenced by semantics. They porposed Syn-QG, a QG system, to implement their obeservation. The system includes a set of transparent syntactic rules that utilize universal dependencies, shallow semantic parsing, lexical resources, and custom rules of transforming declarative sentences into question-answer pairs. The authors apply back-translation over the rules to improve syntactic fluency and eliminate grammatical errors at a slight cost of generating irrelevant questions. The crowd-sourced evaluations result shows that thier system can generate a larger number of grammatically correct and relevant questions than previous QG systems.

Questions also serve the need of acquiring information.Majumder et al. (2021) believe that the ability to generate questions that identify useful missing information in a given context is important, and to identify these information, humans compare global view consists of previous experience with similar contexts to the given context. The authors propose a model for clarification question generation in which "what is missing" is identified first by comparing the global and the local view and then a model identifies what is useful and generate a question about it. Qi et al. (2020) dedicate their research to the scenario in which the questioner is given the shared conversation history but not the context from which answers are drawn, thus must ask questions to obtain new information. To generate pragmatic questions, the authors use reinforcement learning to optimize an informativeness metric they propose, along with a reward function which encourages more specific questions.

In this paragraph, we will briefly introduce the researches aiming to generate question-answer pairs or obtaining training data for QA. Alberti et al. (2019) introduce a novel method of generating synthetic question answering corpora by combining models of question generation and answer extraction, and filtering the results to ensure roundtrip consistency. Significant improvements were obtained after pretraining on the resulting corpora. The authors also describe a variant that does full sequence-to-sequence pretraining for question generation, obtaining outstanding performance on SQuAD 2.0 (Rajpurkar et al., 2018). Fabbri et al. (2020) demonstrate that generating questions for QA training by applying a simple template on a related, retrieved sentence rather than the original context sentence allows the model to learn more complex context-question relationships thus improves unsupervised QA. To cope with the scarcity of question-answer pairs for a specific domain with human annotation, Lee et al. (2020) propose a hierarchical conditional variational auto encoder (HCVAE) for generating QA pairs from unstructured texts given as context and maximizing mutual information between generated QA pairs to ensure consistency.

3.4 Evaluation

According to Amidei et al. (2018), currently, the evaluation of automatic question generation includes a wide variety of both intrinsic and extrinsic evaluation methodologies. Since the evaluation of AQG has no exclusive, commonly agreed metric, most literature adopts multiple evaluation metrics. The statistics of our survey are provided in Table 5. Unlike the results of the review reported by Kurdi et al. (2020) (Table 1), the most common evaluation method is comparison with manually written ground truth questions. Since there is no common framework for evaluating AQG systems, researchers use n-gram models including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004). Note that none of the mentioned metrics were created specifically for the evaluation of AQG, BLEU and METEOR were designed for evaluating machine translation, while ROUGE aims to evaluate text summarization. Nema and Khapra (2018) has delineated that the evaluation of natural language generation systems, including those of AQG, using the aforementioned n-gram based similarity metrics sometimes shows poor correlation with human judgments in terms of answerability.

On the other hand, the variety of datasets used for evaluation also makes comparison between different models more difficult (Amidei et al., 2018). We noticed that except Pan et al. (2020b) and Cho et al. (2019a), other studies included the SQuAD (Rajpurkar et al., 2016) in their evaluation datasets or used it as the only source (See Table 4 for the details). Nevertheless, SQuAD 2.0 contains unanswerable questions written by crowdworkers while SQuAD 1.1 does not, which can affect the result of evaluation. Comparison with another generator remains the second most popular. 8 of the studies compare the results of automatic evaluation with baseline models and 6 studies compare with other models through human evaluation. The most common dimensions include fluency, relevance, syntactic or grammar correctness, with occurrences of four, three, and three, respectively.

