Survey Paper On Question Generation

Baozheng Li labeerlee@gmail.com
Qichao Jiang jiangqichao564@gmail.com
Zimo Wang zimo2021gradschool@gmail.com

Abstract

Question Generation has received a lot of academic attention over the past few years, and hundreds of papers have been developed in the Question Generation field. Our literature review analyzes and compares technical details of all the methods employed in the modelling process including context preprocessing, encoding, decoding. Also we summarize the datasets and metrics that are used for model evaluation. Through our analysis and comparisons, we find that most current research has been trying to refine the modelling architectures in order to gain a better performance evaluated by the popular metrics, BLEU for example. However, the datasets can be biased. What's more, popular metrics cannot perfectly measure the quality of generated questions, and somtimes human evaluators are needed.. Consequently, we propose future work could focus on coming up with some unbiased datasets and smart metrics for evaluation in the field of question generation.

Introduction

Question Generation (QG) from text aims to automatically construct questions from textual input (Heilman and Smith, 2010). And a variety of Question Generation styles have been developed over the past few years. For example, answer-aware style QG, dialogue style QG, etc. QG has received increasing attention from research communities recently, due to its broad applications in scenarios of dialogue system and educational reading comprehension (Piwek et al., 2007). It can also help to augment the question set to enhance the performance of question answering systems.

We searched for over 30 papers published from 2018 to 2022 and investigated the technical aspects of these papers, aiming to provide a summary of the most recent trend in the QG field. Overall, we developed our work following the encoder-decoder modelling process. In other words, our literature review covers all aspects of the whole process to build a question generation model. In each workflow segmentation, our literature review shows different methods for that specific procedure. We make comparisons across methods and anlalyze the trends and the differences for handling different issues. And finally, we summarize the performance of these methods and propose a few potential research directions that might make improvements in the field of question generation.

Summary of Previous Reviews

Previous reviews were organised similarly. They first provided an overview of different types of questions that can be generated, then followed by the objective of the review. Then, the methodology used in practice is summarised through the comprehensive data collection, preprocessing or word embedding for example, and analysis of the intended purposes. After that, they illustrated the major phases for the proposed models. An evaluation on results, metrics and limitations was also conducted based on the review, and future work that can be extended and further explored was also implied in this section. Finally, they concluded the paper with overall findings.

Previous work suggested that by jointly training and learning between Question Answering and Question Generation, it would open up a wide field of study that gives promising direction and novel ideas(Wang et al., 2017); Question Generation can benefit from relevant context information, the overall objective of built systems aimed to make improvement on strong extractive Question Answering systems (Song et al., 2018); To be specific, Zhao et al. (2018) proposed a model that used paragraph-level context, and outperformed the results from state of the art at that time with sentence-level context. In terms of workflow, previous work by CH and Saha (2020) proposed a generic workflow consisting of six phases: Pre-processing, Sentence selection, Key selection, Question formation, Distractor generation, and Post-processing. In this paper we would adapt from this and combine with the information gathered to present a thorough review in an adequate format.

Review Objective

We investigated various models used in the past five years in chronological order, including basic encoder-decoder (seq2seq), Bi-LSTM with attention mechanism, Transformer and BERT-based models. From previous work, we managed to gather the final results in readable spreadsheet format to compare different models / metrics horizontally.

Review Contents

Overall Trends

Through our analysis, we find that the papers in the field of question generation modelling have formulated a structure for question generation tasks. Over the past several years, the data processing architecture has adapted significantly: The basic model that was used widely in earlier works is encoder-decoder model with fine-tuning, as more attention and effort being made in accomplishing the task, pre-trained models came in practice, namely the GloVe word vectors, GPT-2(Radford et al., 2019) and later T5 language model (Raffel et al., 2020).

The architectures

For most papers, the architecture follows an encoder-decoder norm, but the papers vary mainly on the encoder methods. And the mainstream papers primarily apply three kinds of encoder models: Seq2Seq, (Bi)LSTM and BERT models. Also, different decoders are applied to help generate the questions. Following is our comparison of the encoder methods and decoder methods:

Preprocessing

Preprocessing is the most attractive part that various methods are proposed to help combat different issues. Unlike the encoder-decoder architecture of the training and prediction process, the frameworks employed in the preprocessing part can be very costumed and vary a lot according to the settings of the problems.

Context

For most researchers, context processing is the most important part to make machine learning models take context showing up around the target words and characters into account to generate better predictions.

(Flor, 2018) Their SRL-based AQG system uses a mostly standard NLP pipeline structure with the following steps: 1) tokenization and sentence boundary detection; 2) POS tagging; 3) detection of verbal groups; 4) semantic role labeling; 5) post processing; 6) question generation.

（Sun et al, 2020）Sun implements a table-to-sequence approach that generates natural language questions from a structured table. Instead of using a sequence as input, their attention model is calculated over the headers, cells and the caption of a table. Ideally, the decoder should learn to focus on a region of the table when generating a word. In

this way, their model takes more context into account to generate a next word.

