KIPS Trans. Softw. and Data Eng.
Vol.11, No.3 pp.133~140
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.3.133

Comparison of Korean Classification Models' Korean Essay Score Range Prediction Performance

Heeryon Cho† · Hyeonyeol Im†† · Yumi Yi††† · Junwoo Cha††††

ABSTRACT

We investigate the performance of deep learning-based Korean language models on a task of predicting the score range of Korean essays written by foreign students. We construct a data set containing a total of 304 essays, which include essays discussing the criteria for choosing a job ('job'), conditions of a happy life ('happ'), relationship between money and happiness ('econ'), and definition of success ('succ'). These essays were labeled according to four letter grades (A, B, C, and D), and a total of eleven essay score range prediction experiments were conducted (i.e., five for predicting the score range of 'job' essays, five for predicting the score range of 'happiness' essays, and one for predicting the score range of mixed topic essays). Three deep learning-based Korean language models, KoBERT, KcBERT, and KR-BERT, were fine-tuned using various training data. Moreover, two traditional probabilistic machine learning classifiers, naive Bayes and logistic regression, were also evaluated. Experiment results show that deep learning-based Korean language models performed better than the two traditional classifiers, with KR-BERT performing the best with 55.83% overall average prediction accuracy. A close second was KcBERT (55.77%) followed by KoBERT (54.91%). The performances of naïve Bayes and logistic regression classifiers were 52.52% and 50.28% respectively. Due to the scarcity of training data and the imbalance in class distribution, the overall prediction performance was not high for all classifiers. Moreover, the classifiers' vocabulary did not explicitly capture the error features that were helpful in correctly grading the Korean essay. By overcoming these two limitations, we expect the score range prediction performance to improve.

Keywords : Deep Learning-Based Korean Language Model, KoBERT, KcBERT, KR-BERT, Document Classification

한국어 학습 모델별 한국어 쓰기 답안지 점수 구간 예측 성능 비교

조 희 련† · 임 현 열†† · 이 유 미††† · 차 준 우††††

요 약

우리는 유학생이 작성한 한국어 쓰기 답안지의 점수 구간을 예측하는 문제에서 세 개의 딥러닝 기반 한국어 언어모델의 예측 성능을 조사한다. 이를 위해 총 304편의 답안지로 구성된 실험 데이터 세트를 구축하였는데, 답안지의 주제는 직업 선택의 기준('직업'), 행복한 삶의 조건('행복'), 돈과 행복('경제'), 성공의 정의('성공')로 다양하다. 이들 답안지는 네 개의 점수 구간으로 구분되어 평어 레이블(A, B, C, D)이 매겨졌고, 총 11건의 점수 구간 예측 실험이 시행되었다. 구체적으로는 5개의 '직업' 답안지 점수 구간(평어) 예측 실험, 5개의 '행복' 답안지 점수 구간 예측 실험, 1개의 혼합 답안지 점수 구간 예측 실험이 시행되었다. 이들 실험에서 세 개의 딥러닝 기반 한국어 언어모델(KoBERT, KcBERT, KR-BERT)이 다양한 훈련 데이터로 미세조정되었다. 또 두 개의 전통적인 확률적 기계학습 분류기(나이브 베이즈와 로지스틱 회귀)도 그 성능이 분석되었다. 실험 결과 딥러닝 기반 한국어 언어모델이 전통적인 기계학습 분류기보다 우수한 성능을 보였으며, 특히 KR-BERT는 전반적인 평균 예측 정확도가 55.83%로 가장 우수한 성능을 보였다. 그 다음은 KcBERT(55.77%)였고 KoBERT(54.91%)가 뒤를 이었다. 나이브 베이즈와 로지스틱 회귀 분류기의 성능은 각각 52.52%와 50.28%였다. 학습된 분류기 모두 훈련 데이터의 부족과 데이터 분포의 불균형 때문에 예측 성능이 별로 높지 않았고, 분류기의 어휘가 글쓰기 답안지의 오류를 제대로 포착하지 못하는 한계가 있었다. 이 두 가지 한계를 극복하면 분류기의 성능이 향상될 것으로 보인다.

키워드 : 한국어 심층학습 언어모델, KoBERT, KcBERT, KR-BERT, 문서 분류

1. 서 론

K-팝, K-드라마, 한국 영화 등 한류 콘텐츠가 세계적인 주목을 받으면서 한국어를 외국어 또는 제2 언어로 배우려는 학습자가 늘고 있다. 외국인 학습자를 위해 발간된 한국어 교재는 약 3,400권에 이르고, 2018년에 실시한 한국어능력시험(TOPIK)에 응시한 사람은 329,224명, 합격한 사람은 185,624명이었다. 1997년에 2,692명이 응시하고 711명이 합격했던 것과 비교해 그 수가 크게 늘었다[2]. 이처럼 한국어를 배우

※ 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A6A3A01078538).
※ 이 논문은 2021년 한국정보처리학회 춘계학술발표대회에서 "KoBERT, 나이브 베이즈, 로지스틱 회귀의 한국어 쓰기 답안지 점수 구간 예측 성능 비교"의 제목으로 발표된 논문을 확장한 것임.
† 종신회원 : 중앙대학교 인문콘텐츠연구소 HK교수
†† 비 회 원 : 중앙대학교 다빈치교양대학 조교수
††† 비 회 원 : 중앙대학교 인문콘텐츠연구소 부교수
†††† 비 회 원 : 중앙대학교 한국어교육원 강사
Manuscript Received : July 2, 2021
First Revision : August 18, 2021
Accepted : August 26, 2021
* Corresponding Author : Heeryon Cho(heeryon@cau.ac.kr)

려는 외국인이 많아지면서 컴퓨터로 한국어 학습을 지원하는 방법에 관한 관심도 높아지고 있다.

