JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, 571-589 (2022)
DOI: 10.6688/JISE.202205_38(3).0005

Question Generation for Reading Comprehension Test Complying with Types of Question*

JUNJIE SHAN[1], YOKO NISHIHARA[2], AKIRA MAEDA[2] AND RYOSUKE YAMANISHI[3]
[1]*Ritsumeikan Global Innovation Research Organization*
[2]*College of Information Science and Engineering*
Ritsumeikan University
Shiga, 525-8577 Japan
[3]*Faculty of Informatics*
Kansai University
Osaka, 569-1095 Japan
E-mail: {shan; nisihara}@fc.ritsumei.ac.jp

In this paper, we proposed a method to generate two different types of reading comprehension questions complying with types of question for language learning tests with the Transformer model and the seq2seq method. In recent years, many approaches have showed good results by treating question generation as a seq2seq task. These approaches were implemented with a question-answering (QA) dataset; however, few studies have considered a reading comprehension-based dataset. Therefore, this paper proposed a method to generate questions appropriate for reading comprehension tests from articles. Moreover, analysis of reading comprehension test questions revealed two primary types of the question's asking style: the commonly-used question (CM question) and the directly-related question (DR question). The characteristic of the two question types was different and therefore needs to design the generation models complying with the type of questions. We proposed a method to separate the two question types in the dataset and used two models to generate both types, comparing the result with the method that generates the two types of questions together. The positive rate for the proposed method's CM questions was 88% and for its DR questions was 49%, compared to 33% and 24%, respectively, for the comparative method. The evaluation showed that the proposed method could generate the two types of reading comprehension questions more effectively, with a positive rate increased by an average of 40%.

Keywords: question generation, reading comprehension tests, language learning, attention mechanism, transformer model, Seq2Seq

1. INTRODUCTION

Question generation for reading comprehension tests is a time-consuming and labor-intensive work, and so has attracted researchers' widespread attention. Classic approaches tried to generate questions for the given reading article through the rule-based method. In recent years, a new method that converts the question generation (QG) into a sequence-to-sequence (seq2seq) task has been widely used, with many positive results. Unlike the rule-based method, the seq2seq approach is based on artificial neural networks and neural language models, and therefore requires large amounts of training data. Almost all studies on the seq2seq method are implemented with a question-answering (QA) based dataset called SQuAD1 because it has enough manually prepared sentence-question pairs for the seq2seq

Received February 1, 2021, revised April 6, 2021; accepted May 18, 2021.
Communicated by Chia-Hui Chang.
* This work is partly supported by a grant from the Kitano Foundation of Lifelong Integrated Education.

model's training. Differing from the reading comprehension test, questions in the QA dataset resemble information retrieval that is, asking about detailed facts of the given article. In contrast, questions in reading comprehension tests, especially in foreign language learning tests, require more information from given articles when answering.

Table 1, for example, shows two questions from the QA dataset and a reading comprehension test of language learning: Question A is from the QA-based dataset and question B is from the reading comprehension test. The answer to question A comes directly from words or phrases in the given article, but to answer question B, examinees need to understand the whole reading material (reading article) and compose the answer in their own words. According to Day's classification [1], reading comprehension questions that could find the answer directly from the reading material are the most basic. However, reading comprehension questions in language learning tests are more likely to emerge from a broader span of the reading material and need to integrate more information from the reading material to answer.

Therefore, we attempted to apply the new seq2seq method to the reading comprehension-based dataset [26]. Questions in reading comprehension tests will not only have a single answer pattern, but also several other characteristics due to the type. For example, reading comprehension questions ask about detailed information, like specific names or places, but also ask about the entire given article, for instance, composing a suitable title for it. Through analysis of the questions in reading comprehension tests, we found there were two primary types of the questions' asking style: the commonly-used (CM) question and the directly-related (DR) question. Each of the types has different focuses on reading comprehension tests and would affect each other in model training. Thus, for improved accuracy in QG, we designed different models for each type of question, that comply with the characteristics of the question types.

In this paper, we proposed a method to generate questions for reading comprehension tests of language learning, using models that comply with the question type2. First, the proposed method classified the two types of questions and separated them from the dataset. Second, the method prepared training data with different strategies: For CM questions, each was placed with an entire given article into an Article-Question pair. For DR questions, the proposed method extracted question-related sentences from the given article and placed them in Sentences-Question pairs. Third, the proposed method trained two Transformer models for QG with the seq2seq method through prepared Article/Sentences-Question pairs. Finally, the proposed method used these two models to generate CM and DR questions separately from the input reading articles. Through the proposed method, we could use the seq2seq method to deal with the question generation of reading comprehension tests with a more complex composition of questions.

Table 1. Question examples from the QA dataset and reading comprehension test. Question A can be answered directly from the underlined sentence, but question B needs to analyze the whole reading material.

Question in QA dataset:
Reading Material: Somerton took over from Ilchester as the county town in the late thirteenth century, but it declined in importance and the status of county town transferred to Taunton about 1366. The county has two cities, Bath and Wells, and 30 towns. ...
Question A: how many cities are there in the county?

[1] https://rajpurkar.github.io/SQuAD-explorer/

[2] This paper's proposed method is based on research published at the TAAI2020 international conference [27].

Question in reading comprehension test:
Reading Material: From Chengdu to Jiuzhaigou, you can travel along the East Line not only the other line going past Dujiangyan and Wenchuan. It only takes about three hours to Jiangyou. ... Pingwu has the best royal Buddhist temples in Ming Dynasty, Bao'en Temple and the historic relics in the period of Three Kingdoms. Here we would like to recommend another hotel in Jiuzhaigou for you, ...
Question B: how many cities are there in the passage?