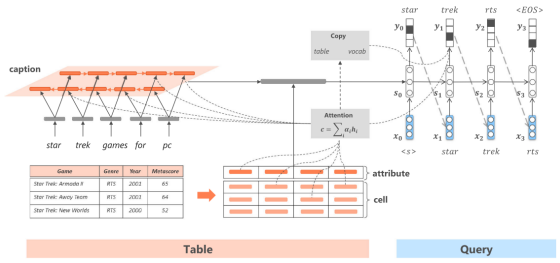

Figure 1: Architecture of table-to-sequence question generation model

(Dan Su, Yan Xu et al., 2021) Su and Xu use an answer-aware context encoder to help train the question generation model. To be specific, they split context and answer into word-level tokens, represent them by the pre-trained GloVe embeddings and append the answer tagging embeddings. Then they feed the context and answer embeddings into two bidirectional LSTM-RNNs separately to obtain their initial contextual representations. In addition, they also employ entity graph constructed with the name entities in context as nodes to recognize name entities from the context.

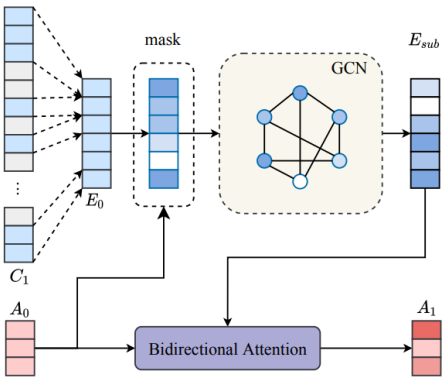

Figure 2: The illustration of GCN-based Entity-aware Answer Encoder.

(Siyuan Wang and Zhongyu Wei et al., 2020) In order to tackle the problem that existing models tend to generate irrelevant and uninformative questions, they employ a query (representation) learner to take a query path as input and learn the query representation. To be specific, they concatenate the embeddings and two BIO (beginning, inside, outside) tags as the input of the context encoder and use a bi-directional LSTM (Huang et al., 2015) to obtain the context states. In this way, their model is able to generate relevant and informative questions.

(Dugan et al., 2022) Instead of the original text, taking automatically generated summaries as input for answer-agnostic QG models results in significant increases in question acceptability (33%→83%), relevance (61%→95%), and in-context interpretability (56% → 94%).

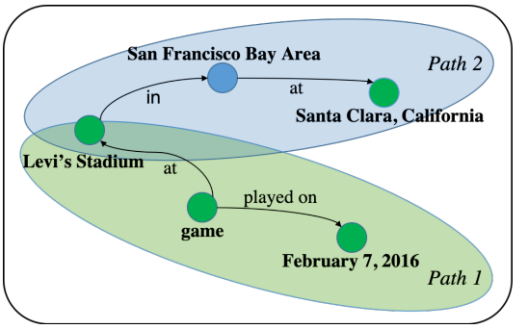

Figure 3: Knowledge graph constructed based on the input text shown in top sub-figure. Two colored ellipsoid are two query paths related to two ground truth questions

Encoders-Decoder

Basic encoder-decoder:

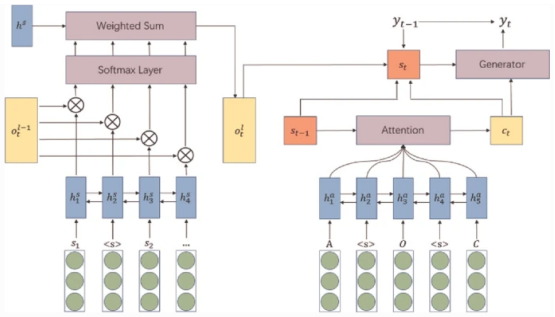

Figure 4: Overview of attribute-attention seq2seq module (Xie et al., 2021)

In the figure above, the bottom left part is the target sentences encoder, the bottom right part is the additional features encoder, and the top half part is the decoder. In the Encoder–Decoder framework proposed by Bahdanau et al. (2015), an encoder reads the input sentence, a sequence of vectors into

a vector. The decoder is often trained to predict the next word given the context vector and all the previously predicted words. It essentially means that the decoder defines a probability over the translation y by decomposing the joint probability into the ordered conditionals. After encoding the additional features and the target sentences by encoders, the decoder uses a one-layer uni-directional LSTM with attention modules to generate the questions employing the above encoded features. (Xie et al., 2021)

Bi-LSTM:

Bidirectional LSTM is absolutely one of the most popular models in the NLP field. Its property that it considers both the forward and backward word vectors make it an ideal tool to process the natural languages because natural languages often need contexts to fully identify the meanings of one single text. Sometimes, the authors may adopt multiple Bi-LSTM models to encode different lays of the text.