한국어 학습 방법 중 한국어 쓰기 연습은 학습자의 한국어 작문 실력을 기르는 데 도움을 준다. 그런데 학습자의 작문 실력을 가늠하기 위해서는 학습자가 작성한 한국어 문장을 교수자가 일일이 평가해야 한다. 이러한 평가 작업은 많은 시간과 노력이 들기 때문에 사람이 한정된 시간에 한정된 분량의 답안지밖에 평가할 수 없고, 채점 기준이 모호하여 일관된 채점이 어렵다[3]. 이러한 문제를 극복하기 위해 컴퓨터를 이용한 한국어 쓰기 답안지의 자동채점을 고려할 수 있다.

지금까지 한국어 주관식 답안을 자동으로 채점하기 위한 연구는 대부분 1~3개 단어로 된 단답형 문항이나 한두 문장으로 이루어진 서답형 문항을 채점하는 데 초점이 맞춰졌다[4-8]. 이들 연구는 전문가에 의해 기술된 정답 템플릿을 이용하거나 벡터 공간 모델을 이용하여 모범 답안과의 유사도를 측정하거나 다양한 기계학습 기법을 이용하여 자동채점을 시도하였다. 그러나 이들 연구는 본 연구에서 다루는 것과 같이 장문의 서술형 답안을 자동채점하는 문제를 다루지는 않았다.

본 연구에서 다루는 서술형 답안은 기존의 서답형 문항 자동채점 연구에서 다루던 답안보다 길이가 훨씬 긴 논술형 답안이기 때문에, 정답 템플릿이나 벡터 공간 모델에 기반한 기계학습 기법을 이용하는 기존 연구 방법으로는 채점에 한계가 있다. 이에 본 연구에서는 사전학습된(pretrained) 심층학습(deep learning) 한국어 언어모델을 활용하여 서술형 답안을 자동으로 채점하고자 한다.

심층학습은 데이터가 충분히 있을 때 주로 효과적인 기계학습 기법이지만, 사전학습모델(pretrained model)을 미세조정(fine-tuning)하는 방법으로, 비교적 적은 데이터로도 의미 있는 결과를 얻을 수도 있다. 우리는 이 논문에서 학습 데이터가 많지는 않지만, 심층학습기반 한국어 언어모델(language model)인 KoBERT[1], KcBERT[2], KR-BERT[3]의 세 가지 모델을 미세조정하는 방법으로 한국어 쓰기 답안지의 점수 구간을 예측해 보고, 그 성능을 확인한다. 또, 이들 성능을 확률적 기계학습 분류기(probabilistic machine learning classifier)인 나이브 베이즈(naive Bayes, 이후 NB)와 로지스틱 회귀(logistic regression, 이후 LR)의 성능과도 비교해 본다. 실험에서 사용할 데이터는 유학생이 작성한 한국어 쓰기 답안지로, 네 가지 주제(‘직업’, ‘행복’, ‘경제’, ‘성공’)에 대해 작문한 텍스트이며, 이 답안지를 네 개의 점수 구간(A, B, C, D)으로 분류하는 실험을 시행한다.

2. 관련 연구

유학생이 작성한 한국어 쓰기 답안지를 사람이 일일이 채점할 경우, 채점자의 주관이 개입될 수 있고, 오랜 시간 동안 사람이 채점할 경우 피로가 쌓일 수 있어 대책이 필요하다. 이때 컴퓨터가 한국어 쓰기 답안지를 정확하게 자동채점할 수만 있다면 짧은 시간에 많은 답안지를 채점할 수 있고, 일관된 채점 결과를 얻을 수 있어서 사람에게 도움이 될 것이다. 여기서 자동채점이란 컴퓨터가 정확한 점수(score)를 예측하는 것을 의미할 수도 있으나, 이 연구에서는 주어진 글쓰기 답안지를 평어(grade)나 정해진 점수 구간(score range)으로 자동분류하는 것을 뜻한다.

기존의 컴퓨터를 이용한 한국어 서답형 문항의 자동채점 연구는 크게 채점 기준을 정답 템플릿에 정의하는 방식[4,5]과 다양한 기계학습 기법을 적용하는 방식[6-8]으로 나뉜다. 그런데 이들 연구가 다루는 한국어 서답형 답안은 몇 개의 단어로 구성된 단답형 답안이나 한두 개의 문장으로 구성된 답안이다. [4]는 채점 전문가가 정답 일치 채점, 개념 기반 채점 등의 정답 템플릿을 정의하여 단어 또는 구 수준의 한국어 서답형 문항을 자동채점하는 방법을 제안하면서 모범 답안에 기반한 채점 결과와 자동채점 결과의 일치도를 카파 계수(kappa coefficient)로 측정하였다. [5]에서는 [4]의 정답 템플릿을 활용하여 사용자 인터페이스가 제공되는 자동채점 시스템을 제안하였다. [6]은 답안으로부터 문장 개수, 어휘 개수, 키워드 유무, 키워드 비율 등 다양한 자질(feature)을 추출한 뒤, 벡터 지지 기계(support vector machine), 결정 트리(decision tree), 최대 엔트로피(maximum entropy)를 학습한 후 투표기법을 이용하여 자동채점의 성능을 평가하였다. [7]은 분류기 학습을 위한 자질을 추출하고, 로지스틱 회귀와 k-NN (k-nearest neighbor) 분류기를 이용하여 미채점 답안을 자동채점한 후, 로지스틱 회귀의 분류 확률이 임곗값(threshold)보다 높으면서 로지스틱 회귀와 k-NN 분류기의 분류 결과가 일치하는 답안을 훈련 데이터로 추가하는 형식의 준지도학습 자동채점 방법을 제시하였다. [8] 또한 세 개의 분류기(로지스틱 회귀, nearest centroid, AdaBoost)를 학습한 후, 로지스틱 회귀의 분류 확률이 임곗값보다 크면서 세 개의 분류기의 결과가 만장일치 하는 경우를 정답으로 채택하였다. 그러나 기존 연구의 대부분은 단어, 구, 문장 등 비교적 짧은 주관식 답안을 자동채점하고 있어, 본 연구와 같이 여러 문단으로 구성된 장문의 서술형 답안을 다루지는 않았다.