2. RELATED WORK

As a classic challenge in natural language processing, QG attracts much attention from researchers. Currently, two primary methods for QG tasks are the rule-based method and the seq2seq-based method. Rule-based methods attempt to reveal rules between the reading article and its corresponding questions and to cooperate with pre-designed templates in the generation task [2-7]. This kind of approach can generate questions with high accuracy and smooth expressions. However, the rule-based method's limitation soon becomes evident. Just like Desai *et al.* mentioned [4], a "heavily annotated corpus" is necessary for the rule-based method, and the types of questions generated are also limited by pre-designed rules and templates. Therefore, it is difficult for the rule-based method to generate questions from a common reading article corpus without annotation. In this research, we attempted to generate reading comprehension questions through the common article corpus of plain text.

Since 2017, with the rapid improvement of the neural language model, an approach based on the seq2seq model has cut a broad figure in QG tasks [8-13]. In the seq2seq method for QG, the input is sentences of plain text from the given reading article, and the output is the generated question corresponding to the input sentences. The seq2seq-based method eliminated the preparation requirement of a heavily annotated corpus and the creation of templates for rule-based methods. After several improvements, methods based on seq2seq can generate questions from a single sentence or a short text span [8, 10], from entire given reading articles [14], or from multiple sentences or paragraphs [15, 16]. To improve the usability, some methods also allow users to select specific words for generating questions [17], while others attempt to detect "important sentences" (or say key sentences) from the given article and use them for question generation [18]. Some recent studies also tried to implement the seq2seq method on the newly proposed Transformer model instead of on the traditional LSTM model [19, 20]. All these studies could effectively generate many questions from input reading materials, but they focused almost exclusively on generating questions that ask about direct, factual information in the article, for instance, appeared names, locations, times, and amounts. Existing studies usually implemented the seq2seq method on the QA-based dataset and therefore generated questions with the QA-like style that always ask about the direct information.

As mentioned in Section 1, early in 2005, Day *et al.* classified reading comprehension questions in language learning into six types and sorted them by difficulty in answering [1]. According to Day's classification, questions ask about "direct facts" (direct information) that could be answered "directly and explicitly" from the reading material are the most basic type of questions in reading comprehension tests. Questions in reading comprehension of language learning tests often could not be answered directly like those in the

QA dataset, but requiring students to answer them through understanding, analysis, inference, or prediction over a relatively long span of text in the given article. However, few studies used the seq2seq method to generate questions with a reading comprehension-based dataset. Considering the QA-based dataset's limitation and inspired by Day *et al.*, in this research, we proposed a method using the seq2seq-based approach to generate questions for the reading comprehension tests.

Previous research attempted to apply the seq2seq-based QG method on the reading comprehension dataset, but the positive rate of the generated questions was much lower than that with the QA dataset [26]. During the analysis of questions, we found there were two primary types of the questions' asking style in reading comprehension tests. These two types of questions have different characteristics that would affect each other during the training step and reduce the training effect of the seq2seq model. We speculated that generating the two types of questions separately would help to improve generation results. Therefore, our proposed method classified question types from the reading comprehension-based dataset and designed different generation models to correspond with each type's characteristics.

3. QUESTION DEFINITION AND DATASET

3.1 Question Definition

As we mentioned before, in this paper, we newly defined a reading comprehension-based question generation task, which is different from those previously proposed QA-based question generation tasks. Reading comprehension-based questions usually do not have manually annotated text span of correct answers or clear correspondences between the question and the given article. Through analysis of the reading comprehension questions in language learning tests, we found two primary question types: the commonly-used (CM) question and the directly-related (DR) question [26]. Usually asking about the given reading article as a whole, the CM question has a broad description that could be used for different articles. DR questions are specific to the given article and usually ask only about part of the article's content. Since CM questions consider the whole article while DR questions consider only partial content, in this paper, we set the generation of one CM question and several DR questions for each given reading article as the target. Table 2 lists examples of CM and DR questions in the reading comprehension-based dataset. The proposed method used two separate models to generate these two types of reading comprehension questions. In order to verify our proposed method, we chose the reading comprehension-based dataset of the RACE, which will be described in the next section.

Table 2. Examples of CM and DR questions in reading comprehension tests.

CM Questions: (more broad description) The passage mainly tells us that _.
According to the article, which of the following is true?
What is the main purpose of the passage?
DR Questions: (more specific content)
Why does the author ask the postman about his baby?
How many kinds of wines were mentioned in the passage?
To fight against hacking, some companies _.

3.2 RACE Dataset

The dataset used to verify the proposed method in this paper was RACE [21], which contains reading comprehension questions for English learning, from exercises and tests in Chinese middle and high schools. Unlike the QA-based SQuAD dataset, the original purpose of the articles and questions in RACE was to examine learners' English reading comprehension skills. Questions in the RACE dataset usually require learners to integrate the information from several sentences in the given article to answer. Meanwhile, questions in RACE do not have manual annotations of answers or clear one-to-one correspondence of sentences-question pairs, so most of the previously proposed state-of-the-art methods could not be directly used to the RACE dataset but need to prepare article/sentences-question pairs specific for the generation model design. Table 3 shows the information about the RACE dataset. The RACE dataset contains 25,135 articles (443,683 sentences) and 87,866 questions. On average, every 5.04 sentences provide one question. Table 4 shows the example of the RACE dataset's articles, questions, and answers (four options per question). All questions in RACE are considered solvable (answerable), which is also different from the SQuAD dataset.

