A Study on Performance Analysis of Question Generation based on Korean Pretrained Language Model

HongYeon Yu
Honam Research Center
ETRI
Gwangju, Korea
keister@etri.re.kr

Jiwon Yang
Honam Research Center
ETRI
Gwangju, Korea
jiwonyang@etri.re.kr

Seunghun Oh
Honam Research Center
ETRI
Gwangju, Korea
osh93@etri.re.kr

Donghoon Son
Honam Research Center
ETRI
Gwangju, Korea
dhson78@etri.re.kr

Aram Lee
Honam Research Center
ETRI
Gwangju, Korea
al246@etri.re.kr

Jeongeun Kim
Honam Research Center
ETRI
Gwangju, Korea
j.kim@etri.re.kr

***Abstract*— This paper presents the results of a performance analysis of the question generation task based on a Korean pretrained language model for assessing elementary school students' reading comprehension using literary texts. The Korean pretrained language model is based on SKT-KoBart and employs transfer learning using literary texts of narrative and expository structures, along with the KorQuAd dataset for question-answering tasks. Through the extraction of factual and inferential assessment items, BLEU scores are measured based on the type of literary text, and the performance analysis results of the downstream task of question generation are presented.**

Keywords*— *KoBart language model, question generation, reading comprehension, narrative texts, expository texts, assessment question

I. INTRODUCTION

Elementary school students' reading comprehension education plays a crucial role in developing and enhancing reading habits and reading comprehension skills. Reading has a profound impact on personal growth in various aspects, including expanding knowledge, enhancing language abilities, and fostering creativity. It serves as an essential skill that individuals must possess, as it can form the foundation for effective communication in the digital age.

For reading education, elementary school teachers measuring reading comprehension is an essential factor in assessing whether students have the ability to understand documents and grasp their content. During lessons, teachers can assess students' reading comprehension through comprehension questions and reading skill tests.

Traditionally, such reading comprehension assessment questions have been manually generated by skilled teachers directly from given the literary text. However, as new texts are integrated into the curriculum, teachers often face the challenge of generating new reading comprehension questions for each new literary text. Furthermore, they spend a considerable amount of time constantly generating questions that are tailored to the specific level of a group of students or individual students and that can assess various aspects of reading comprehension. Therefore, to enhance the reading comprehension education environment and alleviate the burden on elementary school teachers in generating assessment questions, there is a need for technology that can automatically generate reading comprehension assessment questions for a given literary text. Especially due to the impact of the COVID-19 pandemic, the necessity for effective online reading education platforms has been emphasized. For such online platform-based education, it is crucial to have the capability of automatically generating a variety of reading comprehension assessment questions related to literary texts, which can be provided to students.

Recently, there has been a growing interest in online reading education service technologies aimed at enhancing elementary students' literacy skills in Korea. In particular, there is a demand for online reading education service technologies that empower students to self-directed enhance their literacy skills. However, for the implementation of such self-directed reading education service systems, it is crucial to have the capability of the system to infer and present questions that can assess students' reading comprehension from various types of literary texts.

However, most existing question generation systems tend to focus on creating factual questions that can be answered solely based on explicitly stated information in the text. Therefore, it is crucial to perform on the ability to answer the questions that require students to use their personal background knowledge or make inferences from the text, especially for elementary school students.

Therefore, in this paper, we present the results of analyzing the performance of existing pretrained language models for the development of an online reading education service based on the generation of assessment questions that can measure factual and inferential reading comprehension according to different literary styles of texts.

The structure of this paper is as follows. In the next chapter, we explain the pretrained models and learning structure related to question generation. In chapter 3, we present the performance analysis results of question generation based on dataset composition. and then in chapter 4, we draw conclusions of this study.

II. RELATED WORKS

A. Pretrained Language Model

A pretrained language model is an artificial intelligence model used for natural language processing tasks by utilizing

large-scale text data. This model is trained on a substantial amount of text data beforehand and is then used to perform various natural language processing tasks based on that training. Pretrained language models enhance language understanding and generation abilities, serving as a foundation for transfer learning methods applicable to diverse natural language processing tasks such as object recognition, translation, document summarization, question-answering, and more.

Some examples of pretrained language models include OpenAI's GPT (Generative Pre-trained Transformer)[1], Google's BERT (Bidirectional Encoder Representations from Transformers)[2], and Facebook's BART (Bidirectional and Auto-Regressive Transformers)[3] These models support fine-tuning for transfer learning, making them applicable to various natural language processing tasks.

The GPT language model is specialized in text generation and primarily possesses the ability to understand context in a unidirectional manner. On the other hand, the BERT language model is capable of understanding context bidirectionally and is mainly applied to various natural language processing tasks through fine-tuning.