(Li, Gao 2019) They use two bidirectional LSTM models separately. They use the first Bi-LSTM to encode the sentence to capture a contextualized representation for each token. For the relation encoder, they firstly join all items in the n-ary relation into a sequence. Then they only take answer position embedding as an extra feature for the sequence. Similarly, they take another bidirectional LSTMs to encode the relation sequence and derive the corresponding contextualized representation.

Wang (2020) also adopts several Bi-LSTM models. They employ one bi-directional LSTM to get the context states which concatenates word embedding and tag embedding matrix. Since each entity or relation in the path is also a sequence of tokens, they take the average pooling of their word embeddings as input and feed the input to one another Bi-LSTM to calculate the sigmoid probability of each component in the path as its contribution weight.

(Nema, 2019) In this paper, they compute a contextualized representation for every word in the passage by passing the word embeddings through a bidirectional-LSTM.

Transformer:

Kriangchaivech and Wangperawong (2019) demonstrated that a transformer model can be trained on SQuAD to generate questions with correct grammar and relevancy to the context passage with answers provided. The model was trained on the inverted Stanford Question Answering Dataset (SQuAD). After training, the question generation model is able to generate simple questions relevant to unseen passages and answers containing an average of 8 words per question. Although the high average word error rate (WER) in evaluation suggests that the questions generated by transformer differ from the original SQuAD questions, the questions generated are mostly grammatically correct and plausible in their own right.

The attention-based model eschews the paradigm of existing recurrent neural networks (RNNs). Compared to RNNs used in prior studies, transformers allow us to more conveniently train and perform inference on longer sequence lengths.

BERT:

Chan and Fan, (2019) investigated the employment of the pre-trained BERT language model trained and evaluated on SQuAD to tackle question generation tasks. The model first revealed the defects of directly using BERT for text generation. The authors restructured the original BERT model and proposed BERT-HLSQG, a new model to generate one word at a time, using the encoded task inputs and the previously generated words as inputs to BERT. The best model outperforms previous RNN-based state-of-the-arts in terms of standard NLG metrics (BLEU, ROUGE, and METEOR)

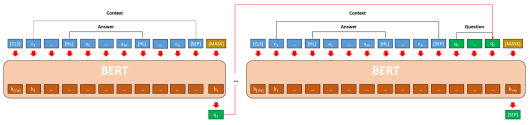

Figure 5: The BERT-HLSQG architecture (Chan and Fan, 2019)

Datasets:

A data-driven approach to reading comprehension traces back to Hirschman et al. (1999), who curated a dataset of 60 simulated news stories, followed by 3rd– 6th grade reading comprehension "5W" questions: who, what, when, where, and why. Later in 2013, Richardson et al. (2013) curated MCTest, which contains 500 stories and 2000 questions created by crowdworkers, with 4 multiple-choice questions per story and 4 answer choices per question. However, many of the questions require commonsense reasoning and reasoning across multiple sentences, making the dataset remain quite challenging for basic models.

(Kushman et al., 2014) collected a new corpus of 514 algebra word problems and associated equation systems gathered from Algebra.com for equation system generation. This is a creative experiment from natural language to mathematical question generation.

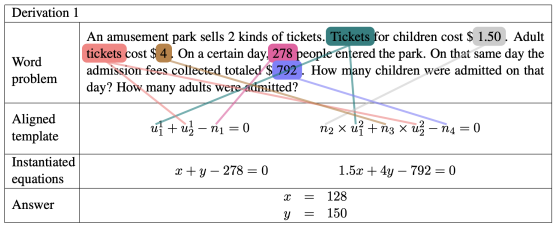

Figure 6: An example algebra word problem (Kushman et al., 2014)

(Weston et al., 2015) used a fully synthetic dataset (BAbI) that is stratified by 20 types of reasoning required to solve each task. The dataset is provided not only in English, but also in Hindi and shuffled English words so they are no longer readable by humans. A good learning algorithm should perform similarly on all three, which would likely not be the case for a method using external resources, a setting intended to mimic a learner being first presented with a language and having to learn from scratch.

Some popular and frequently used data sets are introduced as below:

NewsQuizQA

NewsQuizQA DATASET was introduced in 2020 by Adam et al., to address the lack of viable training data for the task of quiz-style QAG. The dataset contains 20K human written quiz-style question and answer pairs coming from 5k news articles between June 2018 to June 2020. These articles are summarised using a PEGASUS model fine-tuned on the CNN/Dailymail summarization dataset (Karl, et al., 2018), a state-of-the-art model for single document summarization. Only the summaries with exactly four human written question-answer pairs are kept for the final dataset. An 80-10-10 split is then randomly sampled from the 5k summaries to produce training, validation, and test sets, respectively. The NewsQuizQA dataset contains four typically diverse reference questions for each input passage, which can be used to calculate the minimum Reference Loss.

