Journal of The Korea Convergence Society
Vol. 13. No. 1, pp. 81-88, 2022
e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2022.13.01.081

대학수학능력시험 독서 영역의 교육 목표를 위한 자연어처리 기법을 통한 검증

이수민[1], 김경민[2], 임희석[3*]
[1]고려대학교 컴퓨터학과 석사과정, [2]고려대학교 컴퓨터학과 석박통합과정, [3]고려대학교 컴퓨터학과 교수

Verification of educational goal of reading area in Korean SAT through natural language processing techniques

Soomin Lee[1], Gyeongmin Kim[2], Heuiseok Lim[3*]
[1]Master Student, Department of Computer Science and Engineering, Korea University
[2]Master&Ph.D Combined Student, Department of Computer Science and Engineering, Korea University
[3*]Professor, Department of Computer Science and Engineering, Korea University

요 약 대학수학능력시험 국어 과목에서 중요한 비중을 차지하는 독서 영역의 주된 교육 목표는 주어진 지문을 온전히 이해할 수 있는가를 평가하는 데에 있다. 따라서 해당 지문에 포함된 질의를 주어진 지문만으로 풀이할 수 있는지는 해당 영역의 교육 목표와 관련이 깊다. 본 연구에서는 처음으로, 교육학 분야와 딥러닝을 접목하여 이러한 교육 목표가 실제로도 타당하게 실현 가능한지를 입증하고자 한다. 대학수학능력시험의 독서 영역의 개별 지문과 그에 수반된 다수의 문장 쌍(sentence pair)을 정제하여 추출하고, 해당 문장 쌍을 주어진 지문에 비추어 적절하거나(T), 적절하지 않은지(F)를 판단하는 이진 분류 태스크(binary classification task)에 적용하여 평가하고자 한다. 그 결과, F1 스코어 기준 59.2%의 human performance를 뛰어넘는 성능을 62.49%의 KoELECTRA를 비롯한 대부분의 언어 모델에서 확인할 수 있었으며, 또한 데이터 전처리 과정에 변화를 줌으로써 언어 모델의 구조적 한계를 극복할 수 있었다.

주제어 : 대학수학능력시험, 딥러닝, 이진 분류 태스크, 언어 모델, 데이터 전처리

Abstract The major educational goal of reading part, which occupies important portion in Korean language in Korean SAT, is to evaluated whether a given text can be fully understood. Therefore given questions in the exam must be able to solely solvable by given text. In this paper we developed a datatset based on Korean SAT's reading part in order to evaluate whether a deep learning language model can classify if the given question is true or false, which is a binary classification task in NLP. In result, by applying language model solely according to the passages in the dataset, we were able to acquire better performance than 59.2% in F1 score for human performance in most of language models, that KoELECTRA scored 62.49% in our experiment. Also we proved that structural limit of language models can be eased by adjusting data preprocess.

Key Words : Korean SAT, Deep learning, Binary classification task, Language models

*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) and supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

*Corresponding Author : HeuiSeok Lim(limhseok@korea.ac.kr)

Received October 7, 2021
Revised November 10, 2021
Accepted January 20, 2022
Published January 28, 2022

1. 서론

1.1 대학수학능력시험

언어를 막론하고 교육의 측면에서 '독서'가 갖는 위상은 높다. 특히 한국에서는 국어교육 중에서도 독서 교육이 매우 특수한 위치를 차지하고 있다. 교육과정개발원 연구 보고서에 따르면 독서 교육을 국어교과의 하위 영역으로 규정하면서도, 동시에 독서가 국어과뿐만 아니라 학교의 모든 교과에서 요구되는 교수-학습의 한 요소라고 지적하고 있다[1]. 비슷한 맥락에서 2015 개정 초중등학교 교육과정에서도 국어가 전 교과를 통틀어 필수로 요구되는 학습의 중요한 토대임을 강조하고 있다[2].

널리 통용되는 대입 시험인 SAT와 달리 한국에서만 통용되는 시험인 대학수학능력시험은 1994년부터 시행된 전국 단위 시험으로 대학 진학을 목표로 하는 수험생들이 사실상 필수적으로 응시해야 하는 시험이다. 시험은 독서, 문학으로 세분화되는 국어 영역, 수학 영역, 영어 영역, 사회탐구 영역, 과학탐구 영역 등으로 세분화된다.