Table 3. RACE dataset information.

Content	Number
Articles (Reading materials)	25,135
Sentences	443,683
Questions	87,866
Sentences/Questions	5.04 (= 443,683/87,866)

Table 4. Example of article, questions, and options in RACE dataset.

Article:
On a small farm in Mexico, there are no schools. A bus is the school! The driver of the bus is the teacher! It is a school bus, but it doesn't take children to school. It just goes round from place to place, and sometimes it comes to this farm. (omitted)
Question:
The bus school will _.
Answer Options: (Correct: D)

A. take children to school	B. stay there for lunch
C. take the fathers and mothers to school	D. go round from place to place

4. PROPOSED METHOD

As Fig. 1 shows, the two models for generating CM and DR questions were implemented in the following four steps. First, the proposed method classified all questions in the RACE dataset into the two types of CM and DR. Second, the method used different strategies to prepare training data for CM and DR questions. For CM questions, the proposed method placed the whole reading article and a CM question into an Article-Question pair (Article-CM pair). For DR questions, the proposed method used the attention mech-

nism to extract each DR question's related sentences from the given article and created a SentencesQuestion pair (Sentences-DR pair). Third, the proposed method used the prepared Article/Sentences-Question pairs to train the generation model of the seq2seq method for each question type. Forth, the method used the obtained two models to generate CM and DR questions separately for reading comprehension tests.

4.1 Question Classification

To address DR and CM questions separately, the proposed method first classified all questions in the RACE dataset into these two types by using a neural network approach. Each question was embedded through the BERT pre-trained model[3] and represented as a sentence vector of 768 dimensions. The neural network has one hidden layer with 128 hidden units. A manually prepared list containing 158 CM and 270 DR questions was used as the training data. The batch size was 64, and the epoch was 800 at the training step. The final accuracy on training set is 1.0 and the mean loss is 0.00668.

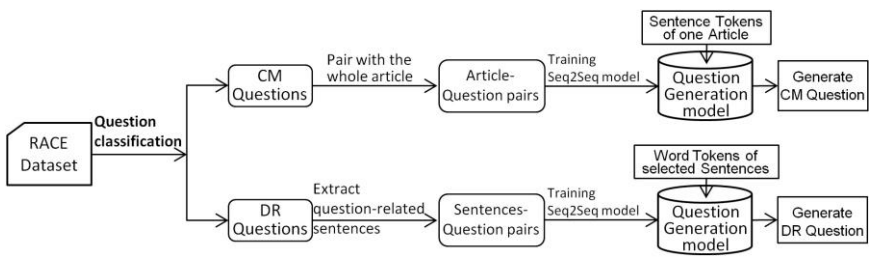

Fig. 1. Overall procedure of the proposed method.

Table 5. Classification results of CM and DR questions.

	Number of Classified	Accuracy of Randomly Selected 200
CM	23,408	98.0%
DR	64,437	97.5%

After the training step of the classification model, all questions were classified into either CM or DR question sets. The preliminary experiment to determine classification accuracy was conducted by randomly selecting 200 questions from each classification and checking them manually. Table 5 shows the results of the questions' classification. There were 23,408 CM questions and 64,437 DR questions that have been classified from the RACE dataset. (A few questions were removed at the data cleaning step because the given article lost or lacked information on the question.) Classification accuracy on the 200 randomly selected questions was 98% for the CM and 97.5% for the DR. We speculated that the reason why the classification model could distinguish the two types of questions in a high accuracy with few of the training data, is benefit from the fine-tuning step of the pre-trained embedding BERT model. The pre-trained BERT model has already had many of the learned features inside the textual embeddings, so it could be easily finetuned to high performance with a small amount of training data. These classified questions were later used to train the proposed question generation method.

[3] BERT-Base, Uncased, https://github.com/google-research/bert/

4.2 Pairing Article Sentences with a Question

For training the seq2seq-based QG models, each question was paired with its related sentences. Analysis of questions revealed that CM questions usually ask about the comprehension of the article as a whole, like its main purpose, topic, or what a suitable title would be. The proposed method placed the entire given article with each CM question as an Article-Question pair (Article-CM pair). DR questions usually ask about part of the article's content, like the facts, reasons, or inferences, *etc.* Therefore, the proposed method tried to extract the related sentences of each DR question and placed them as a Sentences-Question pair (Sentences-DR pair).

Table 6. Correspondence between article content and the CM questions.

Article Content	Examples of CM Questions
Statements Introduction	Which of the following is true? What can be inferred from the passage?
Topic News Report	What is the passage mainly about? What is the best title of this passage? What is the purpose of this passage?
Story Knowledge Situation	What can we learn from the passage? From the text, we can know _. We can conclude from the passage that _.
Viewpoint Feeling	What is the main idea of this passage? The main purpose of the passage is _. Why did the author write this passage?
Advertisement Announcement Information	Which of the following is true? This passage is probably taken from _.

4.2.1 Article-CM pairs

The manual check of CM questions found that there was a correspondence between each CM question and the content of the given article. Table 6 shows the correspondences between article content and the CM questions. If the given article was an introduction or contained many statement sentences, the CM question might have been expressed as, "which of the following is true?" or "what can be inferred from the passage?" Similarly, if the article was a news report or had a main topic, the CM question might have been "what is the passage mainly about?" or "what is the best title of this passage?" or "what is the purpose of this passage?"