한편 장문의 답안을 채점하는 또 다른 방법으로, 주어진 문서를 심층학습 언어모델을 이용하여 자동분류하는 방법을 생각할 수 있는데, 지금까지 한국어 심층학습 언어모델을 텍스트 분류에 이용한 기존 연구로는 KoBERT 언어모델로 16만 건의 한국어 온라인 상품평을 긍정 상품평과 부정 상품평으로 자동분류하여 90% 후반대의 정확도를 낸 연구[9]와, 한국어 기술문서 7,108건을 KoBERT를 이용하여 33개 국가 R&D 과제 중분류 코드로 분류하여 0.5 이상의 F-score를 기록한 연구[10]가 있다. 또 KoBERT를 이용하여 네이버 영화평의 긍정/부정 감성분석을 시행하거나[11], KoBERT를 네이버 영화평과 뉴스 댓글로 미세조정한 뒤, YouTube 댓글의 긍정/부정 감성을 분류하거나[12], KoBERT를 미세조정하여 단발성 대화 데이터셋의 일곱 가지 감정(중립, 분노,

1) https://github.com/SKTBrain/KoBERT
2) https://github.com/Beomi/KcBERT
3) https://github.com/snunlp/KR-BERT

Table 1. Dataset

Score Range	Grade	Job	Happiness	Economic	Success	Total
24-30	A	24	28	8	10	70
18-23	B	45	53	26	13	137
15-17	C	17	6	13	5	41
6-14	D	15	9	15	17	56
Total		101	96	62	45	304

혐오, 공포, 행복, 슬픔, 놀람)에 대한 감정분석을 진행하거나[13], KoBERT를 이용하여 15개의 일상 대화의 주제를 분류하거나[14], 한국어 가사 데이터로 KoBERT를 미세조정한 뒤 사랑 노래와 이별 노래를 자동분류한[15] 연구가 있다.

그러나 한국어 쓰기 답안지의 자동채점에 한국어 심층학습 언어모델을 적용한 연구는 우리가 발표한 이전 연구[16] 이외에는 우리가 조사한 바로는 없었으며, 또 수천에서 수만 개의 학습 데이터를 이용한 기존 연구와 달리, 이 연구처럼 500건 이하의 적은 양의 문서 데이터로 KoBERT, KcBERT, KR-BERT를 미세조정한 후, 문서 자동분류 성능을 살펴본 연구는 발견하지 못했다. 실세계의 산업 현장에서는 충분한 양의 데이터를 확보하지 못하는 경우가 많아서, 이 연구는 그런 현장에서 참고가 되는 결론을 제시할 수 있을 것으로 보인다. 이후 우리는 이 논문에서 사전학습된 세 가지 한국어 언어모델인 KoBERT, KcBERT, KR-BERT를 이용하여 유학생의 한국어 쓰기 답안지를 네 개의 점수 구간으로 자동 분류하는 텍스트 분류문제를 다룬다.

3. 실 험

3.1 데이터

실험에 사용한 유학생 한국어 쓰기 답안지는 주제별로 '직업의 조건(직업)' 100편, '행복의 조건(행복)' 95편, '경제와 행복(경제)' 61편, '성공의 기준(성공)' 44편과 교수자가 주제별로 작성한 모범 답안 4편으로 모두 304편이었다. 학습 데이터를 하나라도 더 늘리기 위해 채점자가 작성한 모범 답안도 데이터 세트에 포함하였다.

하나의 답안지는 0점부터 30점까지의 점수를 가질 수 있는데, 실제로 실험에 사용한 답안지는 최저 점수가 6점, 최고 점수가 30점이었다. 우리는 이 답안지들을 Table 1과 같이 네 개의 평어(A, B, C, D)로 구분하여 실험에 사용하였다. 평어에 대응하는 점수 구간을 Table 1(Score Range)처럼 정한 이유는 [16]을 참조하기 바란다. 유학생이 작성한 답안지의 예를 Fig. 1에 제시한다. 답안지의 오른쪽 맨 아래를 보면 채점자가 답안지를 27점(8-6-6-4-3/27)으로 평가하고 있다.

3.2 비교 모델

이번 유학생 한국어 쓰기 답안지의 점수 구간 예측 실험에서는 총 7개의 자동분류 모델을 구축하여 분류 정확도를 조사한다.