However, there is one challenge for the model to use the dataset: the size of the dataset is quite small in the number of news article summaries covered so it's hard to train a satisfying neural network model with only NewsQuizQA. Adam et al. introduced a mid-training phase to tackle the low-resource setting. In particular, they combine SQuAD, Natural Questions, and NewsQA and first fine-tune PEGASUS on this combined dataset. Although individually none of these datasets are suitable for the application, the combined examples do cover several qualities that are desired: length, quality, domain, and self-containment. They borrowed inspiration from T5 and prefixed all inputs with a label denoting the style they expect the model to generate. e.g. "Style SQuAD:", "Style NQ:", "Style NewsQA:". The model then is able to learn to associate certain styles of generated questions with certain labels in the input and does not have to infer this information from the contents of the input, which may be prone to overfitting. This is a very innovative method and will gain its popularity in the future.

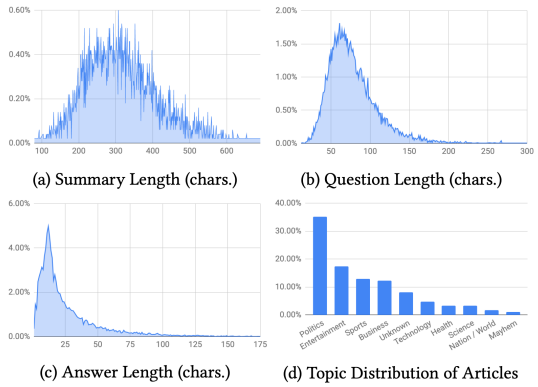

Figure 7: Visualisation of NewsQuizQA data statistics (Adam et al., 2020)

SQuAD

(SQuAD) is a new reading comprehension dataset consisting of 100,000+ questions posed by crowdworkers on 536 Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage.

Dataset	Question source	Formulation	Size
SQuAD	**crowdsourced**	**RC, spans in passage**	**100K**
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855
WikiQA (Yang et al., 2015)	query logs	IR, sentence selection	3047
TREC-QA (Voorhees and Tice, 2000)	query logs + human editor	IR, free form	1479
CNN/Daily Mail (Hermann et al., 2015)	summary + cloze	RC, fill in single entity	1.4M
CBT (Hill et al., 2015)	cloze	RC, fill in single word	688K

Table 1: A survey of several reading comprehension and question answering datasets. (Rajpurkar, Zhang, Lopyrev and Liang, 2016)

There are three major challenges when training model on SQuAD:

1. The dataset is diverse. Dates and other numbers make up 19.8% of the data; 32.6% of the answers are proper nouns of three different types; 31.8% are common noun phrases answers; and the remaining 15.8% are made up of adjective phrases, verb phrases, clauses and other types.
2. Reasoning required to answer questions. The authors (Rajpurkar, Zhang, Lopyrev and Liang, 2016) sampled 4 questions from each of the 48 articles in the development set, and then manually labelled the examples with the reasoning categories. The results show that all examples have some sort of lexical or syntactic divergence between the question and the answer in the passage.
3. Another challenging aspect of the dataset is the syntactic divergence between the question and answer sentence. The figure below shows that the more divergence there is, the lower the performance of the logistic regression model. Interestingly, humans do not seem to be sensitive to syntactic divergence, suggesting that deep understanding is not distracted by superficial differences. Measuring the degree of degradation could therefore be useful in determining the extent to which a model is generalising in the right way.

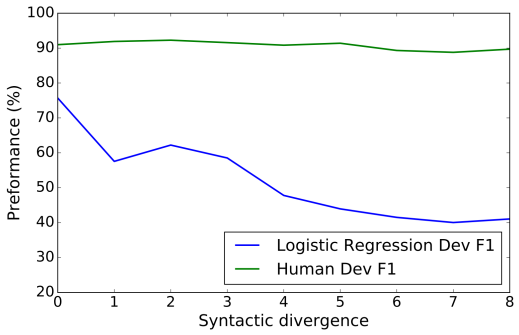

Figure 8: Performance of logistic regression model stratified by syntactic divergence of questions and sentences. (Rajpurkar, Zhang, Lopyrev and Liang, 2016)

WikiQA

Microsoft (Yang, Yih and Meek, 2015) created the WikiQA dataset, which, like SQuAD, uses Wikipedia passages as a source of answers, but the answer of WikiQA is sentence selection, while SQuAD requires selecting a specific span in the sentence. The dataset contains 3,047 questions originally sampled from Bing query logs. WikiQA offers us the opportunity to evaluate QA systems on answer triggering, a new challenge for the question answering problem, which requires QA systems to: (1) detect whether there is at least one correct answer in the set of candidate sentences for

the question; (2) if yes, select one of the correct answer sentences from the candidate sentence set. Both SQuAD and WikiQA have driven significant advances in reading comprehension, but systems now outperform humans and harder challenges are needed.