4 Conclusion

In our survey, analysis of 15 AQG conference papers from PwC reported between 2019 and early 2021 is provided, taking the survey by Kurdi et al. (2020) as reference and tracking the development of the AQG field. Focusing on the purposes, methods, and evaluation of AQG, our findings are as follow:

Evaluation Method	No. of Studies
Compare with manually written ground truth through automatic evaluation	8
Compare with another generator	8
Crowd sourcing	4
Human review	3

Table 4: Evaluation Methods. Multiple evaluation methods can be implemented in one study. The statistics demonstrate that using ground truth written manually for evaluation, or using the answers from other QG generator models for comparison, is the mainstream evaluating method in recent years.

(1) Purposes of AQG

Recent studies tend to focus on data augmentation of QA. 6 of the 15 papers we review use AQG to generate QA training data.

(2) Generation Method

When it comes to the level of understanding, most AQG systems take semantic information into consideration since it provides the systems with more understanding to answer complicated questions. Regarding the procedure of transformation, Statistical methods have become more popular for the AQG task. Since Transformer provides self-attention and parallelization thus significantly boosts accuracy and efficiency, respectively, it is attracting increasing interest.

(4) Evaluation

Despite there being no widely acknowledged evaluation metric for AQG, researchers adopt automatic evaluation metrics for other NLP tasks to compare with human-authored questions and different models.

(5) Evolvement of AQG since Kurdi et al. (2020) s' survey

The results of our review differ from that of Kurdi et al. (2020). Kurdi et al. (2020) when it comes to the purpose of using AQG and the process of creating inquisitive sentences. We found out that recent researches tend to focus on data augmentation of QA systems instead of generating assessments, and using templates to convert input text into questions is gradually replaced by implementing RNN-Based methods and Transformer.

Dataset	Source	Development method	Content	OCC
SQuAD1.1	Wikipedia	Crowdsourcing	Questions and paragraph-answer pairs	9
SQuAD2.0	Wikipedia	Crowdsourcing	SQuAD1.1 plus unanswerable questions	2
Hotpot QA	Wikipedia	Crowdsourcing	QA pairs and evidence documents	4
Natural Questions (NQ)	Search queries issued to Google search engine	Crowdsourcing	Questions corresponding Wikipedia page, a long response and a short one	2
HarvestingQA	Wikipedia	Automatic	QA pairs and Wikipedia articles	1
TriviaQA	Web, Wikipedia	Crowdsourcing	QA pairs and evidence documents	1
DROP	Wikipedia	Crowdsourcing	Questions	1
Amazon Review	Amazon.com	Not specified	Relationships between objects, an image and a category label	1
Amazon Question-answering	Amazon.com	Collecting and labeling	Questions and answers about products	1
HybridQA	Wikipedia	Crowdsourcing	Multi-hop questions, Wikipedia table and passages linked with it	1
NEWSQA	News articles from CNN	Crowdsourcing	Questions and answers	1
MS-MARCO QA	Search queries issued to Bing or Cortana, web pages	Crowdsourcing	Questions, related web pages, crowdsourced answer and supporting information if answerable	1
QuAC	Wikimedia foundation	Crowdsourcing	Information-seeking QA dialogues	1

Table 5: Information of datasets used in reviewed studies. Of the 15 papers, a total of 13 datasets are used, including SQuAD, HotpotQA, Natural Question (Kwiatkowski et al., 2019), HarvestingQA (Du and Cardie, 2018), TriviaQA (Joshi et al., 2017), DROP (Dua et al., 2019), Amazon Review (McAuley et al., 2015), AmazonQuestion-answering (McAuley and Yang, 2016), HybridQA (Chen et al., 2020), NEWSQA (Trischler et al., 2016), MS-MARCO QA (Nguyen et al., 2016), QuAC (Choi et al., 2018). We also provide their data source, develop method, and content description of the data.

References

Renlong Ai, Sebastian Krause, Walter Kasper, Feiyu Xu, and Hans Uszkoreit. 2015. Semi-automatic generation of multiple-choice tests from mentions of semantic relations. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 26–33.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416*.

Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. 2016. Ontology-based multiple choice question generation. *KI-Künstliche Intelligenz*, 30(2):183–188.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. 2020. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2019. Reinforcement learning based graph-to-sequence model for natural question generation. *arXiv preprint arXiv:1908.04942*.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019a. Mixture content selection for diverse sequence generation. *arXiv preprint arXiv:1909.01953*.

Woon Sang Cho, Yizhe Zhang, Sudha Rao, Asli Celikyilmaz, Chenyan Xiong, Jianfeng Gao, Mengdi Wang, and Bill Dolan. 2019b. Contrastive multi-document question generation. *arXiv preprint arXiv:1911.03047*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Kaustubh D Dhole and Christopher D Manning. 2020. Syn-qg: Syntactic and shallow semantic rules for question generation. *arXiv preprint arXiv:2004.08694*.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. *arXiv preprint arXiv:1805.05942*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Alexander R Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. *arXiv preprint arXiv:2004.11892*.

Deepak Gupta, Hardik Chauhan, Akella Ravi Tej, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Reinforced multi-task approach for multi-hop question generation. *arXiv preprint arXiv:2004.02143*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. *arXiv preprint arXiv:1906.02525*.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent qa pairs from contexts with information-maximizing hierarchical conditional vaes. *arXiv preprint arXiv:2005.13837*.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2020. Transformer-based end-to-end question generation. *arXiv preprint arXiv:2005.01107*.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. *arXiv preprint arXiv:2104.06828*.

Lluís Màrquez, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635.

Preksha Nema and Mitesh M Khapra. 2018. Towards a better metric for evaluating question generation systems. *arXiv preprint arXiv:1808.10192*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2020a. Unsupervised multi-hop question answering by question generation. *arXiv preprint arXiv:2010.12623*.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020b. Semantic graphs for generating deep questions. *arXiv preprint arXiv:2004.12704*.

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

Peng Qi, Yuhao Zhang, and Christopher D Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. *arXiv preprint arXiv:2004.14530*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Melissa Roemmele, Deep Sidhpura, Steve DeNeefe, and Ling Tsou. 2021. Answerquest: A system for generating question-answer items from multi-paragraph documents. *arXiv preprint arXiv:2103.03820*.

Vasile Rus, Zhiqiang Cai, and Art Graesser. 2008. Question generation: Example of a multi-year evaluation campaign. *Proc WS on the QG-STEC*.

Mourad Sarrouti, Asma Ben Abacha, and Dina Demner-Fushman. 2020. Visual question generation from radiology images. In *Proceedings of the First Workshop on Advances in Language and Vision Research*, pages 12–18.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*.

Liana Stanescu, Cosmin Stoica Spahiu, Anca Ion, and Andrei Spahiu. 2008. Question generation for learning evaluation. In *2008 International Multiconference on Computer Science and Information Technology*, pages 509–513. IEEE.

Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020. Multi-hop question generation with graph convolutional network. *arXiv preprint arXiv:2010.09240*.

Will Thalheimer. 2003. The learning benefits of questions. *Work Learning Research*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Luu Anh Tuan, Darsh J. Shah, and Regina Barzilay. 2019. Capturing greater context for question generation. *CoRR*, abs/1910.10274.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Siyuan Wang, Zhongyu Wei, Zhihao Fan, Zengfeng Huang, Weijian Sun, Qi Zhang, and Xuan-Jing Huang. 2020. Pathqg: Neural question generation from facts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9066–9075.

Kun Xu, Lingfei Wu, Zhiguo Wang, Yansong Feng, Michael Witbrock, and Vadim Sheinin. 2018. Graph2seq: Graph to sequence learning with attention-based neural networks. *arXiv preprint arXiv:1804.00823*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. *arXiv preprint arXiv:1909.06356*.