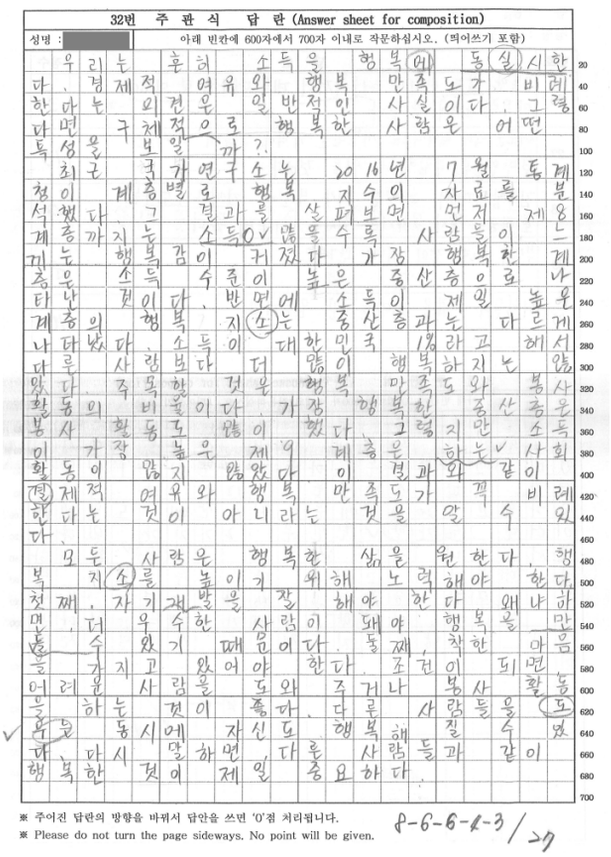

Fig. 1. Sample Handwritten Korean Essay

a) KoBERT: SK텔레콤이 개발한 한국어 딥러닝 언어모델이다. 한국어의 분석, 이해, 활용에 특화됐다. 영어를 위해 개발된 BERT[17]의 오픈소스 트랜스포머 아키텍처를 활용하고 있다. 약 9천 2백 만개의 파라미터를 가진다.

b) KcBERT[18]: 네이버 뉴스 댓글과 대댓글을 수집해 토크나이저와 BERT 모델을 학습한 모델로, 신조어와 구어체에 강인한 사전학습 언어모델이다. 실험에서는 1억 8백만 개의 파라미터를 갖는 bert-base 모델을 사용하였다.

c) KR-BERT[19]: 한국어 위키피디아와 신문 기사를 학습한 모델로 9천 9백만 개의 파라미터를 가진다.

d) 나이브 베이즈(NB): 클래스(class) 내 단어 간의 조건부 독립을 가정하는 확률기반 텍스트 분류 모델이다. 훈련 데이터가 적어도 비교적 준수한 성능을 내는 모델로 알려져 있다[20]. 이번 실험에서도 적은 양의 훈련 데이터로 분류 모델을 학습하는데, NB가 과연 다른 모델에 비해 유리한지를 확인하려고 한다.

e) NB': 상기 d)와는 다르게 형태소분석기로 특징 단어를 추출하지 않고, KR-BERT의 단어 목록(vocabulary)으로 특징 단어를 추출한 후 분류기를 구축한다.

f) 로지스틱 회귀(LR): 선형 회귀와 시그모이드 또는 소프트맥스 함수의 계산을 통해, 데이터가 어떤 범주에 속할

확률을 계산하고, 이들 중 최고의 확률을 가지는 범주로 데이터를 분류하는 모델이다. 분류 모델의 가중치(weight)를 참고하여 중요한 특징 단어를 확인할 수 있다. 클래스별로 중요한 단어가 무엇인지를 구축된 모델로 확인한다.

g) LR': 상기 e)와 같이 KR-BERT의 단어 목록을 이용하여 LR' 분류기를 구축하여 실험한다.

3.3 실험 구성

평가를 위한 테스트 데이터로 우리는 크게 '직업(job)', '행복(happ)', '통합(all)'의 세 가지 데이터를 사용하였다. 여기서 '통합'은 네 개의 주제['직업', '행복', '경제(econ)', '성공(succ)']의 데이터를 혼합한 데이터를 가리킨다. 우리는 다음과 같이 다양한 주제의 데이터를 섞어 훈련 데이터를 만들고, 훈련 데이터로 분류 모델을 미세조정한 뒤, 테스트 데이터로 분류 모델의 점수 구간 예측 성능을 조사하였다.

① [훈련: '직업'] → [테스트: '직업']
② [훈련: '직업+경제'] → [테스트: '직업']
③ [훈련: '직업+성공'] → [테스트: '직업']
④ [훈련: '직업+행복'] → [테스트: '직업']
⑤ [훈련: '직업+경제+성공+행복'] → [테스트: '직업']
⑥ [훈련: '행복'] → [테스트: '행복']
⑦ [훈련: '행복+경제'] → [테스트: '행복']
⑧ [훈련: '행복+성공'] → [테스트: '행복']
⑨ [훈련: '행복+직업'] → [테스트: '행복']
⑩ [훈련: '행복+경제+성공+직업'] → [테스트: '행복']
⑪ [훈련: '통합'] → [테스트: '통합']

이렇게 훈련 데이터를 다양하게 설정한 이유는, 서로 다른 주제의 한국어 쓰기 데이터를 혼합하여 언어모델을 미세조정했을 때, 분류 성능이 과연 향상하는지를 확인하기 위해서이다. 만약 이러한 방법이 도움이 된다면, 데이터가 불충분한 상황에서 데이터를 혼합하는 것이 하나의 해결책이 될 수 있다는 결론을 얻을 수 있다.

각 실험은 학습 데이터의 양이 적은 점을 고려하여 7-겹 교차검증(7-fold cross validation)으로 시행되었다. 매 실험에서는 72%를 훈련(training) 데이터로, 14%를 검증(validation) 데이터로, 14%를 테스트(test)로 설정했다.

