mathematics

MDPI

Article

AI Student: A Machine Reading Comprehension System for the Korean College Scholastic Ability Test

Gyeongmin Kim [1], Soomin Lee [1], Chanjun Park [1] and Jaechoon Jo [2,*]

[1] Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea; totoro4007@korea.ac.kr (G.K.); skyop27@korea.ac.kr (S.L.); bcj1210@korea.ac.kr (C.P.)
[2] Division of Computer Engineering, Hanshin University, Osan 18101, Korea
* Correspondence: jaechoon@hs.ac.kr

Abstract: Machine reading comprehension is a question answering mechanism in which a machine reads, understands, and answers questions from a given text. These reasoning skills can be sufficiently grafted into the Korean College Scholastic Ability Test (CSAT) to bring about new scientific and educational advances. In this paper, we propose a novel Korean CSAT Question and Answering (KCQA) model and effectively utilize four easy data augmentation strategies with round trip translation to augment the insufficient data in the training dataset. To evaluate the effectiveness of KCQA, 30 students appeared for the test under conditions identical to the proposed model. Our qualitative and quantitative analysis along with experimental results revealed that KCQA achieved better performance than humans with a higher F1 score of 3.86.

Keywords: academic reading skills; Korean College Scholastic Ability Test; Korean CSAT question and answering; machine reading comprehension

MSC: 68T50

Citation: Kim, G.; Lee, S.; Park, C.; Jo, J. AI Student: A Machine Reading Comprehension System for the Korean College Scholastic Ability Test. *Mathematics* **2022**, *10*, 1486. https://doi.org/10.3390/math10091486

Academic Editors: Heui Seok Lim, Sanghyuk Lee, Yeongwook Yang and Imatitikua Aiyanyo

Received: 7 April 2022
Accepted: 27 April 2022
Published: 29 April 2022

1. Introduction

Machine Reading Comprehension (MRC) aims to teach machines to read and answer questions after understanding the given text passages, which is a fundamental goal of natural language understanding [1,2]. This sequential process of the MRC model resembles a human who solves reading comprehension questions in the Korean language subject on the College Scholastic Ability Test (CSAT). The CSAT is a standardized test in Korea. It is designed to evaluate the scholastic ability of students required for a college education. The CSAT focuses on high-order thinking skills based on the prior knowledge of each student in the subject and how they answer the exam. For the Korean language subject in the CSAT, the MRC model should evaluate the effectiveness of whether the model can classify answers as True (*T*) or False (*F*) given the passage, which is approximately 1700 characters and 15 multiple-choice questions on average. However, research on analyzing the CSAT using Natural Language Processing (NLP) remains limited.

Over the past few years, there has been a significant change in the learning of NLP models. Pre-trained Language Models (PLMs) [3–5], which are pre-trained with large-scale text corpus through unsupervised objectives similar to masked language modeling such as BERT, GPT, RoBERTa, and T5, are adapted to various downstream tasks by training and fine-tuning parameters with task-specific objective functions [3–7]. Recently, there have been several attempts to adopt PLM-based MRC models in other domains, the MRC has domain application cases such as biomedicine and cybersecurity [8–11]. However, to the best of our knowledge, there has been limited research on exploiting the MRC technique in the domain of the Korean language reading in the CSAT. It is crucial to utilize PLMs to convert given passages and questions into a classification problem, and the intrinsic reasoning skills of the PLMs to test exams depend on their prior knowledge, similar to humans.

Therefore, in this paper, we propose *AI student*, *a novel* ***K****orean* ***C****SAT* ***Q****uestion and* ***A****nswering* ***(KCQA)*** model that assesses the scholastic reading ability in the Korean language CSAT. KCQA evaluates the effectiveness of language models with given questions and passages under identical conditions experienced by students. In addition, Round-Trip Translation (RTT), which is a method mainly utilized to alleviate insufficient training data in neural machine translation, is applied and simultaneously obtains practical knowledge by exploiting the Easy Data Augmentation (EDA) method to effectively augment CSAT corpus with wordnet, which includes synonym and antonym relationships between words [12–15]. This approach suggests the possibility of examining the contribution of prior knowledge of certain subjects to the understanding of the related passage by adopting language models, which was impossible to quantify in the field of education.

The contributions are summarized as follows:

- We obtained insights into the meaningful effects of our KCQA models, which assessed scholastic reading ability.
- To this end, we demonstrated the effectiveness of using various Korean and multilingual language models with four data augmentation strategies for practical learning to alleviate insufficient training data problems.
- For human performance, we employed 30 test students preparing for the CSAT with conditions identical to the PLMs. The results proved the superior performance of the proposed KCQA.
- We comprehensively conducted qualitative and quantitative analyses based on this by deriving concrete experimental results in both aspects of educational assessment and deep learning.