If CM questions have some fixed patterns or descriptions and each article consists of only one exact statement of content, we could use the rule-based templates to generate them. However, every CM question has lots of variant descriptions with the same meaning (Table 7) and the reading articles also contained several statements of content at the same time, so the rule-based method might not have been adequate for obtaining appropriate CM questions. A generation model could analyze the content of the given article and generate the appropriate question automatically.

4.2.2 Sentences-DR pairs

For DR questions, we used a method based on the attention mechanism to extract the

question-related sentences from the given article. In the attention mechanism, an indicator called the "attention score" shows the degree of relevance of input tokens to output tokens. The attention score was used to estimate the degree of relevance between each DR question and its related sentences.

Table 7. Variants of CM questions with same meaning.

Which of the following is (not) true?
Which of the following statements does not agree with the passage?
According to the story, which of the following is true?
What can be inferred from the passage? We can infer from the passage that _.

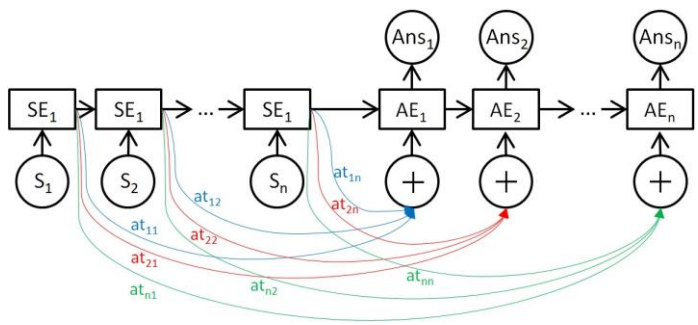

Fig. 2. Sentences question pairing using the attention mechanism. Each DR question was placed with sentences extracted from each article; S: Sentence; SE: Sentence Embedding; AE: Answer Embedding; Ans: Answer of Question; at: Attention Score.

Fig. 2 illustrates the structure of the attention mechanism-based method in creating Sentences-Question pairs. At the input part of the attention model, the given article was divided into sentences (s_1, ..., s_n) as tokens. At the output part, each question's correct answer was used to represent the original question, and the tokens were each of the correct answers (Ans_1, ..., Ans_n). The *correct answer* means to complete the question through the correct option. For example, the correct answer of the question "The bus school will _." in Table 4 would be "The bus school will go round from place to place _." The correct answer was used to represent the question because the correct answer's content was closer to the original article's wording so that related sentences could be extracted more accurately. The maximum length of tokens was set to 40 for input and 5 for output. The input/output excess the length was cut off, and insufficient places were filled with [PAD] tokens.

Each answer token of the output would have an attention score at_{ij} relating to all input sentence tokens. The attention scores were used as relevant degrees between the answer and all sentences to extract related sentences. The relevant degree of the s_j to the Ans_i $R(s_j|Ans_i)$ is shown as Eq. (1)

$$R(s_j|Ans_i) = at_{ij}. \quad (1)$$

In 41 manually checked DR questions, there were 89 related sentences. For each DR question, the number of related sentences was an average of about 2.2 (= 89/41). Therefore,

the proposed method chose to extract the top three sentences with the highest $R(s_j|Ans_i)$ value as the related sentences for each DR question. The set of extracted sentences was placed with the corresponding DR question to make a SentencesQuestion (Sentences-DR) pair.[4]

4.3 Training of Question Generation Model

After the Article/Sentences-Question pairs were prepared, the proposed method used them to train the Transformer model with the seq2seq method for question generation. The Transformer model, proposed by Google in 2017 [23], is a neural network model for dealing with the sequence task.

Fig. 3 shows the construction of the Transformer model that was used for the generation of CM questions. Because CM questions usually ask about the comprehension of the article as a whole, the proposed method used the entire article as the input of the CM generation model. The input article was divided into *sentence-level tokens* before placement into the encoder. "Sentence-level token" means the unit of the encoder sequence is each sentence, instead of each word, in the input article. Sentence-level tokens were used as the encoder sequence because they better reflect the holistic features of an article. The output CM question was divided into word tokens, as in the common way, and embedded at the word-level for training and generation.

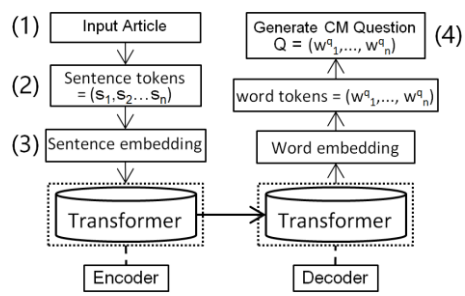

Fig. 3. Procedures of making a question generation model. This figure shows the case of the CM question generation. For DR question generation, the four blocks (1), (2), (3), and (4) must be changed: (1) Input three Sentences; (2) Word tokens; (3) Word embedding; and (4) Generate DR Question.

On the other hand, in training the generation model for DR questions, the encoder's input was the extracted and paired three-sentence set. Different from CM questions, DR questions ask about more specific content instead of the whole article. Therefore, the proposed method divided the input three sentences into word-level tokens to capture detailed information. The output DR question was also divided into word tokens and embedded at the word-level. Subsequent procedures were the same as those for CM questions. Table 8 shows information of the parameters' setting of the Transformer model.

Through the results and experience we obtained in [26], we chose the 4-layers model for CM questions and the 6-layers model for DR questions according to the positive rate of the generated questions. The *n*-layers means that there were each *n* layers of blocks in the Transformer model's encoder and decoder.

[4] Detailed content for using the attention mechanism to extract question-related sentences was published at the TAAI2020 international conference [27].