3.4 평가 척도

평가척도로는 7-겹 교차검증 결과의 평균 정확도(average accuracy)를 사용하였다. 각 실험의 정확도(accuracy)는 Equation (1)로 계산하였다.

$$\text{정확도}(\%) = \frac{\text{올바르게 예측한 테스트 데이터의 개수}}{\text{전체 테스트 데이터의 개수}} \times 100 \quad (1)$$

7-겹 교차검증의 평균 정확도는 Equation (2)와 같이 계산하였다.

$$\text{평균 정확도} = \frac{1}{7}\sum_{k=1}^{7} \text{정확도}_k \quad (2)$$

3.5 모델 학습

한국어 심층학습 기반 언어모델은 맨 마지막에 다차원의 임베딩 벡터(embedding vector)를 출력하는데, 우리는 여기에 완전 연결 계층(fully-connected layer)을 연결한 후, 모델이 과적합(overfit) 되는 것을 방지하기 위해 dropout (0.5)을 적용하고 전체 언어모델을 미세조정하였다. 사전학습된 언어모델의 가중치를 고정(freeze)한 후 완전 연결 계층만 미세조정하는 방법도 실험해봤으나, 전체 모델을 미세조정하는 쪽이 성능이 나아서 전체 모델을 미세조정하였다.

모델의 최적화 함수에는 AdamW optimizer를, 손실함수(loss function)로는 cross entropy loss를 사용했고, 학습률(learning rate)은 1e-5로 설정하였다. 또 batch size는 24, epoch는 50으로 정했는데, epoch마다 검증 데이터로 분류 모델의 성능을 확인한 후, 검증 데이터의 성능이 최고가 되는 분류 모델의 테스트 데이터 분류 성능을 취합하여 그 평균을 계산했다(7-겹 교차검증의 평균). 언어모델의 미세조정에는 HuggingFace의 transformer[4] 패키지를 사용하였다. 실험의 재현성을 보장하기 위해 seed를 설정하였다.

한편, NB와 LR의 경우, scikit-learn[5]의 기본 파라미터값으로 모델을 구축하고 7-겹 교차검증을 시행하였다. 이때 답안지 속 단어들을 KoNLPy[6]의 Komoran 형태소분석기로 토큰화하고 scikit-learn의 CountVectorizer로 단어 빈도를 계산하여 문서-단어 행렬을 만들었다. 이 문서-단어 행렬의 특징 단어의 개수는 적게는 1,100여 개, 많게는 2,600여 개였다. 더불어 KR-BERT 모델의 토크나이저를 이용하여 특징 단어를 추출하고 NB'와 LR'의 실험도 진행하였다(3.2 비교 모델 e와 g). 실험 장비로는 32GB 메모리와 RTX 3090 GPU를 내장한 컴퓨터를 사용하였다.

3.6 실험 결과

'직업', '행복', '통합' 테스트 데이터에 대한 실험 결과를 각각 ['직업': Fig. 2, Table 2], ['행복': Fig. 3, Table 3], ['통합': Fig. 4, Table 4]에 정리한다. Table 2와 Table 3의 분류기별 최고 정확도와 분류기들 중 최고의 5개 실험 평균 정확도를 굵은 글씨로 표시했다.

먼저 '직업' 테스트 데이터 분류 실험[Fig. 2, Table 2]에서는 전반적으로 KcBERT가 상대적으로 높은 예측 성능을 보였는데, 실험 ①~⑤의 평균 정확도가 47.65%로 가장 높았다. 가장 높은 예측 성능을 보인 개별 모델도 KcBERT였는데, '⑤직업+경제+성공+행복' 훈련 데이터로 미세조정한

4) https://huggingface.co/transformers/
5) https://scikit-learn.org/stable/
6) https://konlpy.org/en/latest/

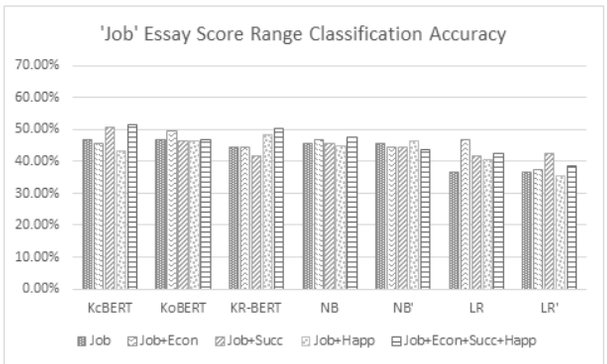

Fig. 2. 'Job' Essay Score Range Classification Accuracy

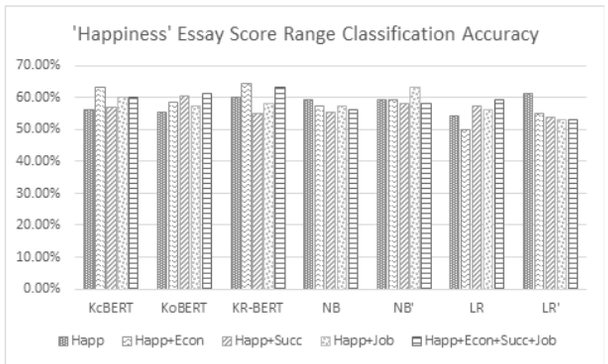

Fig. 3. 'Happiness' Essay Score Range Classification Accuracy

Table 2. 'Job' Essay Score Range Classification Accuracy

	KcBERT	KoBERT	KR-BERT	NB	NB'	LR	LR'
Job	46.67%	46.67%	44.49%	45.65%	45.51%	36.60%	36.53%
Job+Econ	45.71%	**49.66%**	44.49%	46.80%	44.49%	**46.67%**	37.62%
Job+Succ	50.75%	46.60%	41.70%	45.51%	44.63%	41.63%	**42.52%**
Job+Happ	43.47%	46.60%	48.50%	44.69%	**46.53%**	40.48%	35.58%
Job+Econ+Succ+Happ	**51.63%**	46.67%	**50.54%**	**47.62%**	43.74%	42.65%	38.64%
Average	**47.65%**	47.24%	45.95%	46.05%	44.98%	41.61%	38.18%