2. Background

The CSAT, which is a high-stakes assessment that serves as a decisive factor for college admission, has been developed and administered by the Korea Institute for Curriculum & Evaluation (https://www.kice.re.kr/ (accessed on 5 April 2022)) since its introduction in 1993 as the most important standardized tool. It is a compulsory exam for students who are aiming to enter college after 12 years of regular education courses [16,17]. Each subject consists of Korean language subjects subdivided into reading and literature, as well as English, Mathematics, Social Studies, Science, Vocational Studies, Foreign Languages, and Chinese Classics.

Regardless of language, "Reading" is highly thought of in education; this is especially true for Korean language education. The reading section in the CSAT aims at "cultivating the ability to accurately understand a certain passage". When evaluating reading comprehension skills, the prior knowledge of each student plays a crucial role in understanding the grammar and literary knowledge for a given passage and question [18]. In this process, reading requires students to classify the given question as *T* or *F* for the given passage.

3. Related Works

3.1. Pre-Trained Language Models

Many recent studies based on transformer models consisted of a pre-training and fine-tuning stage and achieved good performances in various downstream tasks, such as named entity recognition, question answering, text generation, and other NLP tasks. It is common practice in the NLP community to fine-tune various tasks instead of learning models from scratch [19–24]. BERT showed superiority over human performance in the Stanford question answering dataset, the most representative question answering dataset [25]. PLMs apply various techniques, such as masked language models, next sentence prediction, and Replaced Token Detection (RTD) to provide contextual information for better quality context-sensitive information within the passage for each downstream task. Multilingual-BERT (mBERT) for conducting research in Korean includes various forms of linguistic information, but it has limitations in that Korean-specific data are not sufficient.

Recently, various Korean PLMs have been released in the Korean research community. They achieved outstanding performances that were better than those of multilingual models with large-scale Korean corpora and vocabulary exploited by reflecting the philological characteristics of the Korean language, which is an underlying factor for the performance of PLMs. KoBERT (https://github.com/SKTBrain/KoBERT (accessed on 5 April 2022)) is pre-trained with 5 million Korean sentences and optimized for the Korean language. KoELECTRA is pre-trained with 34 GB worth of Korean sentences from news articles, wikidata, and the National Institute of Korean Language, which is an institution that establishes the norm for Korean linguistics (https://corpus.korean.go.kr/ (accessed on 5 April 2022)). The model adopted RTD to change a certain ratio of tokens into masking tokens and let the generator generate suitable tokens to fill in the masking tokens [26]. In this process, the discriminator trains by determining which token has been replaced based on the output of the generator. Moreover, KcELECTRA is an ELECTRA-based model, but there is a significant difference in the nature and scale of the training data. It is optimized for news comments and movie reviews where tokens have colloquial features, including numerous new words and informal expressions such as typos. Approximately 17 GB of news comments were used for pre-training.

3.2. Adoption of Deep Learning in CSAT

A study that extracted keywords, which accounts for 11% of the Korean language section CSAT to visualize as a word cloud, and to analyze the language network with a term-document matrix method was conducted [27]. However, considering only the frequency of the keywords, describing the contextualized feature representation and understanding the meaning of linguistic representations could not be achieved. Furthermore, a study was conducted on the English section of CSAT, which compared vocabulary complexity using a vocab profile with the reading comprehension of SAT2 in ETS, and grammatical complexity using the L2 syntactic complex analyzer [28]. However, this approach also did not graft to the PLMs, which mainly dealt with in NLP. Although the various subjects of the CSAT were used to evaluate the ability the understanding of a given passage, no research has been conducted.

3.3. Data Augmentation Strategy

Enhancing text data to overcome the lack of training corpus in a relatively low-resource language has been suggested in various research studies as an effective way to increase the contextual understanding of the language model [29–31]. For data augmentation, the original training corpus must be transformed into a suitable form within a range that does not damage the original meaning. There were several challenges in applying data augmentation to NLP, as compared to other fields; however, this presents an opportunity to introduce the EDA to augment text-based datasets effectively. EDA is a technique for augmenting given sentences according to four categories: Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD). SR is a method in which specific tokens are replaced with synonym tokens referred to as wordnet. RI is a method of inserting a random token at a specific location. RS is a method of swapping two random token positions. Finally, RD is a method of deleting a random token from a sentence. In the process of data augmentation for text classification, even if the size of the dataset has been increased effectively, the issue of the existing correct answer being contaminated should not exist. For example, when the number of sentences with labels classified as *T* is augmented, it becomes a failure if many of them are classified as *F*. However, EDA prevents the label from being reversed from *T* to *F* and vice versa.

4. Enhanced Dataset Strategies for CSAT

4.1. Reading Form in CSAT

Reading consists of philosophy, social studies, science, technology, and art, and convergent passages deal with two different fields: philosophy and science. Each field includes

three or four passages, and each passage includes four or five questions. The question includes five answer candidates and must be classified whether each answer candidate is *T* or *F*. The test format is demonstrated in Figure 1. Passages and questions in the actual example of CSAT are shown in Table 1.