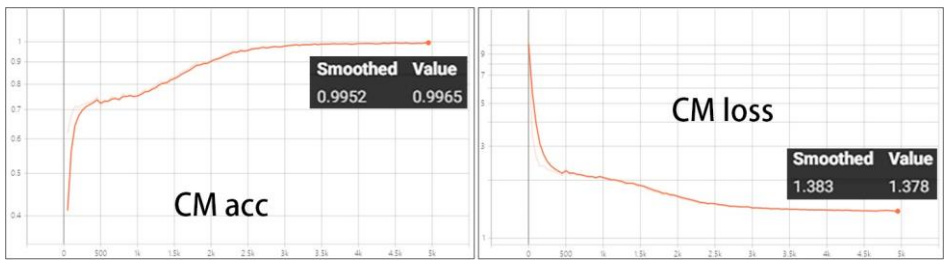

Fig. 4. Accuracy and mean loss curves of CM generation model's training process.

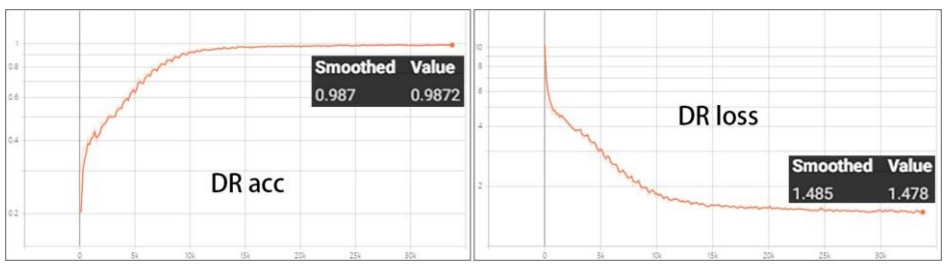

Fig. 5. Accuracy and mean loss curves of DR generation model's training process.

Figs. 4 and 5 show the change of the accuracy and mean loss of CM and DR questions' generation model during the training process. The final accuracy and mean loss of the CM generation model were 0.9965 and 1.378, while 0.9872 and 1.478 for the DR generation model. After the training step, we used these two models to generate CM and DR questions for the test set.

5. EVALUATION EXPERIMENTS

We evaluated the separately generated CM and DR questions and compared them with the generation result from a previously proposed method [26], which attempted to generate CM and DR questions together.

Table 8. Parameter settings of the transformer model.

Setting	Value
Input Max Length (Tokens)	40 for CM; 100 for DR
Output Max Length (Tokens)	30
Vocabulary Size	28,379
Attention Heads	8
Hidden Unit	768 for CM; 512 for DR
Learning Rate	0.0001
Dropout Rate	0.1
Training Epoch	40

5.1 Procedures

Questions in Table 5 were used as the training data for the generation model. After the training step, the test set of the RACE dataset was used to generate questions. For the

CM generation model, randomly selected 100 articles from the test set were input to the model to produce 100 CM questions. Similarly, the input to the DR generation model was a three-sentence set (the same as in the training setting). The three-sentence set was composed of a randomly selected sentence and its preceding and following sentences. We chose to use adjacent sentences instead of randomly selecting three sentences to avoid the selected sentences' content being irrelevant. The evaluation was conducted at randomly generated 100 CM questions and 100 DR questions. We chose the amount of 100 generated questions in evaluation through the experience from previously published related researches [8, 15, 16, 19].

5.2 Evaluation of Generated Questions

5.2.1 Basic evaluation of usefulness

First, we made a basic evaluation to check the usefulness of generated questions. The usefulness of each generated question was judged according to the following two aspects:

1. **Grammatical correctness:** whether the description of the generated question is a correct English sentence.
2. **Relevance with the given article:** whether the content of the generated question is related to the input reading article.
   - For generated CM questions, whether the generated questions correspond with the content of the input article was evaluated. The arranged correspondences between article content and CM questions (shown in Table 6) were also used to guide the evaluation of generated CM questions.
   - For generated DR questions, we evaluated whether the generated questions were meaningful and solvable according to the input three sentence set.

Each generated question needs to satisfy the above two aspects to get the "positive" judgment, questions that failed at either aspect would be judged as "negative" results. As all questions were considered solvable (answerable) in the RACE dataset, the generated un-solvable (unanswerable) question was also evaluated as a "negative" result. All generated questions were manually and individually checked to see whether each was positive or negative. The positive rate was used to evaluate the usefulness of generated questions. The evaluation results were confirmed by two researchers in natural language processing.

5.2.2 Detailed evaluation of grammatical, make sense and solvable

In order to obtain an extensive sight about the quality of generated questions, we conducted a detailed evaluation for generated questions through three aspects of "Grammatical", "Make Sense", and "Solvable". Each of the generated questions was evaluated according to the following generally-used criteria [15, 18, 19].

- **Grammatical:** regardless of the given article/sentences, whether the description of the generated question is correct in English grammar. Evaluators were asked to give a rating from a 1-5 scale to each question's grammatical correctness (5 for the best, 1 for the

worst.).
- **Make Sense:** whether the content of the generated question is in common sense and makes sense to the input article/sentences. The rating on this aspect was also from a 1-5 scale, same as the "Grammatical".
- **Solvable:** whether the generated question is solvable (answerable) according to the given article/sentences. Rating of the "Solvable" was from a 1-3 scale (3 for solvable, 1 for unsolvable).

We asked three English teachers to give their rating scores of the above three aspects for each generated question and made the evaluation results.