Table 3. 'Happiness' Essay Score Range Classification Accuracy

	KcBERT	KoBERT	KR-BERT	NB	NB'	LR	LR'
Happ	56.12%	55.34%	60.28%	**59.26%**	59.26%	54.24%	**61.46%**
Happ+Econ	**63.34%**	58.40%	**64.60%**	57.22%	59.26%	50.00%	55.18%
Happ+Succ	57.14%	60.36%	55.10%	55.26%	58.24%	57.30%	54.00%
Happ+Job	60.20%	57.38%	58.24%	57.22%	**63.42%**	56.36%	52.98%
Happ+Econ+Succ+Job	60.13%	**61.46%**	63.27%	56.20%	58.16%	**59.42%**	52.98%
Average	59.39%	58.59%	**60.30%**	57.03%	59.67%	55.46%	55.32%

Table 4. Overall Score Range Classification Accuracy

	KcBERT	KoBERT	KR-BERT	NB	NB'	LR	LR'
Job (Max.)	**51.63%**	49.66%	50.54%	47.62%	46.53%	46.67%	42.52%
Happiness (Max.)	63.34%	61.46%	**64.60%**	59.26%	63.42%	59.42%	61.46%
All (Exp.⑪)	52.32%	**53.60%**	52.35%	50.66%	47.69%	44.75%	43.40%
Overall Average	55.77%	54.91%	**55.83%**	52.52%	52.55%	50.28%	49.13%

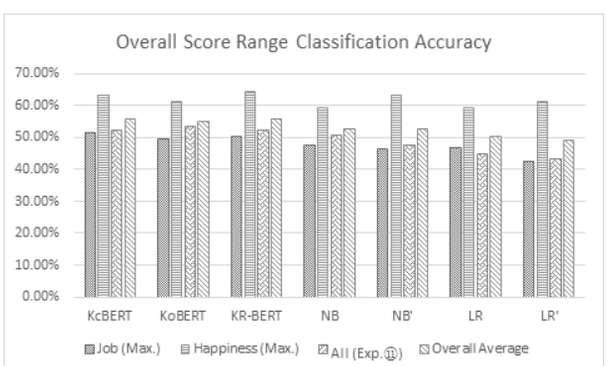

Fig. 4. Overall Score Range Classification Accuracy

KcBERT가 51.63%의 정확도를 나타냈다. 심층학습 언어모델 분류기가 50% 정도의 정확도를 나타낸 데 비해 NB나 LR 분류기는 46~47%대의 정확도를 나타냈다. 특히 LR이 전체적으로 낮은 성능을 보였다.

'행복' 테스트 데이터의 분류 실험[Fig. 3, Table 3]에서는 실험 ⑥~⑩의 평균 정확도가 60.30%인 KR-BERT가 전반적으로 높은 성능을 보였다. 또 개별 분류기 중에서도 '⑦행복+경제' 훈련 데이터로 미세조정한 KR-BERT가 64.60%로 가장 높은 정확도를 보였다. 한편 KcBERT와 NB' 분류기도 각각 63.34%와 63.42%로 높은 성능을 나타냈다. 특기할만한 점은 KR-BERT에서 추출한 단어 목록으로 구축한 NB'가 63.42%

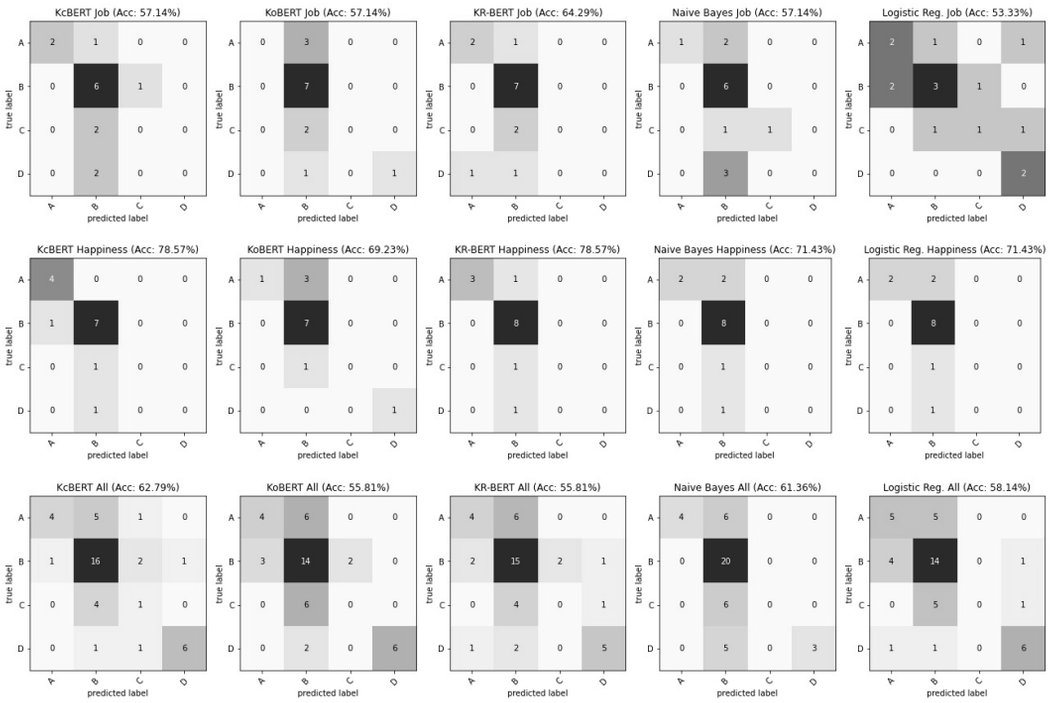

Fig. 5. Confusion Matrix Analyses (Top: 'Job' Result, Middle: 'Happiness' Result, Bottom: 'All' Result)

로 심층학습 언어모델 분류기와 비슷한 성능을 나타냈다는 점이다. 이에 비해 형태소 분석기를 이용하여 전통적인 방법으로 구축한 NB는 59.26%로 상대적으로 낮은 성능을 보였다.