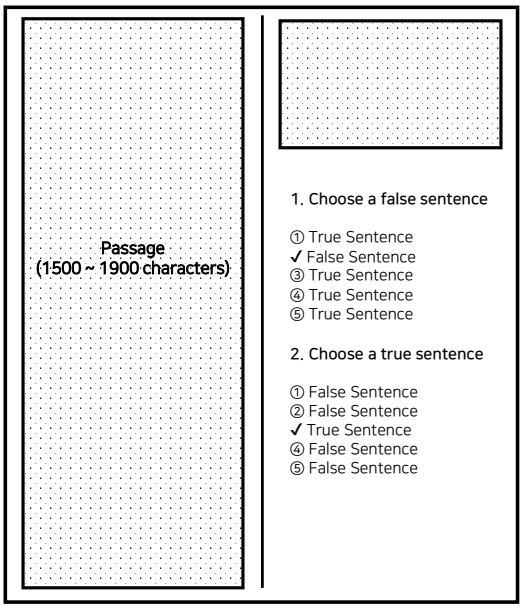

Figure 1. Description of the Korean CSAT. To solve the given questions, a student must be able to select which question is *T*/*F* out of the multiple-choice questions.

Table 1. Example of passage and question in the Korean CSAT.

Passage	a semiconductor substrate board, which is similar to the process of making an engraving. Just as countless engravings can be made from an original plate on paper, in the case of photolithography, a single plate called a mask is made, and then the pattern of the exact same shape is repeatedly copied on the substrate using a laser to make a large number of patterns. Compared to making the original plate using an engraving knife, in the case of photolithography, the size of the mask pattern is very small, so it is made using a laser.
Question	The size of the pattern engraved on the mask is smaller than the size of the pattern created on the substrate board.

To determine the correct answer corresponding to the question, students must be capable of understanding not only the superficial information but also the semantic and contextual information. In Table 1, a student pays attention to the superficial information of *the size of the mask pattern is very small*, and thus can easily misclassify the given question as *T*. However, if a student pays attention to the figurative semantic information implied in *Just as countless engravings can be made from an original plate on paper*, the student can correctly classify the given question as *F*.

4.2. Reading Section of Korean Language

A total of 285 passages were used, and each passage was given an index according to the admission year and area of the reading section in the CSAT. The index consisted of an academic area that focused on philosophy, social studies, art, science, technology, convergent, and language, which have not appeared in the CSAT since 2012. In Table 2, the average number of questions in augmented train and test datasets for each academic area are shown in detail. The passages corresponding to each category are indicated in parentheses. For example, the 58 passages in the field of philosophy have on average 120.86 sentences on the augmented training dataset, and the average number of questions that are *T* or *F* is 6.16 and 5.71 in the test datasets, respectively.

Table 2. Main categories and their average number of sentences (in parentheses, marked #) in reading.

Categories (#)	Augmented Train Dataset	Test Dataset	
		(True, T)	(False, F)
philosophy (58)	120.86	6.16	5.71
social studies (55)	115.54	6.53	5.33
science (53)	120.36	7.36	6.86
technology (43)	119.00	6.95	6.31
art (45)	115.69	6.11	5.71
language (23)	109.00	4.83	4.26
convolution (8)	203.78	10.25	9.75

This is closely related to a downstream task in NLP that recent claims that the MRC model truly understands a given passage. Extracting the span corresponding to the answer based on the given text tended to have an increasingly degrading performance in the presence of discouragements unrelated to the passage [32]. It is considered that models rely on superficial information and do not understand the passage in context. Our CSAT corpus can be classified correctly only when the models fully understand the passages and questions. Therefore, by capturing the CSAT in the form of a dataset to which NLP can be applied, it is possible to present methods to solve problems for the MRC task.

4.3. Strategy for CSAT Corpus Augmentation

We used the passages and their corresponding questions from the CSAT as training data. Despite the fact that the average number of sentences increased to 30.76 after the 2015 revised curriculum, the amount of CSAT corpus is insufficient for the practical learning process. Furthermore, since the test aims to evaluate the reading ability of the student in a limited time, unsophisticatedly extending the length of the passages is not an appropriate solution. To overcome this, we applied the four training data augmentation strategies proposed by EDA [14], which are effective even when the size of data is relatively small. We applied SR, RI, RS, and RD by 10% ratio, independently, in order to augment the sentences by five times.

Wordnet was referred to replace tokens included in a given sentence with synonym tokens or to insert random tokens in the middle of sentences. Considering that wordnet is not universally available publicly, unlike English, RTT is an effective method that splits the training data to each sentence before carrying out Ko–En translation and then En–Ko translation back again. The overall procedure of the data preprocessing is shown in Figure 2.