5.3 Comparative Method

Overall, the proposed method has two modules in the procedure: the classification and the generation. In order to evaluate the effect of the proposed method that has the classification module and used two models to generate CM and DR questions separately, we chose the method that tried to generate CM and DR questions without the classification step as the comparison. A previously proposed method [26] was used for comparison with the currently proposed two-model method. The procedure of the previous method resembled that of the proposed method, but the previous method treated CM and DR questions together through a single-model without classification. The single-model method extracted the question-related sentences for both two types of questions and prepared the Sentences-Question pairs, which were then used to train the single-model of the question generation. The training set and the test set of the single-model method were the same as those of the two-model method. After the training step, randomly selected sentences set from the test data was input to the single-model method to generate questions. The generated questions were manually classified into CM and DR questions for comparison with the two-model method's generation results. We generated and evaluated each 100 questions from the proposed two-model method and the comparative single-model method.

Table 9. Evaluation of generated 100 CM questions.

Method	# of Pos. CM	# of Neg. CM	Total
Proposed	88	12	100
Comparative	33	67	100

Table 10. Evaluation of generated 100 DR questions.

Method	# of Pos. DR	# of Neg. DR	Total
Proposed	49	51	100
Comparative	24	76	100

5.4 Evaluation Results

5.4.1 Evaluation result of usefulness

Through the evaluation of all generated results, we could find that the questions gen-

erated from the CM model were all CM-like questions and those from the DR model were all DR-like questions.

Table 9 displays the comparison of evaluation results of CM questions generated by each method. Along with a higher positive rate of generation results, the proposed method obtained 88 positive questions, while the comparative method obtained 33 positive questions.

Table 10 displays the comparison of DR questions generated by both methods. Along with a higher positive rate of the generation results, the proposed method obtained 49 positive questions, while the comparative method obtained 24 positive questions.

The positive rate of the generated two types of questions increased by 40% on average (= (88% − 33% + 49% − 24%)/2).

5.4.2 Evaluation result of grammatical, make sense and solvable

Table 11 shows average rating scores of the detailed evaluation of Grammatical, Make Sense, and Solvable for each 100 generated CM questions from the proposed and comparative method. The average rating scores of Grammatical, Make Sense, and Solvable of the proposed method is 4.74, 4.58, and 2.29, compared to 4.11, 3.83, and 1.7 of the comparative method.

Table 11. Average rating scores of grammatical, make sense and solvable of generated 100 CM questions. (Two-tailed *t*-test was applied for the proposed method to the comparative method for each evaluation aspect ($p < 0.01$).)**

Method	Grammatical	Make Sense	Solvable
Proposed	4.74**	4.58**	2.29**
Comparative	4.11	3.83	1.7

Table 12. Average rating scores of grammatical, make sense and solvable of generated 100 DR questions. (Two-tailed *t*-test was applied for the proposed method to the comparative method for each evaluation aspect ($p < 0.01$).)**

Method	Grammatical	Make Sense	Solvable
Proposed	4.21	4.05**	2.05**
Comparative	4.16	3.77	1.48

Table 12 shows average rating scores of the detailed evaluation of 100 generated DR questions from the proposed and comparative method. The average rating scores of Grammatical, Make Sense, and Solvable of the proposed method is 4.21, 4.05, and 2.05, compared to 4.16, 3.77, and 1.48 of the comparative method.

Through the results, we could find that the proposed method made improvements at all three aspects of Grammatical, Make Sense, and Solvable for both CM and DR questions' generation. Two-tailed t-tests between the proposed method and the comparative method also showed that the significance of improvement was considerable at the Grammatical, Make Sense, and Solvable for generated CM questions. For generated DR questions, two-tailed *t*-tests showed that the improvement of the Make Sense and Solvable of the proposed method was significant compared to the comparative method.

6. DISCUSSION

6.1 Results of CM Question Generation

Table 13 shows the example of generated questions from the proposed CM method and the comparative single-model method when input the same article. For the proposed CM method, the generated question was "what is the main idea of the passage?" The input article has the sentence conclusively expressed that writer's viewpoint, so a corresponding CM question asking about the "main idea" could be positive. The writer of the article shared his experience and thinking processes prior to the conclusion, so readers needed to read the entire article to check which sentence could reflect the main idea.

Table 13. Example of CM question's generation through proposed and comparative methods.

Input Article:
Several years ago, I worked in an office which had a locust tree growing outside the window. It had grown into a tall tree and must have been there for a long time. The tree blocked the view and made the office seem dull, unfortunately this happened to be the room assigned to me and I was depressed by it for quite some time. When the first spring came, everything was green except the locust tree. Secretly, I was very happy because I had thought it had died. One morning when I opened the window for fresh air, I unexpectedly smelt a familiar sweet scent floating into my office but I couldn't name it. Suddenly, I realized it was the locust tree! The tree that I thought dead was blossoming in full glory. From then on, I usually came to the office very early to see dew forming on the locust tree blossoms and every year was eager to see spring again. During summer, the green leaves provided shade protecting me from the harsh sunlight, but allowing enough light in to make it pleasant. In autumn, the leaves turned into many different colors. Its beauty touched my soul. Many times, I thought to take photos but never did. Then I had to leave in a hurry. Later in life it became a great regret that I had not done so. Actually, many times in our lives, we think we own something therefore we don't cherish it. We don't feel regretful until we lose it one day! Moreover, sometimes, we have to accept the things we don't want and need to discover the hidden beauty to find unexpected joy!
Proposed: (Positive)
What is the main idea of the passage?
Comparative: (Negative)
How did the writer feel when he first saw the teacher's first little girl?