'통합(All)' 테스트 데이터 분류 실험[Fig. 4, Table 4]에서는 KoBERT가 53.60%로 가장 높은 성능을 나타냈다. LR'이 KoBERT보다 10% 넘게 차이가 나면서 43.40%로 가장 낮은 성능을 보였다. 일곱 개의 분류기에 대하여 '직업'과 '행복' 실험에서 가장 높은 예측 성능을 선택한 후 '통합(All)' 결과를 포함하여 세 가지 정확도의 평균을 낸 결과, KR-BERT가 55.83%로, 55.77%인 KcBERT보다 약간 더 우세한 성능을 나타냈다.

한편 한국어 쓰기의 주제에 따라 성능에 큰 차이가 나는 것을 확인하였는데, '행복' 테스트 데이터 분류 실험(⑥~⑩)이 '직업' 테스트 데이터 분류 실험(①~⑤)보다 10% 넘게 더 높은 분류 정확도를 나타냈다.

3.7 혼동 행렬 분석

이번 연구에서는 실험 데이터 총 304개 중 평어가 B인 데이터가 137개로 가장 많아(45.1%) 데이터의 분포가 균등하지 않다는 문제가 있었다. 이에 개별 평어(A, B, C, D)가 어떻게 분류되고 있는지를 혼동행렬(confusion matrix)을 통해 확인하였다. Fig. 5는 다섯 개의 분류기(왼쪽부터 KcBERT, KoBERT, KR-BERT, NB, LR)와 '직업(Job)', '행복(Happiness)', '통합(All)' 테스트 데이터의 분류 결과(위에서부터)를 각각 혼동행렬로 나타낸 그림이다. 실험에서는 다양한 훈련 데이터로 7-겹 교차검증을 시행했기 때문에, '직업', '행복', '통합' 테스트 데이터의 분류에서 분류기 전반적으로 비교적 높은 성능을 보인 개별 실험 결과를 하나씩 선택하여 분류기별로 각 평어의 분류 성능을 살펴보았다.

분석 결과 가장 많은 데이터를 구성하는 평어 B가 역시 분류 결과에 영향을 끼치고 있음을 알 수 있었다. 그러나 평어 B 데이터의 절반 정도 되는 평어 A 데이터(70개)도 어느 정도 올바르게 분류되고 있어, 단순히 평어 B로 모든 데이터를 분류하고 있지는 않았다. 또 '통합(All)' 테스트 데이터에서는 평어 D 데이터도 일부 올바르게 분류되고 있어, 평어별 데이터를 균등하게 늘린 후 분류기를 다시 구축하면 더 높은 예측 성능을 기대할 수 있을 것으로 보인다.

3.8 특징 단어 비교

총 열 한 개의 실험 중 일곱 개의 분류기의 성능의 표준편차가 가장 작았던 실험은 '⑦ [훈련: '행복+경제'] → [테스트: '행복']' 실험이었는데[7], 이러한 결과가 나왔을 때 과연 어떤 단어가 중요하게 작용했는지를 확인하기 위해 로지스틱 회귀 분류기에서 높은 가중치를 가졌던 상위 10개의 특징 단어를 클래스(A, B, C, D)별로 추출하여 살펴보았다.

Table 5에는 형태소분석기로 추출한 특징 단어를, Table 6에는 KR-BERT의 토크나이저로 추출한 특징 단어를 제시하였다. 두 경우 모두 예측 정확도가 57.14%였던 실험 회차를 골랐는데, 해당 실험 회차에서의 특징 단어의 총 개수는 각각 1,595개(형태소분석기)와 2,380개(KR-BERT 토크나이저)였다.

Table 5를 보면 평어 C와 평어 D의 특징 단어로 '그래서'와 '라서'라는 구어적 표현이 들어가 있는 것을 볼 수 있다.

7) 〈Table 3〉 Happ+Succ의 분류기 성능들의 표준편차는 0.0217로 열 한 개의 실험 결과 중 가장 작았다.

Table 5. Top 10 Weighted Feature Words for 'Happiness' Essay Classification Using Logistic Regression [LR] (Words Extracted from Morphological Analyzer)

Class	A	B	C	D
Feature Word	어떤	생활	ㄴ다	아서
	느끼	자신	자기	ㅂ니다
	기본	생각	에게	부모
	과정	아니	그래서	실패
	ㄴ다면	명예	프로그램	사이
	욕심	다른	품행	라서
	조건	좋아하	세상	필요
	는지	가족	어야	즐겁
	태도	라고	이런	예상
	능력	마다	문제	없이

Table 6. Top 10 Weighted Feature Words for 'Happiness' Essay Classification Using Logistic Regression [LR'] (Words Extracted from KR-BERT Tokenizer)

Class	A	B	C	D
Feature Word	나는	생활	##하고	##고
	##에	자신의	##하게	마음
	삶	행복	좋은	##분
	##적인	생각한다	한다	##다
	어떤	일	자기	##니다
	경제적	명예	##은	필요한
	조건	목표	어려운	##은
	##감	게	##면	##서
	필요하다	자신이	##를	위해서
	이루	##이라고	있다	부모님