NewsQA

(Trischler et al., 2017) presented NewsQA, a challenging machine comprehension dataset of over 100,000 human-generated question-answer pairs. Crowdworkers supply questions and answers based on a set of over 10,000 news articles from CNN, with answers consisting of spans of text in the articles.

NewsQA is closely related to the SQuAD dataset:

1. It is crowdsourced, with answers given by spans of text within an article.
2. There are no candidate answers from which to choose.

The challenging characteristics of NewsQA that distinguish it from SQuAD are as follows:

1. Articles in NewsQA are significantly longer (6x on average) and come from a distinct domain.
2. The collection process of NewsQA encourages lexical and syntactic divergence between questions and answers.
3. A greater proportion of questions requires reasoning beyond simple word and context matching.
4. A significant proportion of questions have no answer in the corresponding article.

NewsQA offers a greater challenge to existing comprehension models. Given their similarities, we believe that SQuAD and NewsQA can be used to complement each other, for instance to explore models' ability to transfer across domains.

Answer type	Example	Proportion (%)
Date/Time	March 12, 2008	2.9
Numeric	24.3 million	9.8
Person	Ludwig van Beethoven	14.8
Location	Torrance, California	7.8
Other Entity	Pew Hispanic Center	5.8
Common Noun Phr.	federal prosecutors	22.2
Adjective Phr.	5-hour	1.9
Verb Phr.	suffered minor damage	1.4
Clause Phr.	trampling on human rights	18.3
Prepositional Phr.	in the attack	3.8
Other	nearly half	11.2

Table 2: The variety of answer types appearing in NewsQA (Trischler et al., 2017)

Natural Questions,

The progress on QA has been hindered by a lack of appropriate training and test data. To address this, Google (Kwiatkowski et al., 2019) presented the Natural Questions corpus. This is the first large publicly available data set to pair real user queries with high-quality annotations of answers in documents. Natural Questions contains quadruples (question, wikipedia page, long answer, short answer) where: the question seeks factual information; the Wikipedia page may or may not contain the information required to answer the question; the long answer is a bounding box on this page containing all information required to infer the answer; and the short answer is one or more entities that give a short answer to the question, or a boolean yes or no. Both the long and short answer can be NULL if no viable candidates exist on the Wikipedia page.

CNN/Daily Mail

The CNN/Daily Mail corpus (Hermann et al., 2015) consists of news articles scraped from those outlets with corresponding cloze-style questions. Cloze questions are constructed synthetically by deleting a single entity from abstractive summary points that accompany each article (written presumably by human authors). As such, determining the correct answer relies mostly on recognizing textual entailment between the article and the question. The named entities within an article are identified and anonymized in a preprocessing step and constitute the set of candidate answers.

Hotpot QA

This dataset is used by Xie et al. (2021) to verify the proposed method's efficiency, originally created by Yang et al. (2018). HotpotQA is a question answering dataset featuring natural, multi-hop questions, with strong supervision for supporting facts to enable more explainable question answering systems. It consists of 113k Wikipedia-based question-answer pairs with four key features: (1) the questions require finding and reasoning over multiple supporting documents to answer; (2) the questions are diverse and not constrained to any pre-existing knowledge bases or knowledge schemas; (3) sentence-level supporting facts required for reasoning, allowing QA systems to reason with strong supervision and explain the predictions; (4) a new type of factoid comparison questions to test QA systems' ability to extract relevant facts and perform necessary comparison.

MS MARCO

This dataset is presented by Nguyen et al. (2016), it stands for MAchine Reading COmprehension. The major goal was to do open-domain question answering over real world data, it consists of segmented data from user logs, with 100k queries, answers from human generated data. It contains over 200K documents and 1M passages. This dataset provides training data with question-answer pairs, where only a single answer text is provided via crowdsourcing, has eliminated major disadvantages in existing datasets, for example the requirement that the answers to questions have to be restricted to an entity or a span from the existing text.

However, as pointed out by Rajpurkar et al. (2016), most datasets suffer from one of two shortcomings: those that are designed explicitly to test comprehension are too small for training data-intensive deep learning models, while those that are sufficiently large for deep learning are generated synthetically, yielding questions that are not posed in natural language and that may not test comprehension directly.
SQuAD and NewsQA overcomes these deficiencies as it contains crowdsourced natural language questions and thus become the most frequently used dataset in AQG in recent years.