한국에는 대학수학능력시험 이외에도 다양한 형태의 전국 단위 시험이 존재하는데, 대표적으로 국가수준 학업 성취도 평가가 있다. 다만 대학수학능력시험은 여타의 전국 단위 시험과 달리 대학 진학을 결정하는 강력한 요인으로 작용한다는 점에서 고부담 평가에 속한다는 특징이 있다[3]. 그 중, 국어 영역은 독서, 문학, 화법과 작문, 언어와 매체 분야로 세분화된다. 독서 분야는 일반적으로 1,700자 내외의 지문에 대한 이해를 토대로 평균 15개의 주어진 문장 형태의 질의가 적절한지(T), 혹은 적절하지 않은지(F)를 판단하기를 요구한다.

1.2 언어 모델을 활용한 대학수학능력시험의 교육적 목표 검증

대학수학능력시험에서 학생들에게 요구하는 이진 분류 태스크는 주어진 지문에 대한 높은 수준의 이해를 전제로 한다. 문제는 지문을 이해하는 과정에서나, 질의를 판단하는 과정에서나 사전 지식이 크게 개입할 수 있다는 점이다. 대표적으로 일부 지문은 특정 교과의 배경지식을 갖고 있는 학생들에게 유리하다는 비판이 제기되었다[4]. 이는 오로지 주어진 지문에 대한 독해 역량을 평가하고자 하는 대학수학능력시험의 교육 목표와 배치되는 면이다. 본 논문에서는 이러한 문제 제기를 컴퓨터 과학 분야에서 활발히 연구되는 사전 학습된 언어 모델을 활용하여 검증하고자 한다.

언어 모델은 대학수학능력시험 독서 영역에서 지문에 대한 질의가 주어졌을 때, 언어 모델이 실제로 수험생과 비슷한 수준의 성능을 보인다면, 주어진 지문만을 토대로 하여도 충분히 질의를 판단할 수 있다는 결론에 도달할 수 있다. 본 논문은 한국어 기반의 PLM을 통해 대학수학능력시험에서 요구하는 것과 동일한 방식으로, 주어진 질의에 대한 적절성 여부를 결정하는 이진분류 태스크를 수행한다. 이를 위해 독서 분야의 지문과 질의, 그리고 그에 대한 정답을 기반으로 본 연구를 KSQA (Korean SAT Question and Answering, KSQA)라 정의한다.

본 논문은 독서 문항에 대응하기 위해 다양한 외부지식(external knowledge)을 학습하는 실제 수험생들의 학습 방법에 착안하여, 데이터 처리 과정에서 영어-한국어 Back-translation[5]을 거쳐 데이터 증강을 수행한다. 그와 동시에 단어 간 유의어, 반의어와 같은 관계를 담고 있는 wordnet을 기반으로 데이터 증강 기법을 함께 적용함으로써 외부지식(external knowledge)을 증강 과정에 포함시킨다. 이와 같은 과정을 거쳐 구축된 KSQA 데이터 셋에 대한 분류 태스크의 성능을 평가하는 과정에서 고난도 한국어 분류 태스크에 적합한 한국어 PLM이 무엇인지, 그리고 트랜스포머 기반 모델의 한계를 극복하기 위해 언어 모델 내에서가 아닌, 데이터 전처리 과정 내에서 어떤 수단이 도입될 수 있는지에 대해 초점을 두어 설명하고자 한다.

본 연구의 순서는 다음과 같다. 2장에서는 교육학 분야와 최근 자연어처리, 그리고 데이터 증강에 관한 관련 연구, 3장에서는 대학수학능력시험과 본 연구에서 구축한 KSQA 데이터셋에 관해 소개하고, 4장에서는 사전 학습 모델 기반 실험, 5장은 본 연구에 대한 분석 및 한계점, 그리고 마지막으로 결론에 대해 논의한다.