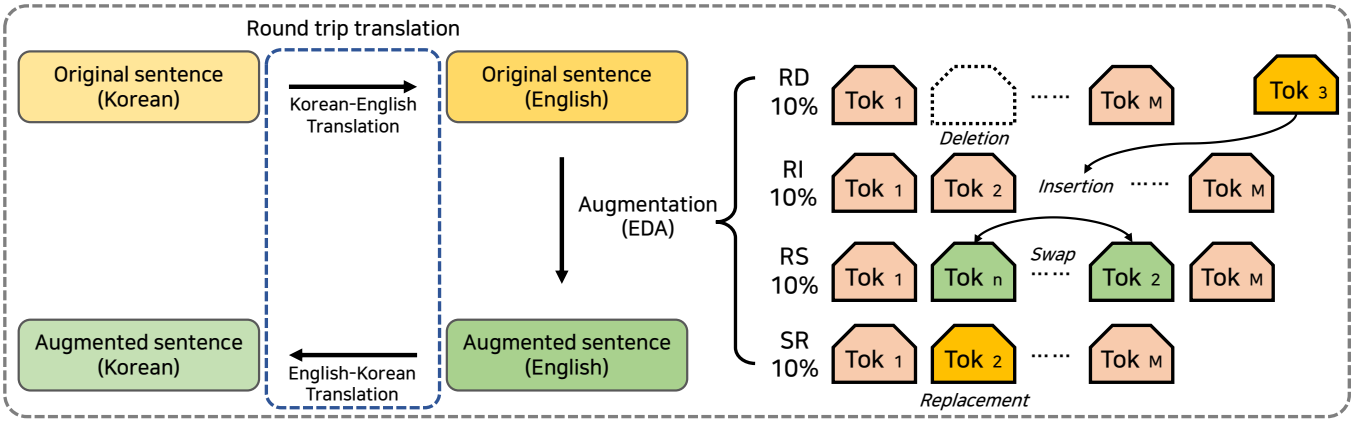

Figure 2. An overview of the RTT process with four EDA strategies, which is applied as SR, RI, RS, and RD by 10%, independently.

4.4. KCQA System

We fine-tuned our KCQA model by leveraging the Korean PLMs with the optimization process. Figure 3 illustrates the overall model architecture. From the entire training corpus, we fed augmented input sentences into the model. The model is trained to correctly classify (CLS) tokens as the final prediction value of the input. In preparation for this process, we randomly composed sentences of multiple passages so that the same number of true-labeled and false-labeled sentences were input into training for each passage. However, to let the model effectively train false-labeled sentences, the sentences with negative expressions were artificially generated based on the passage and were preferentially included as the input. The method was devised from the fact that the most negative expressions in Korean sentences are found in the terminus of the word token, and also for the reason that in the case of the Korean language section of the CSAT, it is helpful to generate negative expressions consistently because the terminus of each question tends to be consistent [33]. Detailed examples are shown in Table 3.