Metrics:

word error rate (WER)

Word error rate is measured from Levenshtein distance between two strings, including substitution, deletion and insertion between reference string and hypothesis string. WER was a useful metric in measuring the performance of speech recognition, in Transformers model it is used to compare the similarity between SQuAD questions and the model-generated questions (Kettip et al., 2019)

ROUGE

Namely Recall-Oriented Understudy for Gisting Evaluation, are the scores for evaluating how the generated text matches the reference outputs. According to (Adam et al., 2021), ROUGE scores have become the most commonly used metric for evaluating natural language generation models, measuring n-gram overlap between the generation and the reference output, in their work, they measured the ROUGE2-F1 scorer of bigram overlap between the generated text and the reference. There are a lot of variations for ROUGE:

1. ROUGE-1 refers to the overlap of unigrams between the system summary and reference summary.
2. ROUGE-2 refers to the overlap of bigrams between the system and reference summary.
3. ROUGE-N measures unigram, bigram, trigram and higher order n-gram overlap.
4. ROUGE-L measures the longest matching sequence of words using LCS (longest common subsequence). An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length.
5. ROUGE-S is any pair of words in a sentence in order, allowing for arbitrary gaps. This can also be called skip-gram occurrence. For example, skip-bigram measures the overlap of word pairs that can have a maximum of two gaps in between words. As an example, for the

phrase “cat in the hat” the skip bigrams would be “cat in, cat the, cat hat, in the, in hat, the hat”.

6. ROUGE 2.0 (Ganesan, 2018) is a Java implementation of ROUGE with improved and updated scoring. It allows capturing of semantic overlap through the use of a synonym dictionary and it also allows for evaluation of specific topics or subset of content.

BLEU

BLEU (Bilingual Evaluation Understudy), a commonly used algorithm designed for evaluating the quality of output text from language that is machine-translated to another. The BLEU scores are frequently used as one of the baseline metrics throughout the review, since there is a high correlation between manual supervision and quality, most of the works suggested that their models can outperform state of the art on this metric.

Perplexity

Perplexity is the multiplicative inverse of the probability assigned to the test set by the language model, normalised by the number of words in the test set. If a language model can predict unseen words from the test set, i.e., the P(a sentence from a test set) is highest; then such a language model is more accurate. As a result, better language models will have lower perplexity values or higher probability values for a test set. Since BLEU only measures a hard matching between references and generated text, (Zhou, Zhang and Wu, 2019) further adopt perplexity and distinct (Li et al., 2016) to judge the quality of generated questions in their paper *Multi-Task Learning with Language Modelling for Question Generation*. Using perplexity as metric helps the model to generate more fluent and readable questions. Besides, the generated questions have better diversity.

METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering), is a metric used to evaluate machine translation output, in practice it calculates the similarity between generations and references by considering synonyms, stemming and paraphrases in (Du et al., 2017). It focuses on recall over precision, throughout the review process, over 35% of works brought this metric into evaluation and the result on the same dataset shows that the correlation with human judgement is significantly higher than BLEU.

Performance for different models

Overall, the transformer and BERT outperforms all the RNN/LSTM models in terms of standard NLG metrics (BLEU, ROUGE, METEOR) and of whether a standard QA model can correctly answer the generated questions in AQG (Chan and Fan, 2019).

The inherent sequential nature of the RNN/LSTM models suffers from the issue of encoding and decoding long context/sequences. In contrast, the most recent models based on BERT and transformer addresses the problem perfectly and the questions generated by these models are more semantically coherent and fluent.

	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	ROUGE-L
SQuAD 73K	NQG-RC	42.54	25.33	16.98	11.86	16.28	39.37
	PLQG	45.07	29.58	21.60	16.38	20.25	44.48
	BERT-QG	37.49	18.32	10.47	6.10	16.80	41.01
	BERT-SQG	**50.00**	34.54	25.98	20.11	23.88	48.12
	BERT-HLSQG	49.73	**34.60**	**26.13**	**20.33**	**23.88**	**48.23**
SQuAD 81K	PLQG	45.69	30.25	22.16	16.85	20.62	44.99
	BERT-QG	32.61	14.50	7.70	4.08	14.18	37.94
	BERT-SQG	50.89	35.49	26.87	21.04	24.25	48.66
	BERT-HLSQG	**51.54**	**36.45**	**27.96**	**22.17**	**24.80**	**49.68**

Table 3: Comparison between RNN-based models (NQG-RC and PLQG) and the BERT-based models using paragraph level context (Chan and Fan, 2019)

- NQG-RC (Du et al., 2017): A seq2seq question generation model based on bidirectional LSTMs.
- PLQG (Zhao et al., 2018): A seq2seq network which contains a gated self-attention encoder and a maxout pointer decoder to enable the capability of handling long text input. The PLQG model is the state-of-the-art model for QG tasks.

Limitations

1. One common issue that appears in almost all the papers is that questions generated by the models are good but with a very low BLEU/WER/ROUGE/METEOR score. The problem for this result comes from that these metrics are token-basis; the generated question is compared with a golden standard based on the token similarity. A generated question might be semantically close to the gold question but syntactically different and thus results in a low score, while two questions with syntactic divergence but similar meaning could result in a high score. In the future, some better metrics for AQG are to be developed.
2. Earlier models suffer from long input sequences due to the drawback of basic encoder-decoder architecture, since it lacks track of previous words in the sequence. With the introduction of attention mechanisms, transformer-based models and pre-trained models, this problem has been gradually eliminated over time.