2. 관련 연구

2.1 트랜스포머 기반 언어 모델의 도입

트랜스포머(Transformer) 모델[6] 기반의 사전 학습 언어 모델(Pre-trained Language Model, PLM), 과 미세 조정 단계(fine-tuning)로 구성된 전이 학습 형태의 언어 모델 학습 절차는 개체명 인식, 질의응답, 텍스트

생성, 텍스트 분류를 비롯한 다양한 자연어처리 (Natural Language Processing, NLP) 분야에서 뛰어난 성능을 보여 왔다. 가장 최초의 양방향 트랜스포머 인코더 기반의 사전 학습 언어 모델인 BERT (Bidirectional Encoder Representations from Transformers, BERT)는 질의응답 태스크의 대표적인 데이터셋인 SQuAD v1.1에서 93.2%의 F1 스코어를 기록하였으며 이는 Human Performance를 2.0%나 앞선 결과이다.

대부분의 NLP task에서 높은 성능을 보이는 최근 Transformer 기반 모델은 MLM (Masked Language Model, MLM), NSP (Next Sentence Prediction, NSP), RTD (Replaced Token Detection, RTD) 와 같은 다양한 기법을 적용하여 각 down-stream task를 위해 텍스트 내의 문맥 정보 (contextual information)을 추출한다. 이전에는 한국어 분야의 연구를 수행하기 위해 다국어 mBERT(multilingual-BERT, mBERT) 모델이 주로 사용되었으나, 이는 한국어 데이터로 사전 학습이 충분히 이루어지지 못했다는 점에서 한계가 있다.

2.2 한국어 기반 사전 학습 언어 모델

최근에는 한국어 커뮤니티에서 다양한 한국어 PLM이 공개되고 있으며, 한국어 데이터에 특화되어 사전 학습이 이루어졌다는 점에서 기존의 다국어 모델을 지원하는 언어 모델인 mBERT 등과 비교하여 탁월한 성능을 보인다. 이는 한국어의 교착어 특성을 반영하여 형태소 단위에서의 처리에 주목할 수 있기 때문이다. 오늘날 대표적인 한국어 언어 모델에는 정제된 언어를 기반으로 사전 학습이 이루어진 KoELECTRA, KoBERT 가 있고, 반대로 정제되지 않은 댓글(comments) 데이터 사전 학습이 이루어진 KcELECTRA 등이 있다. 그중에서도 KoELECTRA는 모두의 말뭉치를 비롯하여 34GB의 상당한 규모의 대용량 데이터를 학습시킴에 따라 오늘날 네이버 영화 리뷰 말뭉치인 NSMC(Naver sentiment movie corpus, NSMC)를 제외한 대부분의 한국어 기반의 down-stream task에서 가장 뛰어난 성능을 보인다.

2.3 교육 분야에서의 NLP 접목 사례

SAT는 사설 기관인 ETS의 주관 하에 진행된다. 더 나아가 ETS는 NLP의 다양한 기법들을 활용하되, 채점 기법의 효율화에 국한되어 있다[7].

반면 대학수학능력시험의 경우 국가 기관인 교육과정평가원의 주관 하에 진행되는 시험으로 역대 출제된 모든 시험들이 공개되어 있음에도 불구하고, 그것을 컴퓨터과학의 측면에서 분석하려는 연구는 아직까지 시도된 바가 없다.

2.4 한국어 PLM 연구

2.4.1 KoBERT

PLM은 대용량 데이터셋을 사전 학습 모델이다. 다국어로 학습한 mBERT의 경우, 한국어로 직접 사전 학습한 언어 모델과 비교하여 적은 문장 비중으로 인해, 한국어 NLP 태스크에 적용되기 어려운 한계점이 존재한다. SKT에서 공개한 KoBERT는 BERT모델 구조를 활용하여 한국어 위키 5백만 개의 문장을 기반으로 사전 학습한다[8].

2.4.2 KoELECTRA

KoELECTRA는 KoBERT와 달리 ELECTRA[9] 모델 기반의 한국어 PLM이다. ELECTRA는 RTD를 채택하는데, 이는 특정 비율의 토큰을 [MASK]로 변환한 후에, generator로 하여금 [MASK]에 입력되기에 적합한 토큰을 생성하도록 한다. 그 과정에서 discriminator는 generator의 출력을 토대로 어느 토큰이 교체되었는지를 판단함으로써 학습한다. KoELECTRA는 뉴스, 한국어 위키, 모두의 말뭉치를 포함하여 약 34G의 한국어 문장을 학습한다[10].