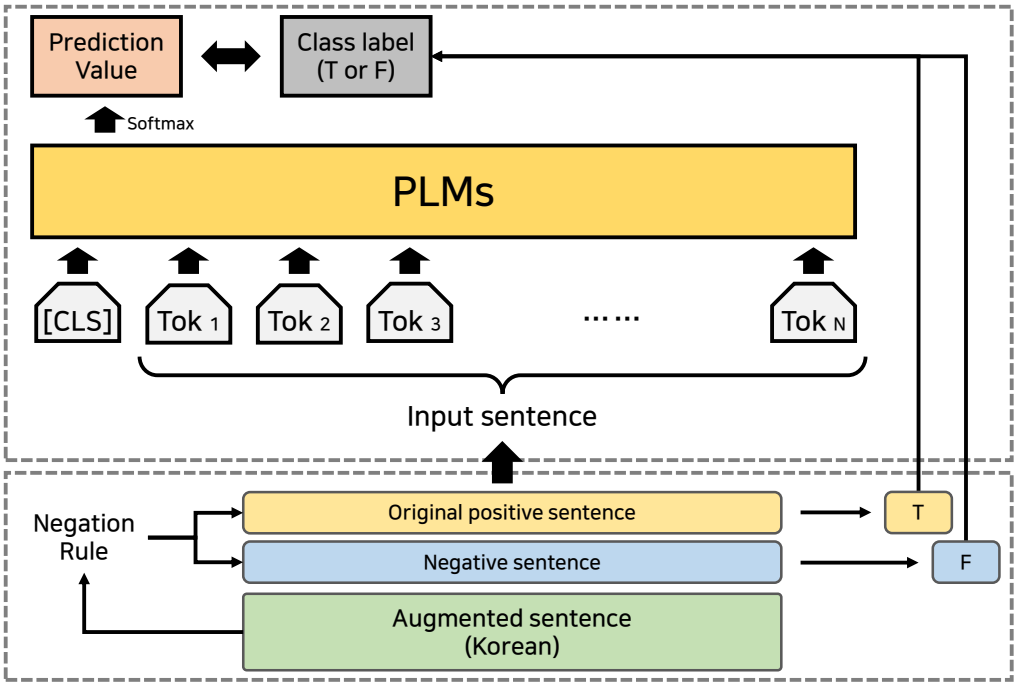

Figure 3. PLMs architecture. In this process, augmented sentences are split into positive and negative sentences using negation rules before being fed into the model. Finally, the prediction value is classified according to whether the answer is *T* or *F*.

Table 3. Examples of sentences according to the negation rules. There are three negative expressions, Verb (V) → don't V, can → can't, and does → doesn't. The underlined words in Korean indicate bold words in English.

Negation Rule	Original True Sentence	Sentence Negation
-이다 (V) → -이 아니다 (don't V)	개인적 동기가 공공성과 상충되는 현상이 나타났던 것<u>이다</u>. (Personal motives **appear** to conflict with public ones.)	개인적 동기가 공공성과 상충되는 현상이 나타났던 것<u>이 아니다</u>. (Personal motives **don't appear** to conflict with public ones.)
-할 수 있다 (can) → -할 수 없다 (can't)	이 교차로 표지가 과정1에 도입되었고 이 표지는 'b' 까지 전달될 수 <u>있다</u>. (This intersecting mark was introduced in 'process 1' and **can** be delivered to 'b')	이 교차로 표지가 과정1에 도입되었고 이 표지는 'b' 까지 전달될 수 <u>없다</u>. (This intersecting mark was introduced in 'process 1' and **can't** be delivered to 'b')
-한다 (does) → -하지 않는다 (doesn't)	아리스토제누스는 귀로 지각된 소리를 근거로 음악의 아름다움을 판단<u>한다</u>. (Aristoxenus **does** determine the beauty of music based on the sound perceived by the ear)	아리스토제누스는 귀로 지각된 소리를 근거로 음악의 아름다움을 판단<u>하지 않는다</u>. (Aristoxenus **doesn't** determine the beauty of music based on the sound perceived by the ear)

5. Experiments

5.1. Dataset

The 285 passages contain training and test datasets. To ensure the quality of the training model, we conducted an evaluation using ten-fold cross-validation, i.e., 10 times for each model, where 90% of the data were provided as training input and 10% as a test to the models. On average, the original training corpus consisted of 24.30 sentences without augmentation and 119.55 sentences with augmentation strategies. A more detailed description of training and testing is shown in Table 4.

Table 4. Dataset details.

Dataset	Data Type	Sentences	Augmented Sentences
CSAT passages (285)	Train	24.3	119.55
	Test	6.58	6.58

5.2. Metrics

To evaluate the effectiveness of our proposed method, we utilize the F1 score, which is a harmonic value of Precision (P) and Recall (R), as a standard indicator to compensate for the weakness caused by using only accuracy for evaluation. The F1 score is calculated by Equation (1).

$$P = \frac{|C|}{|E_p|}, R = \frac{|C|}{|E_r|}, F1 = (\frac{R^{-1} + P^{-1}}{2})^{-1} = 2 \cdot \frac{P \cdot R}{P + R} \quad (1)$$

where E_p represents the set of predicted correct answers, E_r denotes the ground-truth answer collection, and $C = E_p \bigcap E_r$ are the correct answers.

5.3. Experimental Results

In the experiment, we leveraged KoBERT, KoELECTRA, and KCELECTRA as Korean representative models, and mBERT and xlm-RoBERTa as multilingual models for baseline model architecture. For accurate performance evaluation, we set consistently fine-tuning hyperparameters with a batch size of 128, max sequence length of 128, max epochs of 40, learning rate of 1×10^{-5}, and a weight decay of 0.1. In Table 5, we describe our quantitative experimental results. Human performance targeted 30 test students preparing for the CSAT. They were required for the CSAT corpus to be referenced in the identical method as the PLM by reading the passages and determining whether the given information was appropriate or not.

Table 5. Performance of human and various data augmented language models. We utilized five Korean and multilingual models for baseline architecture. As a result, all of the Korean models with data augmentation strategies achieve higher F1 scores.