Conclusions and Future Work

Through the research above, we attempted to dig into the details of how various of the methods are employed in previous literature. And through these papers, we obtain a high level understanding on how particular methods fit into specific circumstances and what outputs will be generated.

In terms of the model architecture, data processing workflow has changed a moderate amount. The basic model that was used widely in earlier works is encoder-decoder model with fine-tuning, as more attention and effort being made in accomplishing the task, pre-trained models came in practice, namely the GloVe word vectors, GPT-2 (Radford et al., 2019) and later T5 language model (Raffel et al., 2020).

In the preprocessing part, we presented a few innovative methods which can help the encoder-decoder machine learning models better understand the meaning of natural languages, including Flor's (2018) SRL-based AQG system, Sun et al's (2020) table-to-sequence approach, Dan Su, Yan Xu et al's (2021) answer-aware context encoder, Dugan et al.'s (2022) answer-agnostic QG models.

For the encoder-decoder models, we classify the models into three popular categories: basic encoder-decoder (Seq2Seq), Bi-LSTM and BERT. We summarised a few papers for each category. In this way, we provided a horizontal overview to the papers that employ these models and comparisons over different circumstances where the models are adapted.

Overall, the modelling architecture has developed over the past years, from simple rule-based preprocessing to complicated named-entity based, knowledge graph based preprocessing methods, from basic encoder decoder models to BERT models. And the model performance has improved greatly.

Future Work

There are a huge amount of Question Generation datasets that are specified for different question generation tasks. Given the fact Neural Network is data consuming, one limitation is that not a single dataset is big enough to train a satisfying model. The strategy is that we can train the model first on weakly supervised data from existing question generation datasets. For example, in the paper from Google last year, the authors introduced a new dataset called NewsQuizQA with only 20K quiz-style question-answer pairs, which is obviously a low-resource dataset, to build a model for quiz-style question generation. They proposed the middle-training phase, they combined SQuAD, Natural Questions, and NewsQA and first fine-tune their model on this combined dataset. Although

individually none of these datasets are suitable for their application, the combined examples do cover several qualities that are desired, in terms of length, quality, and domain. And they borrowed inspiration from T5 to prefix all inputs with a label denoting the style to prevent overfitting. This is a very innovative way to make use of the existing datasets to tackle the low-resource Question Generation setting.

Gold	Generated	WER	BLEU-Score(1gram)
where was PERSON 2 born?	what is the birth place of PERSON 2?	4	3/9 = 0.33
where was PERSON 2 born?	when was PERSON 2 born?	1	5/6 = 0.83

Table 4: Two illustrations on metric problems

Another limitation that appears in almost all the papers is that questions generated by the models have a very low evaluation score even though the questions are good. In the first example in the table, the generated question might be semantically close to the gold question but syntactically different and thus results in a high word error rate and a low BLEU score. In the second example, two questions with semantic divergence but similar grammer could result in a low word error rate and high BLEU score. The problem for this issue comes from that these metrics are token-basis. As a result, the value of these popular metrics is limited and the evaluation relies a lot on the human raters, which is expensive and subjective. In the future, some better metrics for question generation are to be developed.

References

[1] Dhawaleswar Rao CH and Sujan Kumar Saha. 2020. Automatic Multiple Choice Question Generation From Text: A Survey. *IEEE Transactions on Learning Technologies*, 13(1):14-25.

[2] Jiayuan Xie, Wenhao Fang, Yi Cai and Zehang Lin. 2021. Comparison Question Generation Based on Potential Compared Attributes Extraction.

[3] Jishnu Ray Chowdhury, Debanjan Mahata and Cornelia Caragea. 2022. On the Evaluation of Answer-Agnostic Paragraph-level Multi-Question Generation.

[4] Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang and Daniel Gildea. 2018. Leveraging Context Information for Natural Question Generation.

[5] Sathish Indurthi, Dinesh Raghu, Mitesh Khapra and Sachindra Joshi. 2017. Generating Natural Language Question-Answer Pairs from a Knowledge Graph Using a RNN Based Question Generation Model.

[6] Sharon Look. 2016. Question Generation. Edition.

[7] Tong Wang, Xingdi Yuan and Adam Trischler. 2017. A Joint Model for Question Answering and Question Generation.

[8] 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.

[9] Xinya Du, Junru Shao and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension.

[10] Yao Zhao, Xiaochuan Ni, Yuanyuan Ding and Qifa Ke. 2018. Paragraph-level Neural Question

Generation with Maxout Pointer and Gated Self-attention Networks.