2.4.3 KcELECTRA

KoELECTRA와 마찬가지로 ELECTRA 기반의 모델이지만 학습 데이터의 성격과 규모 면에서 큰 차이가 있다. KcELECTRA는 뉴스 댓글과 같이 정제되지 않은 한국어 텍스트에 적용되는 다양한 NLP 태스크에 특화된 모델이다. 대표적으로 NSMC는 실제 사용자들의 영화 리뷰 데이터로, 구어체적 특성을 갖는다. KcELECTRA는 네이버 뉴스의 댓글 약 17GB를 수집하여 학습하였기 때문에 상대적으로 정제되지 않은 데이터(unrefined corpus)를 다수 학습하였다. 실제로 정제된 데이터(refined corpus)를 사전 학습한 여타의 모델에 비해 NSMC에서 F1 스코어 기준 91.71%라는 가장 탁월한 성능을 보인다[11]. 다양한 한국어 PLM 비교에 대한

요약은 Table 1에서 확인할 수 있다.

Table 1. Korean PLM

Model	Architecture	Trained Data
KoBERT	BERT	refined corpus
KoELECTRA	ELECTRA	refined corpus
KcELECTRA	ELECTRA	unrefined corpus

2.5 데이터 증강 기법

데이터 증강은 학습 데이터가 부족한 때에 학습 데이터의 규모를 증강시키는 기법이다. 효과적인 데이터 증강 수행을 위해서는 그 기초가 되는 학습데이터를 변형시키되, 기존 학습데이터의 의미를 훼손하지 않는 범위 내에서 이루어져야 한다[12]. 따라서, 데이터 증강을 컴퓨터 비전 분야에서와 달리 NLP 분야에 본격적으로 적용하기에는 여러 한계점이 존재했으나, 이는 해당 연구가 텍스트 기반의 학습 데이터를 효과적으로 증강하는 EDA 기법을 도입하는 계기가 되었다.

컴퓨터 비전의 데이터 증강 기법에 영감을 받은 EDA는 주어진 문장을 동의어 대체(Synonym Replacement, SR), 무작위 삽입(Random Insertion, RI), 무작위 교체(Random Swap, RS), 무작위 삭제(Random Deletion, RD)라는 네 가지 카테고리에 의하여 증강시키는 기법이다. SR은 토큰이 wordnet을 참조하여 유의어 토큰으로 대체하는 방법이다. RI는 특정 위치에 무작위 토큰을 삽입하는 방법이다. RS는 두 임의의 토큰 위치를 교체하는 방법이다. 마지막으로 RD는 임의의 토큰을 문장에서 삭제하는 방법이다. 또한, 해당 연구에서는 데이터 증강을 수행할 경우 최대 0.8% 성능 향상이 있음을 보였다.

3. 대학수학능력시험과 KSQA

3.1 대학수학능력시험의 데이터 추출

본 연구에서는 대학수학능력시험 영역 중에서도 국어 영역의 독서 영역에 초점을 두고자 한다. 독서 영역은 3-4개의 지문으로 구성되어 있으며, 각 지문은 4-5개의 문항을 포함한다. 각 문항에는 5개의 질의가 존재하며 수험생은 5개의 질의 중 적절하지 않은 1개의 문장 혹은 적절한 1개의 문장을 선택할 수 있어야 한다. 이는 Fig 1.에서 확인할 수 있다.