Models (*Augmented*)	F1 Score	Recall	Precision
Human Performance	58.24	62.31	59.24
KoBERT	58.46 (+0.22)	64.08 (+1.77)	57.24 (−2.00)
KoELECTRA	**61.78 (+3.54)**	67.67 (+5.36)	59.24 (+0.00)
KcELECTRA	**62.10 (+3.86)**	68.22 (+5.91)	59.10 (−0.14)
mBERT	58.08 (−0.16)	62.36 (+0.05)	58.20 (−1.04)
xlm−RoBERTa	48.48 (−9.76)	55.71 (−6.60)	47.63 (−11.61)

In the case of the two multilingual augmented PLMs, mBERT and xlm-RoBERTa performed less favorably than the humans; specifically, xlm-RoBERTa recorded a decrease of −9.76 on the average F1 score. The language models based on ELECTRA architecture pre-trained with large-capacity, formal and informal written styles showed the most effective philological understanding compared to the other models; this was verified through the model performance. In addition, the similar composition of the linguistic system between

the Korean language of CSAT and our KCQA proves that the result implies our models can make correct assessments about questions with the given passages, which shows that educational goals of evaluating the reading comprehension in the Korean language section can be achieved.

6. Analysis

In this section, we provide a detailed analysis of augmenting our KCQA model by describing additional experiments. To further analyze how our data augmented representations influence the model performance, we scrutinized the differences between *P* and *R* of language models trained in augmented and non-augmented strategies. We addressed CSAT by exploiting the same model architectures described above in the corresponding sections.

In Table 6, all of the PLMs with data augmentations enable stable learning as the differences between *P* and *R* is alleviated from 41.14 to 7.33 on average. Based on the *T* values predicted by the model in Equation (1), *P*, which refers to the *T*, and *R*, which the model predicts as *T*, complementarily show better performances when both indicators are higher. For the case where augmentation strategies were not applied, the differences between *P* and *R* were relatively significant because *T* and *F* sentences were predicted as *T* since they were superficially similar to the expression of the given passages. The performance declined after applying augmentation, but it rather indicates a better understanding of the given text since the differences between *P* and *R* decreased after augmentation.

Table 6. Differences between recall and precision, the gaps in our data augmented representations are relatively insignificant.

PLMs	Recall-Precision (*Original*)	Recall-Precision (*Augmented*)
KoBERT	38.23	**6.84**
KoELECTRA	42.50	**8.43**
KcELECTRA	41.44	**9.12**
mBERT	40.54	**4.16**
xlm-RoBERTa	43.01	**8.08**

Thus, we prove its effectiveness by articulating that the differences of both indicators have been mitigated after the augmentation strategy, and confirm that the augmented PLMs can understand the given questions and passages of the CSAT corpus beyond superficial knowledge.

In Figure 4, each field constituting the CSAT corpus was described as a visualized confusion matrix for the data augmented language model performance. Considering that the difference in numerical values represented by each field deepened the understanding of each passage during the augmentation process, it can be interpreted that the required degree of inference and understanding varies for each field.

Furthermore, it can be assumed that philosophy and art areas showed relatively poor performance because those passages required accepting perspectives of specific artists and philosophers. These passages with high abstractness showed minor performance improvements after data augmentation. Table 7 lists the top and bottom ten passages based on the F1 score of each passage. The two fields of philosophy and technology account for 70% of the bottom ten passages with the lowest performances, and in the case of the science field, applicable content was dealt with in detail, but the range of improvement was low. The most probable cause is that domain-specific additional knowledge is considered regardless of the degree of reasoning required by the passage itself.