The BART model employs an encoder-decoder structure, utilizing a bidirectional transformer encoder similar to BERT for the encoder and a unidirectional transformer decoder similar to GPT for the decoder. This Seq2Seq structure allows BART to not only comprehend natural language but also generate it. Furthermore, BART learns data with added noise using a denoising function to extract word meanings, effectively removing noise from the text.

During the pretraining process of BART, the encoder is fed with variously noised versions of the original text while the decoder is given the original text as input. The encoder conveys information along with the noised text to the decoder, which learns to predict the original text. This allows the pretrained model to learn how to reconstruct the original text from the corrupted version.

BART employs various methods to transform the original text, including token masking(randomly masking tokens), sentence permutation(splitting and shuffling sentences), document rotation(randomly selecting tokens to shift the document's start), token deletion(removing tokens while obfuscating their positions), and infilling (deleting text spans). Among these, the infilling method demonstrated the best performance.

B. *Question Generation Task Model*

This paper aims to perform a performance analysis of pre-trained language models for the purpose of question generation to assess reading comprehension among Korean elementary school students. To achieve this, the study employs the SKT-AI developed KoBart[4], a pre-trained Korean language model, to carry out the task of question generation. KoBart is trained using over 40GB of Korean text data and applies the same infilling noise function as the BART pre-trained model, showing robust performance in Korean language-based natural language processing.

III. PERFORMANCE ANALYSIS

A. *DataSets*

This paper focuses on generating questions to measure the reading comprehension of Korean elementary school students. A dataset was collected consisting of 1,139 texts with explanatory and narrative structures, as shown in Figure 1, from literary texts. For each text, a dataset of 7,342 pairs of factual and inferential questions and answers was gathered. Additionally, for Korean machine reading comprehension, combined the KorQuAD 1.0[5] dataset consisting of 10,645 texts with 66,181 pairs, as introduced in a previous study [6], to train the question generation model.

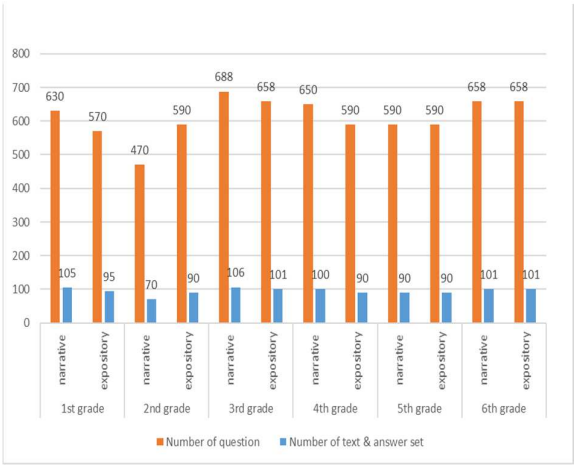

Figure 1. Collected datasets by literary text types

B. *Training Question Generation Model*

Question generation is a task that involves generating questions that can lead to deriving answers when given a text and a question-answer pair that allows for inference.

To learn the characteristics of different literary text styles in reading comprehension questions using the KoBART pre-trained model, in this paper combines the KorQuAD dataset with narrative and expository text style datasets. The KoBART's transfer learning dataset is utilized by inserting tokens to distinguish paragraph boundaries between the context and answer, creating inputs structured as {context, [SEP], answer}, {question} pairs. The model training parameters were set according to the configuration in the existing paper [7]. Gradient clipping was applied to prevent gradient exploding, and the training used the AdamW optimizer and a cosine warm-up scheduler to adjust the initial learning rate.

TABLE I. EXPERIMENT RESULTS

Model(dataset)		BLEU-1	BLEU-2	BLEU-3	BLEU-4
KorQuAd	**FQ**	35.99	22.52	17.11	15.19
	IQ	30.66	17.00	12.95	11.75
KorQuAd Narrative	**FQ**	34.95	20.88	15.32	13.18
	IQ	29.87	16.54	12.48	11.16
KorQuAd Expository	**FQ**	42.59	29.31	23.67	20.59
	IQ	32.22	18.24	14.27	12.74
KorQuAd Narrative Expository	**FQ**	36.03	22.48	17.25	15.35
	IQ	29.66	16.57	12.86	11.86

C. *Experimental Results*

The test data for evaluating the performance of the trained question generation model consists of 7,342 pairs of collected datasets, each containing narrative texts and expository texts. From these pairs, each of 300 pairs of factual questions (FQ)

and inferential questions (IQ) are extracted and used for assessment. The performance analysis will compare three models: a question generation model trained solely on the KorQuAD dataset, a question generation model trained on the KorQuAD dataset as well as narrative and expository datasets separately, and a question generation model trained on the integrated dataset of KorQuAD, narrative and expository texts.