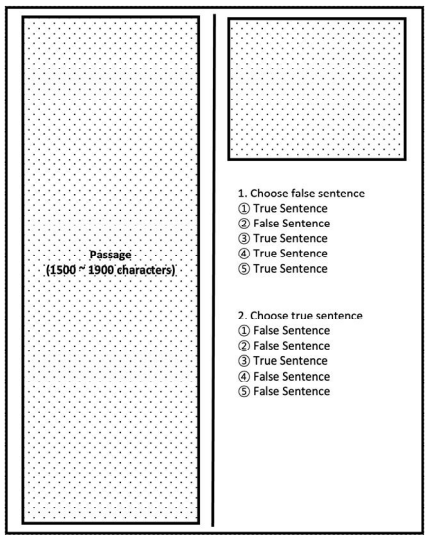

Fig. 1. Configuration of Korean SAT reading

각 선지의 적절성 여부를 판단하기 위해, 함께 주어진 지문의 표면적 정보는 물론, 의미론적 정보와 문맥적 정보까지 읽어낼 수 있는 능력을 요구한다. 이는 최근 기계독해가 진정으로 주어진 텍스트를 이해하고 있는가를 문제제기하는 NLP 분야에서의 과제와도 관련이 깊다. 대표적으로 SQuAD 1.0[13]은 질의응답 데이터셋의 일종으로 단순히 주어진 텍스트를 기반으로, 정답에 대응하는 스팬(span)의 위치를 추출해내는 작업을 수행하는데, 주어진 텍스트에 지문과 무관한 '교란 정보'가 개입할 경우, 기계독해의 성능이 급격히 떨어진 연구 결과가 있다[14].

실제로 평가 주체인 교육과정평가원은 대학수학능력시험과 SAT의 문항별 부담을 비교한 한 연구[15]에 따르면, SAT에서 문항별 주어지는 시간이 약 54초일 경우, 대학수학능력시험의 경우에는 약 88초를 요구한다. 즉, 대학수학능력시험이 그만큼 높은 수준의 추론적 사고를 요구한다는 것을 알 수 있다.

3.3 KSQA 데이터 전처리 과정

본 연구에서 활용하는 학습 데이터는 대학수학능력시험에서 출제한 지문과 그에 수반된 질의들이다. 그러나 2015년 개정교육과정 이후 늘어난 문장의 길이가 평균 30.76문장인 것을 고려하여도, 충분한 학습 데이터로 활용될 수 없다. 또한, 실제 시험에서 제한된 시간에서의 독해 능력을 묻는 것을 목표로 하므로, 지문의 길이

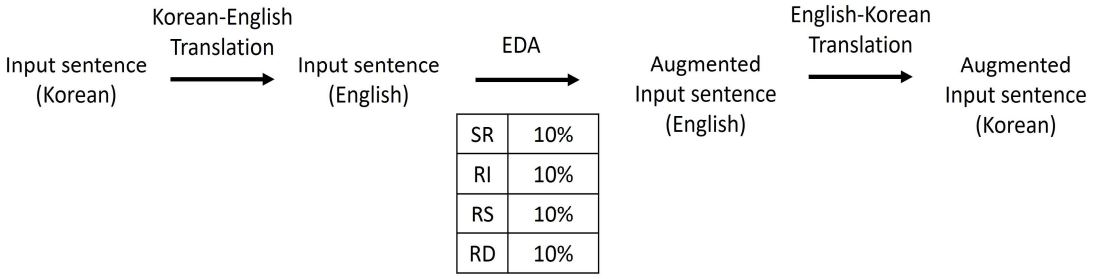

Fig. 2. KSQA preprocess method

확장에 한계가 있다.

본 논문은 이러한 한계점을 극복하기 위해 EDA에서 사용한 4가지 학습 데이터 증강 기법을 적용한다. 이는 해당 기법이 NLP 분야에서 효과적이기도 하지만, 특히 학습 데이터의 규모가 작은 경우에 효과적이다. SR, RI, RS, RD 네 가지 기법을 모두 주어진 문장에 10%의 동일한 비율을 적용하여 문장을 변형시키며, 해당 방식으로 기존 학습 데이터를 10배 증강시킨다. 이때, 주어진 문장에 포함된 단어가 유의어로 교체되거나 문장의 중간에 무작위 단어가 삽입되기 위해서는 wordnet[16]을 참조해야 하는데, 문제는 EDA 기법에서 활용한 영어 wordnet과 달리 한국어 wordnet은 보편적으로 갖추어져 있지 않다는 점이다. 본 연구에서는 주어진 학습 데이터를 문장별로 분할한 이후에 먼저 한국어-영어 번역 (Korean-to-English Translation)을 수행한 다음, 해당 문장에 대해 EDA 기법을 수행하고 증강된 학습 데이터에 대해 다시 영어-한국어 번역 (English-to-Korean Translation)을 수행하는 방법을 채택한다. 이를 통해 영어 wordnet을 활용하여 효과적으로 데이터 증폭을 수행할 수 있다. 이러한 방식은 기계번역에서의 품질 예측 (Quality Estimation, QE)을 위한 순환 번역 (Round-Trip Translation, RTT)에서도 사용된다[17]. 전처리 과정의 순서는 Fig. 2와 같다.