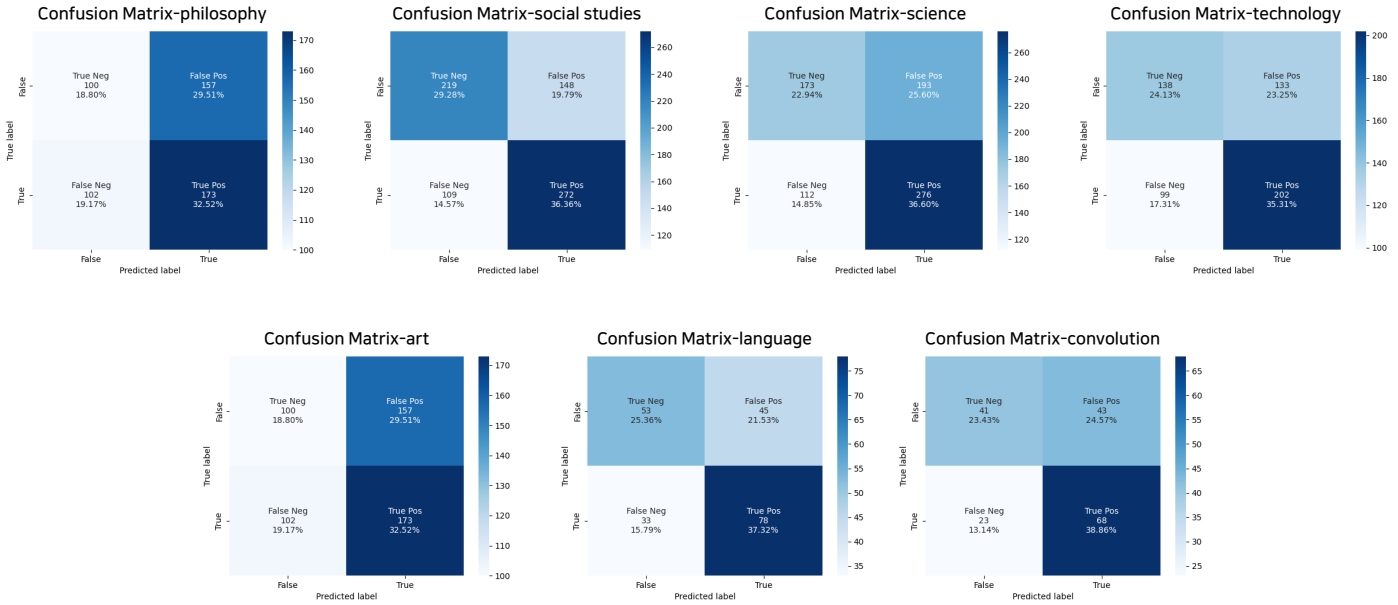

Figure 4. Visualization of augmented language models with confusion matrices of each field.

Table 7. Comparison of the best and worst scores in passages. The passage index in the left side column is organized in the order of year-month-domain, and A and B (e.g., 9A, 6B) describe that it was performed twice in month.

Passage Index	F1 Score	Recall	Precision	Recall-Precision
2013-9-tech	100.00%	100.00%	100.00%	0.00%
2006-6-phil	100.00%	100.00%	100.00%	0.00%
2016-9A-soc	94.12%	88.89%	100.00%	−11.11%
2010-9-sci	94.12%	100.00%	88.89%	11.11%
2004-6-art	94.12%	100.00%	88.89%	11.11%
2015-6B-phil	93.33%	87.50%	100.00%	−12.50%
2011-11-tech	90.91%	100.00%	83.33%	16.67%
2004-6-phil	90.91%	100.00%	83.33%	16.67%
2015-6A-phil	88.89%	100.00%	80.00%	20.00%
2013-11-tech	88.89%	100.00%	80.00%	20.00%
Average	**93.53%**	**97.64%**	**90.44%**	**7.19%**
2013-11-art	20.00%	33.33%	14.29%	19.05%
2010-6-art	20.00%	20.00%	20.00%	0.00%
2003-11-art	20.00%	20.00%	20.00%	0.00%
1994-11-soc	18.18%	20.00%	16.67%	3.33%
2014-5A-phil	18.18%	16.67%	20.00%	−3.33%
2013-11-phil	15.38%	20.00%	12.50%	7.50%
2014-6A-tech	0.00%	0.00%	0.00%	0.00%
2005-9-tech	0.00%	0.00%	0.00%	0.00%
2009-9-phil	0.00%	0.00%	0.00%	0.00%
1996-11-phil	0.00%	0.00%	0.00%	0.00%
Average	**11.17%**	**13.00%**	**10.35%**	**2.65%**

Considering the experimental results, multiple-choice questions can only be judged according to the given passages. However, practically, more detailed, professional knowledge is required, and the limit in fully understanding the context only by the knowledge of the

language models will be obvious. This is an argument that supports the fact that students should endeavor to make efforts to learn related subject knowledge for some fields when studying for the CSAT.

7. Discussion

In this study, we applied data augmentation to improve the skills of truly understanding the given passage via language models. As we can see in Figure 4 described in previous sections, certain deviations exist depending on which domain the passage belongs to. Furthermore, as we can see in Table 7, the different degrees of inference required for each passage led to differences in performance in each passage. Combining the above results shows that the current language model, which already shows outstanding performance in most MRC datasets including a Korean question answering dataset, has not reached the performance of *understanding* the given passage to the same level as humans [34].

8. Conclusions

In this study, we proposed a KCQA model called AI student, which is a Korean reading comprehension system, to evaluate the possibility of the intrinsic reasoning skills of PLMs via the CSAT to that of humans. We performed data augmentation with four EDA-based strategies and verified whether the CSAT corpus determined a question with given passages from the pedagogical perspective by exploiting various Korean and multilingual PLMs. The results demonstrated that the proposed KCQA could determine the appropriateness of the given questions based on the passages and that the EDA-based data augmentation process had the potential to improve the understanding of a passage when access to domain-specific knowledge was allowed for a particular subject. Although the non-augmented model exhibited better performance, it was difficult to judge when a given passage was truly understood. Finally, we expect that the performance of models can be evaluated based on a high-order educational CSAT corpus in terms of pedagogy in the future.