TABLE II. QUESTION GENERATION RESULTS(EXAMPLE)

English (context &answer & teacher-generated question)	**Context**: <…> However, even if we cut them, we don't feel pain. But why are fingernails and toenails not soft and instead hard? The reason is that they became hard to protect our hands and feet. <…> We can tell if a person is healthy or not by looking at their fingernails. Healthy fingernails are glossy and show a white crescent shape at the root. If fingernails are bent or have wrinkles, it's a sign that the person's health is not good. That's why fingernails are called the mirror of health. **Answer(factual):** To protect hands and feet **Teacher-generated question(factual):** Why have fingernails and toenails become hard? **Answer(inferential):** You can tell a person's health condition by looking at their fingernails. **Teacher-generated question(inferential):** Why are fingernails called the mirror of health?
English (model-generated question)	**Model-generated question(factual):** Why have fingernails and toenails become hard? **Model-generated question(inferential):** What can you tell from nails that appear glossy and show a white half-moon shape at the base?
Korean (context &answer & teacher-generated question)	**Context:** <…> 그러나 그것을 잘라 내도 우리는 아픔을 느끼지 못하지요. 그런데 왜 손톱과 발톱은 말랑말랑하지 않고 딱딱할까요? 그 이유는 손과 발을 보호하기 위해 딱딱하게 변했기 때문이랍니다. <…> 손톱을 보면 그 사람이 건강한지 그렇지 않는지 알 수 있어요. 건강 한 사람의 손톱은 윤기가 나고 뿌리 부분에 하얀 반달 모양이 보여요. 손톱이 찌그러지거나 주름이 잡혀 있으면 건강이 좋지 않다는 신호 예요. 그래서 손톱을 건강의 거울이라고 한답니다. **Answer(factual):** 손과 발을 보호하기 위해서 **Teacher-generated question(factual):** 손톱과 발톱이 딱딱하게 변한 까닭은 무엇인가요? **Answer(inferential):** 손톱을 보면 그 사람의 건강 상태를 알 수 있기 때문이다. **Teacher-generated question(inferential):** 손톱을 건강의 거울이라고 하는 까닭은 무엇일까요?
Korean (model-generated question)	**Model-generated question(factual):** 손톱과 발톱이 딱딱하게 변한 이유는? **Model-generated question(inferential):** 손톱이 윤기가 나고 뿌리 부분에 하얀 반달 모양이 보이는 것을 보고 무엇을 알 수 있나요?

The performance analysis employs the BLEU[7] (Bilingual Evaluation Understudy) score as the evaluation metric. The measurement of the BLEU score utilizes n-grams with equal weights to quantify how similar the questions generated by the model are to those generated by teachers, expressed as a ratio.

The experimental results, as shown in Table 1 and Table 2, demonstrated that the performance of factual question generation was superior to inferential question generation. Additionally, the results indicated that factual question generation performed better in expository text types compared to narrative text types. Furthermore, the experiment illustrated that question generation performance varied based on the text type. This ultimately highlights the need to create question generation task models based on the literary text types.

IV. CONCLUTION

In this paper, the analysis of question generation performance using the KoBart pre-trained language model was conducted to assess the reading comprehension abilities of elementary school students. The performance analysis revealed that the KoBart pre-trained Korean language model is well-suited for the question generation task. However, there is an issue with generating inaccurate inferential questions, as seen in "answer (inferential)" in Table 2, when sentences not included in the context are used. Therefore, Research focused on developing models that can generate accurate inferential questions by leveraging such background knowledge in conjunction with given text is required.

ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government(23ZK1100, Honam region regional industry-based ICT convergence technology advancement support project)

REFERENCES

[1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Imporving Language Understanding by Generative Pre-Training," Technical report, OpenAI, 2018.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, Minneapolis, Minnesota, pp.4171-4186, Jun. 2019.

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension," ACL, Online, pp.7871-7880, Jul. 2020..

[4] https://github.com/SKT-AI/KOBART(accessed Aug. 02. 2023)

[5] Seungyoung Lim, Myungji Kim, and Jooyoul Lee., "KorQuAD1.0: Korean QA Dataset for Machine Reading Comprehension," 2019. arXiv:21909.07005.

[6] Gyu-Min Park, Seong-Eun Hong, Seong-Bae Park, "Post-Training with Interrogative Sentences for Enhancing BART-based Korean Question Generator," In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, vol. 2., pp.202-209, Nov. 2022.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318. July 2022.