3.4 KSQA 언어 모델

본 논문에서는 실험을 위해 한국어 PLM을 기반으로 한다. 전체 학습 데이터에서 먼저 순차적으로 하나의 지문에 포함된 데이터 증강 과정을 거친 입력 문장 (augmented input sentence)에 대해 모델에 입력한다. 이때 입력된 문장의 [CLS] 토큰을 정답에 해당하는 적절한 문장으로 학습하도록 PLM을 학습한다. 특정 지문의 문장들이 입력될 때 그와 동일한 수의 적절하지 않은 문장을 다른 지문의 문장들로 무작위로 구성함으로써, 각 지문을 학습하는 과정에서 동일한 수의 적절한 문장과 그렇지 않은 문장이 입력되도록 한다. 각 지문을 학습하였을 때의 성능은 해당 지문에 포함된 질의의 레이블을 올바르게 분류할 수 있는지를 통해 확인한다. 각 질의는 사전에 적절한 문장인지, 적절하지 않은 문장인지를 입력받는다. 모델의 구조는 Fig. 3과 같다.

4. 실험

4.1 KSQA 구성

학습에 동원된 지문의 개수는 총 269개이다. 각각의 지문은 해당 시험의 입학년도-분야에 따라 index를 부여받는다. 이때의 분야는 해당 지문이 다루는 학문의 영역으로 인문, 사회, 예술, 과학, 기술, 그리고 2012년 이후로 출제된 바가 없는 언어 분야가 있다. 각각의 지문은 훈련 데이터를 구성하고, 각 지문에 귀속된 적절하거나(T) 적절하지 않은(F) 질의들로 시험 데이터를 구성한다. 전체 데이터의 규모는 각 학문 영역별 훈련 데이터의 평균 문장 개수, 시험 데이터의 적절한, 적절하지 않은 질의의 평균 문장 개수를 Table 2에서 확인할 수 있다.

Table 2. KSQA Dataset

Branch(#)	Train data (augmented)	Test data (True, #)	Test data (False, #)
인문(56)	214.63	6.16	5.71
사회(57)	209.07	6.53	5.33
과학(52)	214.08	7.36	6.86
기술(42)	213.60	6.95	6.31
예술(45)	204.96	6.11	5.71
언어(23)	194.70	4.83	4.26
융합(8)	399.13	10.25	9.75

4.2 평가 준거

본 연구에서는 각 지문별 성능을 평가하기 위해서 F1 스코어를 척도로 활용한다. F1 스코어는 정밀도(Recall, R)와 재현율 (Precision, P)을 조화평균 한 것으로 개체명 인식 태스크에서 활용되는 평가 기준이다. 이는 단순히 정확도 (Accuracy)를 활용했을 때 발생하는 한계를 극복하기 위한 방안으로 도입되었다. 이는 아래 수식을 통해 구할 수 있다. 수식에서 S_p는 학습 과정에서 예측한 적절한 문장의 집합을 의미하고, S_r는 실제 적절한 문장의 집합을, $C = S_p \cap S_r$는 적절한 문장들을 의미한다.

$$P = \frac{|C|}{|S_P|}, R = \frac{|C|}{|S_r|}, F1 = \frac{2PR}{P+R}$$

4.3 실험 결과

실험을 위해 사용된 모델들은 KoELECTRA, Xlm-Roberta, KoBERT, KcELECTRA, MBERT로 정확한 성능 평가를 위해 모두 동일한 하이퍼파라미터 (Hyperparameter) 환경에서 학습을 수행하였다. 하이퍼파라미터 환경은 Table 3에서 확인할 수 있다.