Author Contributions: Conceptualization, G.K. and J.J.; data curation, G.K., S.L. and C.P.; formal analysis, G.K. and C.P.; funding acquisition, G.K. and J.J.; investigation, S.L.; methodology, G.K.; resources, S.L.; software, G.K.; supervision, J.J.; validation, G.K.; writing—original draft, G.K., S.L. and C.P.; writing—review and editing, G.K. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-0-01405) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rajpurkar, P.; Jia, R.; Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 2, pp. 784–789. [CrossRef]
2. Zhang, Z.; Zhao, H.; Wang, R. Machine reading comprehension: The role of contextualized language models and beyond. *arXiv* **2020**, arXiv:2005.06249.
3. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186. [CrossRef]

4. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2020**, arXiv:1907.11692.
5. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* **2019**, arXiv:1910.10683.
6. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (accessed on 5 April 2022).
7. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
8. Li, D.; Hu, B.; Chen, Q.; Peng, W.; Wang, A. Towards Medical Machine Reading Comprehension with Structural Knowledge and Plain Text. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 1427–1438. [CrossRef]
9. Li, J.; Zhong, S.; Chen, K. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, 7–11 November 2021; pp. 8862–8874. [CrossRef]
10. Lin, G.; Miao, Y.; Hu, Y.; Shen, Z. Support Cybersecurity Risk Public Awareness with AI Machine Comprehension. *Int. J. Inf. Technol.* **2019**, *1*, 22.
11. Musser, M.; Garriott, A. *Machine Learning and Cybersecurity*; Center for Security and Emerging Technology: Washington, DC, USA, 2021. Available online: https://cset.georgetown.edu/wp-content/uploads/Machine-Learning-and-Cybersecurity.pdf (accessed on 5 April 2022).
12. Sennrich, R.; Haddow, B.; Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 86–96. [CrossRef]
13. Moon, J.; Cho, H.; Park, E.L. Revisiting Round-trip Translation for Quality Estimation. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisboa, Portugal, 3–5 November 2020; pp. 91–104.
14. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 6382–6388. [CrossRef]
15. Fellbaum, C. *WordNet and Wordnets*; Princeton University: Princeton, NJ, USA, 2005.
16. Kim, S. Analysis of Students' Recognition of National Scholastic Aptitude Test for University Admission—With Focus on the 'Korean Language Section'. *J. Cheong Ram Korean Lang. Educ.* **2014**, *49*, 135–164. [CrossRef]
17. Kwon, S.K.; Lee, M.; Shin, D. Educational assessment in the Republic of Korea: Lights and shadows of high-stake exam-based education system. *Assess. Educ. Princ. Policy Pract.* **2017**, *24*, 60–77. [CrossRef]
18. Park, K. An Analysis and Improvement of the Korean Language Section of CSAT. *J. Cheong Ram Korean Lang. Educ.* **2014** , *49*, 31–50. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
20. Kim, G.; Lee, C.; Jo, J.; Lim, H. Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 2341–2355. [CrossRef]
21. Kim, G.; Son, J.; Kim, J.; Lee, H.; Lim, H. Enhancing Korean Named Entity Recognition with Linguistic Tokenization Strategies. *IEEE Access* **2021**, *9*, 151814–151823. [CrossRef]
22. Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; Hajishirzi, H. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 1896–1907. [CrossRef]
23. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
24. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [CrossRef]
25. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2383–2392. [CrossRef]
26. Park, J. KoELECTRA: Pretrained ELECTRA Model for Korean. 2020. Available online: https://github.com/monologg/KoELECTRA (accessed on 5 April 2022).
27. Kwon, T.; Kim, S. A Research on the Quality of 'Speech and Writing' Evaluation Items in College Scholastic Ability Test (CSAT). *Cheong Ram Korean Lang. Educ.* **2019**, *71*, 161–194.
28. Oh, J.I.; Shin, Y. A corpus-based analysis of the linguistic complexity levels of reading passages in the Korean college entrance examination. *Engl. Lang. Teach.* **2020**, *32*, 109–126.
29. Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding Back-Translation at Scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 489–500. [CrossRef]

30. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Text data augmentation for deep learning. *J. Big Data* **2021**, *8*, 1–34. [CrossRef] [PubMed]
31. Huang, L.; Pan, W.; Zhang, Y.; Qian, L.; Gao, N.; Wu, Y. Data augmentation for deep learning-based radio modulation classification. *IEEE Access* **2019**, *8*, 1498–1506. [CrossRef]
32. Yang, Y.; Kang, S.; Seo, J. Improved machine reading comprehension using data validation for weakly labeled data. *IEEE Access* **2020**, *8*, 5667–5677. [CrossRef]
33. Nam, Y. An ERP study on the processing of Syntactic and lexical negation in Korean. *Korean J. Cogn. Sci.* **2016**, *27*, 469–499.
34. Lim, S.; Kim, M.; Lee, J. KorQuAD1.0: Korean QA Dataset for Machine Reading Comprehension. *arXiv* **2019**, arXiv:cs.CL/1909.07005.