In contrast, the comparative method often failed to generate positive CM questions. For the comparative method, the generated question was "how did the writer feel when he first saw the teacher's first little girl?" The former part of the generated question ("how did the writer feel") is partially similar to the characteristic of CM questions and corresponds with the content of the input article. However, the latter part ("when he first saw the teacher's first little girl?") does not correspond with the input article, and, indeed, it describes a specific condition for answering that frequently appears in DR questions. This comparative method's generated question did not make sense to the input article, so it was

recognized as negative. Because the comparative method did not classify CM and DR questions but trained a single-model for question generation, its generated questions tended to include the features of both CM and DR questions.

Meanwhile, the proposed method generated some new CM questions that did not appear in the RACE training set (Table 14). Since the proposed method was trained to learn the relations between reading articles and CM questions, it could generate CM questions with the appropriate description to the article content.

In brief, these results verified that the proposed two-model method, which appropriately classified CM and DR questions, can separately and effectively generate more accurate CM questions.

6.2 Results of DR Question Generation

Table 15 shows the example of DR questions' generation from the proposed DR method and the comparative single-model method through the same input sentences. The question generated by the proposed DR generation method was "how many steps are you studying in the letter?". The question should be answered by understanding and inference from several sentences instead of finding the answer directly. This was a predictable benefit of the proposed method that is, integrated information from several sentences to generate questions.

Table 14. Example of generated CM questions not contained in RACE.

Why did the author write this passage? We can tell form the passage that _. According to the text, we can know _. From the passage we can conclude _.

Table 15. Example of DR question's generation through proposed and comparative methods.

Input Sentences: First, you can have a general conversation with her about your career plans. Second, ask for your manager's approval to talk with the training department about any program the company offers. Third, you could say to your boss, "If there is anything you need my help with, I'd love it if you let me know." **Proposed: (Positive)** how many steps are you studying in the letter? **Comparative: (Negative)** the saying "you are just a plan" in the passage means.

In contrast, the question generated by the comparative method was "the saying 'you are just a plan' in the passage means." The comparative method's question could not be answered and did not relate to the input article's sentences. Although the comparative method attempted to extract question-related sentences for the CM and DR questions uniformly with a single model, the distribution of related sentences differed between CM and

DR questions. CM questions involve a larger scope of the reading article, while DR questions relate just to some specific parts of articles. The comparative method would adjust the model to satisfy the different distributions of CM and DR question-related sentences, which decreased the accuracy of extracted sentences related to DR questions. The preparation of Sentences-Question pairs would influence the training effect of the generation model. This result verified that the accuracy of question-related sentences' extraction would influence the generation of DR questions.

For generated DR questions of the proposed method, the positive rate of 49/100 was considerably lower than that of the CM questions (88/100). This might be due to the inadequate selection of the input sentences. For DR questions, we assumed that every three sentences from an article would generate one DR question. However, some of the selected three-sentence sets had a low mutual correlation even from the adjacent position. For example, Table 16 shows a generated negative DR question from the proposed method. It could be found that although the generation model tried to integrate the information from input sentences and to generate the question, due to a lack of correlation among input sentences, the model failed to generate a positive question from the input three sentence set. How to get appropriate and meaningful selections of input sentences is also an important part of DR questions' generation.

In brief, the evaluation results confirmed that the proposed method can generate reading comprehension questions with a higher positive rate than the comparative method.

Table 16. Example of negative output from low correlation input sentences.

Input Sentences:
Perhaps the lesson was difficult, but you tried to understand it. Write these achievements in your diary.
You must be active in learning English and believe that you can do it.
Proposed: (Negative)
What is the main reason why you could fail to understand your diary?

7. CONCLUSION

This paper proposed a method using two different models to generate reading comprehension questions for language learning tests. The reading comprehension-based questions were classified into two types: CM and DR questions. We paired the corresponding related-sentences to each type of question as the training data, and each generation model was designed according to the characteristic of the question type. The proposed method obtained two question generation models for generating the two types of CM and DR questions separately. The generation results of the proposed two-model method were evaluated by comparing with the comparative method, which tried to generate both question types with a single model. The positive rate of CM questions generated from the proposed method was 88% while that of the comparative method was 33%. The positive rate of DR questions generated from the proposed method was 49% while that of the comparative method was 24%. Experimental results proved that the proposed two-model method could generate CM and DR questions more accurately and successfully than the single-model method.

Finally, as a newly proposed task for generating reading comprehension-based questions, many ideas and settings could be applied to improve the future results. For example, a conditional seq2seq model could be used to avoid reducing the training set due to classification, while a generative adversarial network (GAN) model could be used to reinforce the limited training data. How to generate more than one CM questions for a given article is also important for the practical application. We are looking forward to continually improve our method to get better and more effective questions' generation results.