Table 3. Hyperparameter

Hyperparameter	Value
Batch size	32
Optimizer	AdamW
number of epochs	40
learning rate	1e-5

Table 4는 모델별 Macro F1 스코어를 비교한 성능을 나타낸다. mBERT와 KoBERT를 제외하면 대부분의 언어모델은 Human Performance을 상회하는 성능을 보였다. Human Performance의 경우 실제로 올해 대학수학능력시험 응시를 목표로 하는 수험생 30명을 대상으로 KSQA 데이터셋을 질의 형태로 변환하여, 실제 시험지와 동일한 조건에서 측정하였다. 이는 주어진 지문만으로도 주어진 질의에 대해 올바른 판단을 내리는 것이 충분히 가능함을 방증하는 것으로, 국어 영역에서의 독해력 평가로서의 교육적 목표가 달성될 수 있음을 보인다.

그중에서도 Xlm-Roberta, KoELECTRA, KcELECTRA에서 가장 좋은 성능을 보였으며, 특히 KoELECTRA에서 재현율과 정밀도 사이의 차이가 가장 작았다. 즉, KoELECTRA 모델이 다른 모델과 비교하여 가장 효과적 문맥 이해도를 보인다고 할 수 있는데, 이는 KoELECTRA가 사전학습 과정에서 다른 모델에 비해 상대적으로 정제된 언어로 학습한 것과 관련이 있다. 대학수학능력시험의 국어 영역 역시 상대적으로 정제된 언어로 구성되어 있다고 볼 수 있기 때문이다.

Table 4. Experimental Results

Model	Macro F1 score	Recall (Average)	Precision (Average)	Recall - Precision (Average)
KoELECTRA	**62.49%**	66.91%	61.22%	**5.7%**
Xlm-Roberta	**62.82%**	68.61%	60.02%	8.59%
KoBERT	57.88%	61.69%	58.69%	3%
KcELECTRA	**62.72%**	68.12%	61.19%	6.93%
mBERT	57.34%	60.84%	59.76%	1.08%
Human Performance	59.2%	62.38%	59.27%	3.11%

Table 5는 데이터 증강 기법을 적용한 성능에 대해 나타낸다. KoELECTRA를 기준으로 했을 때 데이터 증강을 수행하지 않은 경우 Macro F1 스코어가 4.15% 높아지지만, 재현율과 정밀도 간의 차이는 36.67%의 큰 폭으로 늘어난다는 사실을 확인할 수 있다. 이는 데이터 증강이 이루어지지 않은 경우에는 적절한(T) 문장은 물론, 적절하지 않은(F) 문장의 경우에도, 그것이 표면적으로 주어진 지문의 표현과 유사하다는 이유만으로 전부 적절한 것으로 예측하기 때문이다.

이를 통해 데이터 증강 기법을 적용함으로써 언어 모델로 하여금 KSQA 데이터셋의 지문을 표면적인 수준을 넘어, 보다 깊이 이해할 수 있도록 할 수 있다는 사실을 확인할 수 있었다.

Table 5. Performance of Data Augmentation

Model (KoELECTRA)	Macro F1 score	Recall (Average)	Precision (Average)	Recall - Precision (Average)
Augmented	62.49%	66.91%	61.22%	5.7%
Non-Augmented	66.64%	94.86%	52.48%	42.37%

5. 결론

오늘날의 BERT를 비롯한 사전 학습 모델에는 몇 가지 문제점이 있다. 첫째로, 현 언어 모델은 일부 특정 태스크에 국한되어서만 사용될 수 있으며, 일반적인 용도로는 사용되기 어렵다는 점이다. 둘째로, 언어에서의 "이해"는 주관적인 요소가 반영되기 때문에, 언어 모델이 주어진 데이터셋을 어떻게 이해하였는지를 일정한 기준에 따라 평가하는 데 한계가 있다. 또한, 언어 모델을 사전 학습하기 위해서는 방대한 하드웨어 리소스를 요구한다는 한계가 있으며, 이는 전처리 과정을 통한 성능 개선에 대한 근거가 될 수 있다.

이에 본 연구에서는 대학수학능력시험 국어 영역을 토대로 하는 KSQA 데이터셋을 생성하여 해당 시험의 수험생들의 학습 방식을 참조하여 데이터 증강을 수행했다. 해당 데이터셋에 다양한 한국어 PLM을 적용하여 과연 대학수학능력시험이 교육적 차원에서 별도의 특정 교과 지식이 없더라도 주어진 지문만을 근거로 판단할 수 있는 질의들로 구성되었는지 검증하고자 했다. 그 결과, 주어진 지문만을 토대로 하여도 충분히 주어진 질의가 적절한지에 대한 여부를 판단할 수 있음을 입증할 수 있었다. 둘째로, 증강을 거치지 않은 경우에 성능 자체는 좋아 보일 수 있어도, 주어진 지문을 진정한 의미에서 이해한 것으로 보기 어렵다는 결론을 도출할 수 있다.