[11] Yibo Sun, Duyu Tang, Nan Duan, Tao Qin, Shujie Liu, Zhao Yan, Ming Zhou, Yuanhua Lv, Wenpeng Yin, Xiaocheng Feng, Bing Qin and Ting Liu. 2020. Joint Learning of Question Answering and Question Generation. *IEEE Transactions on Knowledge and Data Engineering*, 32(5):971-982.

[12] Adam D. Lelkes, Vinh Q. Tran, and Cong Yu. 2020. Quiz-Style Question Generation for News Stories. In Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia. ACM, New York, NY, USA, 11 pages.

[13] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (Montreal, Canada) (NIPS'15). MIT Press, Cambridge, MA, USA, 1693–1701.

[14] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, 2383–2392.

[15] Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep Read: A Reading Comprehension System. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 325–332, College Park, Maryland, USA. Association for Computational Linguistics.

[16] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

[17] Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to Automatically Solve Algebra Word Problems. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 271–281, Baltimore, Maryland. Association for Computational Linguistics.

[18] Jason Weston, Antoine Bordes, Sumit Chopra, Tomas Mikolov, Alexander M. Rush, Bart van Merriënboer, "Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks", arXiv:1502.05698 [cs.AI].

[19] Clark, P., & Etzioni, O. (2016). My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI. AI Magazine, 37(1), 5-12.

[20] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov and Christopher D. Manning. 2018. HotpotQA: A

Dataset for Diverse, Explainable Multi-hop Question Answering.

[21] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

[22] Dugan, L., Miltsakaki, E., Upadhyay, S., Ginsberg, E., Gonzalez, H., Choi, D., Yuan, C. and Callison-Burch, C., 2022. A Feasibility Study of Answer-Agnostic Question Generation for Education. [online] arXiv.org. Available at: <https://arxiv.org/abs/2203.08685> .

[23] Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P. and Suleman, K., 2017. NewsQA: A Machine Comprehension Dataset. [online] Microsoft Research. Available at: <https://www.microsoft.com/en-us/research/publication/newsqa-machine-comprehension-dataset/> .

[24] Hermann, K., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M. and Blunsom, P., 2015. Teaching Machines to Read and Comprehend. [online] arXiv.org. Available at: <https://arxiv.org/abs/1506.03340> .

[25] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M., Dai, A., Uszkoreit, J., Le, Q. and Petrov, S., 2019. Natural Questions: A Benchmark for Question Answering Research.

[26] Ying-Hong Chan and Yao-Chung Fan. 2019. A Recurrent BERT-based Model for Question Generation. In Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

[27] Wang, S., Wei, Z., Fan, Z., Huang, Z., Sun, W., Zhang, Q. and Huang, X., 2020. PathQG: Neural Question Generation from Facts. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),.

[28] Flor, M. and Riordan, B., 2018. A Semantic Role-based Approach to Open-Domain Automatic Question Generation. Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications,.

[29] Li, J., Gao, Y., Bing, L., King, I. and Lyu, M., 2022. Improving Question Generation With to the Point Context. [online] arXiv.org. Available at: <https://arxiv.org/abs/1910.06036v2> .

[30] Nema, P., Mohankumar, A., Khapra, M., Srinivasan, B. and Ravindran, B., 2022. Let's Ask Again: Refine Network for Automatic Question Generation. [online] arXiv.org. Available at: <https://arxiv.org/abs/1909.05355> .

[31] Chali, Y. and Baghaee, T., 2018. Automatic Opinion Question Generation. Proceedings of the 11th International Conference on Natural Language Generation,.

[32] Su, D., Xu, Y., Dai, W., Ji, Z., Yu, T. and Fung, P., 2020. Multi-hop Question Generation with Graph Convolutional Network. Findings of the Association for Computational Linguistics: EMNLP 2020,.

[33] Song, L., Wang, Z. and Hamza, W., 2022. A Unified Query-based Generative Model for Question Generation and Question Answering. [online] arXiv.org. Available at: <https://arxiv.org/abs/1709.01058> .

[34] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P., 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. [online] arXiv.org. Available at: <https://arxiv.org/abs/1606.05250> .

[35] Ganesan, K., 2018. ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. [online] arXiv.org. Available at: <https://arxiv.org/abs/1803.01937> .

[35] Zhou, W., Zhang, M. and Wu, Y., 2019. Multi-Task Learning with Language Modeling for Question Generation. [online] arXiv.org. Available at: <https://arxiv.org/abs/1908.11813v1> .

[36] Heilman, M. and Smith, N., 2010. Good Question! Statistical Ranking for Question Generation. [online] ACL Anthology. Available at: <https://aclanthology.org/N10-1086> [Accessed 28 April 2022].

[37] Piwek, P., Hernault, H., Prendinger, H. and Ishizuka, M., n.d. T2D: Generating Dialogues Between Virtual Agents Automatically from Text. Intelligent Virtual Agents, pp.161-174.