REFERENCES

1. R. R. Day and J. Park, "Developing reading comprehension questions," *Reading in a Foreign Language*, Vol. 17, 2005, pp. 60-73.
2. E. Sumita, F. Sugaya, and S. Yamamoto, "Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions," in *Proceedings of the ACL 2nd Workshop on Building Educational Applications Using NLP*, 2005, pp. 61-68.
3. Y. Susanti, R. Iida, and T. Tokunaga, "Automatic generation of English vocabulary tests," in *Proceedings of the 7th International Conference on Computer Supported Education*, 2015, pp. 77-87.
4. T. Desai, P. Dakle, and D. Moldovan, "Generating questions for reading comprehension using coherence relations," in *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 2018, pp. 1-10.
5. M. Heilman and N. A. Smith, "Good question! statistical ranking for question generation," in *Proceedings of Annual Conference on Human Language Technologies of the North American Chapter*, 2010, pp. 609-617.
6. M. Agarwal, R. Shah, and P. Mannem, "Automatic question generation using discourse cues," in *Proceedings of the 6th ALC Workshop on Innovative Use of NLP for Building Educational Applications*, 2011, pp. 1-9.
7. J. Araki, D. Rajagopal, S. Sankaranarayanan, *et al.*, "Generating questions and multiple-choice answers using semantic analysis of texts," in *Proceedings of the 26th International Conference on Computational Linguistics*, Technical Papers, 2016, pp. 1125-1136.
8. X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 1342-1352.
9. X. Du and C. Cardie, "Identifying where to focus in reading comprehension for neural question generation," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2067-2073.
10. N. Duan, D. Tang, P. Chen, *et al.*, "Question generation for question answering," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 866-874.
11. T. Wang, X. Yuan, and A. Trischler, "A joint model for question answering and question generation," *arXiv Preprint*, 2017, arXiv:1706.01450.
12. D. Tang, N. Duan, T. Qin, *et al.*, "Question answering and question generation as dual tasks," *arXiv Preprint*, 2017, arXiv:1706.02027.

13. D. Golub, P. S. Huang, X. He, *et al.*, "Two-stage synthesis networks for transfer learning in machine comprehension," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 835-844.
14. Y. Zhao, *et al.*, "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3901-3910.
15. X. Du and C. Cardie, "Harvesting paragraph-level question-answer pairs from Wikipedia," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 1907-1917.
16. D. B. Lee, *et al.*, "Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 208-224.
17. S. Muneeswaran, G. Ramakrishnan, and Y.-F. Li, "ParaQG: A system for generating questions and answers from paragraphs," in *Proceedings of the 9th International Joint Conference on Natural Language Processing: System Demonstrations*, 2019, pp. 175-180.
18. G. Chen, J. Yang, and D. Gasevic, "A comparative study on question-worthy sentence selection strategies for educational question generation," in *Proceedings of International Conference on Artificial Intelligence in Education*, 2019, pp. 59-70.
19. B. Wang, X. Wang, T. Tao, *et al.*, "Neural question generation with answer pivot," in *Proceedings of AAAI Conference on Artificial Intelligence*, Vol. 34, 2020, No. 05.
20. S. Varanasi, S. Amin, and G. Neumann, "CopyBERT: A unified approach to question generation with self-attention," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational*, 2020, pp. 25-31.
21. G. Lai, Q. Xie, H. Liu, *et al.*, "Race: Large-scale reading comprehension dataset from examinations," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 785-794.
22. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv Preprint*, 2014, arXiv:1409.0473.
23. A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.
24. X. Du and C. Cardie, "Identifying where to focus in reading comprehension for neural question generation," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2067-2073.
25. J. Devlin, *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of Conference of the North American Chapter of the ACL: Human Language Technologies*, Vol. 1, 2019, pp. 4171-4186.
26. J. Shan, Y. Nishihara, A. Maeda, and R. Yamanishi, "Question generation for reading comprehension of language learning test: A method using Seq2Seq approach with transformer model," in *Proceedings of IEEE International Conference on Technologies and Applications of Artificial Intelligence*, 2019, pp. 1-6.
27. J. Shan, Y. Nishihara, A. Maeda, and R. Yamanishi, "Extraction of question related sentences for reading comprehension tests via attention mechanism," in *Proceedings of IEEE International Conference on Technologies and Applications of Artificial Intelligence*, 2020, pp. 23-28.

Junjie Shan received the B.E. degree from Nankai University, China, in 2012 and received M.E. and Dr. degree from Ritsumeikan University, Japan, in 2018 and 2021. He is now a Senior Researcher (PD) in Ritsumeikan global innovation research organization (R-giro). His research interests include e-learning, edutainment, human-computer interaction, natural language processing, and artificial neural network.

Yoko Nishihara is a Professor at the College of Information Science and Engineering, Ritsumeikan University. She received her each B.E, M.E, and Dr. from Osaka University, Japan, in 2003, 2005, and 2007, respectively. She was a JSPS research fellowship for young scientists (DC1 and PD). She was an Assistant Professor of Faculty of Engineering, the University of Tokyo from 2008 and a Lecturer from 2009. She is interested in human computer interaction and natural language processing. She is a member of IPSJ and JSAI.

Akira Maeda is a Professor at the College of Information Science and Engineering, Ritsumeikan University. He received B.A. and M.A. degrees in Library and Information Science from the University of Library and Information Science in 1995 and 1997 and received the Ph.D. degree in Engineering from Nara Institute of Science and Technology in 2000. His research interests include digital libraries, digital humanities, information retrieval, and multilingual information processing.

Ryosuke Yamanishi received his each B.E, M.E and Ph.D. from Nagoya Institute of Technology, Japan, in 2007, 2009 and 2012, respectively. He joined in College of Information Science and Engineering, Ritsumeikan University as a Research Associate in 2012, a Research Assistant Professor in 2013, an Assistant Professor in 2014 and a Lecturer in 2018. During this period, he visited Laboratory of Computational Intelligence of UBC, Canada as a Visiting Assistant Professor. In 2020, he has joined in Faculty of Informatics, Kansai University as an Associate Professor. He is interested in the content-oriented computational culture and arts like music informatics, comic computing, game informatics and computational gastronomy foods and eating activities. He is a member of IEICE, IPSJ, JSAI, JSKE, ASJ, SAS, ACM and ACL.