REFERENCES

[1] M. H. Roh. (2011). Reading: the concept and important issues of education. *KEDI Research Paper, 43(3)*, 1-43.

[2] Ministry of Education. (2015. January). Korean Language Curriculum. *2015 Revised Curriculum, 5*, 1-178.

[3] S. H. Kim. (2014). Analysis of Students' Recognition of National Scholastic Aptitude Test for University Admission -With Focus on the 'Korean Language Section'-, *Journal of CheongRam Korean Language Education, 49*, 135-164.

[4] S. Y. Ryu. (2019). Critical Examination About CSAT Korean Language and Its Developmental Directions - Toward the Recovery of the Nature of the CSAT Evaluation -. *New Language Education, 121,* 353-380.

[5] R. Sennrich, B. Haddow & A. Birch. (2016). Improving Neural Machine Translation Models with Monolingual Data. *Association for Computational Linguistics, 1,* 86-96.

[6] A. Vaswani et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 6000-6010.

[7] D. Powers, D. Escoffery, & M. Duchnowski. (2015). Validating Automated Essay Scoring: A (Modest) Refinement of the "Gold Standard". *Applied Measurement in Education, 28(2)*, 130-142.

[8] SKT Brain. (2019). *KoBERT*. Github Repository. https://github.com/SKTBRain/KoBERT

[9] K. Clark, M. Luong, Q. Le, & C. Manning. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

[10] J. W. Park. (2020). *KoELECTRA: Pretrained ELECTRA Model for Korean*. Github Repository. https://github.com/monologg/KoELECTRA

[11] J. B. Lee. (2021). *KcELECTRA : Korean comments ELECTRA*. Github Repository. https://github.com/Beomi/KcELECTRA

[12] J. Wei & K. Zou. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance. *Association for Computational Linguistics, 1,* 6382-6388. DOI : 10.18653/v1/D19-1670

[13] P. Rajpurkar, J. Zhang, K. Lopyrev & P. Liang. (2016). SQuAD: 100, 000+ Questions for Machine Comprehension of Text, *EMNLP, 1,* 2383-2392. DOI : 10.18653/v1/d16-1264

[14] Y. Y. Yang, S. W. Kang, J. Y. Seo. (2019). Improved Machine Reading Comprehension Using Data Validation for Weakly Labeled Data. *IEEE, 8,* 5667-5677. DOI : 10.1109/ACCESS.2019.2963569

[15] Y. R. Lee et al. (2009. July.). *Analysis of SAT and ACT*, Seoul : KICE.

[16] C. Fellbaum. (2005). *WordNet and wordnets*, Oxford : Elsevier.

[17] J. H. Moon, H. C. Cho & E. J. Park. (2020). Revisiting Round-Trip Translation for Quality Estimation. http://arxiv.org/abs/2004.13937

이 수 민(Soomin Lee) [정회원]

· 2018년 8월 : 고려대학교 경영학과 (경영학사)
· 2018년 ~ 현재 : 고려대학교 컴퓨터학과 석사과정
· 관심분야 : 인공지능, 자연어처리, 기계독해
· E-Mail : skyop27@korea.ac.kr

김 경 민(Gyeong-Min Kim) [학생회원]

· 2017년 8월 : 백석대학교 정보통신학부 정보보호학과(공학사)
· 2018년 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : 인공지능, 자연어처리, 기계독해, 지식표현
· E-Mail : totoro4007@korea.ac.kr

임 희 석(Heuiseok Lim) [종신회원]

· 1992년 2월 : 고려대학교 컴퓨터학과 (이학학사)
· 1994년 2월 : 고려대학교 컴퓨터학과 (이학석사)
· 1997년 2월 : 고려대학교 컴퓨터학과 (이학박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 뇌신경 언어 정보 처리
· E-Mail : limhseok@korea.ac.kr