그러나 전반적으로 Table 5와 Table 6 모두 평어별 상위 10위의 높은 가중치를 가지는 단어들 사이에 두드러진 차이는 없었다. 이번 문서 자동분류의 목적은 한국어 쓰기 수준을 평가하는 것이기 때문에, 오탈자, 문법 오류, 표현 오류 등의 오류 표현들이 특징 단어로 취급되는 것이 필요했다. 그러나 실험에서는 형태소분석기를 이용하여 기본적인 단어를 취득하거나 언어모델의 기본적인 토크나이저로 특징 단어를 추출했기 때문에, 오류 표현을 제대로 다루지 못했다. 따라서 한국어 심층학습 기반 언어모델의 분류 성능이 예상보다 낮았던 이유는 부족한 학습 데이터 때문이기도 하지만, 이처럼 오류 표현을 제대로 다루지 못한 부분에서 찾을 수 있다. 왜냐하면, 사전학습된 언어모델은 오류 표현이 거의 없는 위키피디아나, 구어체이면서 단문의 댓글과 같은 텍스트로부터 구축되었기 때문이다. 이들 텍스트에서는 유학생의 글쓰기 텍스트가 가지는 오류 표현들을 찾기 어렵다. 따라서 앞으로 더 정확한 글쓰기 자동 평가를 실현하려면, 더 많은 오류 표현 데이터가 필요하고, 이러한 데이터로부터 오류 표현을 반영한 단어 목록(vocabulary)을 추출하여 언어모델을 학습할 필요가 있다.

4. 결 론

데이터가 불충분한 상황에서 심층학습 기법을 이용하는 것은 별로 권장할만한 방법은 아니지만, 우리는 이 논문에서 세 종류의 심층학습 기반이면서 BERT에 기반한 한국어 사전학습 모델을 미세조정함으로써 적은 데이터로도 나이브베이즈나 로지스틱 회귀보다 조금 더 높은 성능을 보이는 심층학습 텍스트 분류 모델을 구축할 수 있음을 확인하였다.

K-팝, K-드라마, 한국 영화 등 한류 콘텐츠가 세계적으로 호응을 얻으면서 한국어를 외국어 또는 제2 언어로 배우려는 학습자가 늘고 있는 가운데 한국어 쓰기 데이터의 자동채점 시스템의 구현은 한국어 확산에 큰 도움을 줄 수 있을 것으로 기대된다. 우리는 앞으로 오류 표현을 명시적으로 정의하여 적은 데이터로도 더 정확하게 글쓰기 점수 구간을 분류할 수 있는 글쓰기 자동채점 방법에 관해 계속 연구하려고 한다.

본 연구에서 사용된 데이터는 연구진의 소속기관 홈페이지[8]에서 내려받을 수 있으며, 실험에서 사용한 Python 코드는 연구자의 GitHub[9] 저장소에서 확인할 수 있다.

References

[1] H. Cho, H. Im, J. Cha, and Y. Yi, "Comparison of automatic score range prediction of Korean essays using KoBERT, Naive Bayes & Logistic Regression," in *Proceedings of the KIPS Spring Conference 2021*, Vol.28, No.1, pp.501-504, 2021.

[2] S. Yoo and K. Yang, "The status of Korean as an international language," *Hallyu Now: Global Hallyu Issue Magazine*, Vol.34, pp.9-16, 2020.

[3] H. J. Park and W. S. Kang, "Design and implementation of a subjective-type evaluation system using natural language processing technique," *The Journal of Korean Association of Computer Education*, Vol.6, No.3, pp.207-216, 2003.

[4] I.-N. Park, S.-S. Kang, E.-H. Noh, M.-H. Kim, and T.-J. Seong, "Automatic scoring of Korean short answers by answer template description," *Journal of KIISE: Computing Practices and Letters*, Vol.19, No.12, pp.630-636, 2013.

[5] S. S. Kang and E. S. Jang, "Automatic scoring system for Korean short answers by student answer analysis and answer template construction," *KIISE Transactions on Computing Practices*, Vol.22, No.5, pp.218-224, 2016.

[6] J. Heo and S.-Y. Park, "Design and implementation of an automatic scoring model using a voting method for descriptive answers," *Journal of the Korea Society of Computer and Information*, Vol.18, No.8, pp.17-25, 2013.

[7] M.-A. Cheon, H.-W. Seo, J.-H. Kim, E.-H. Noh, K.-H. Sung, and E. Young Lim, "Semi-automatic scoring for short Korean free-text responses using semi-supervised learning," *Korean Journal of Cognitive Science*, Vol.26, No.2, pp.147-165, 2015.

8) http://aihumanities.org/ko/archive/data/?vid=1
9) https://github.com/heeryoncho/three_korean_bert_LM_comparison

[8] M.-A. Cheon, C.-H. Kim, J.-H. Kim, E.-H. Noh, K.-H. Sung, and M.-Y. Song, "Automated scoring system for Korean short-answer questions using predictability and unanimity," *KIPS Transactions on Software and Data Engineering*, Vol.5, No.11, pp.527-534, 2016.

[9] J.-Y. Choi and H.-S. Lim, "E-commerce data based Sentiment analysis model implementation using natural language processing model," *Journal of the Korea Convergence Society*, Vo.11, No.11, pp.33-39, 2020.

[10] S. Hwang and D. Kim, "BERT-based classification model for Korean documents," *The Journal of Society for e-Business Studies*, Vol.25, No.1, pp.203-214, 2020.

[11] T.-H. Kim, D.-B. Cho, H.-Y. Lee, H.-J. Won, and S.-S. Kang, "Sentiment analysis system by using BERT language model," in *Proceedings of the KIPS Spring Conference 2020*, Vol.27, No.2, pp.975-977, 2020.

[12] S. Park, H. Yang, M. Choe, M. Ha, K. Chung, and M. Koo, "Sentimental analysis of YouTube Korean comments using KoBERT," in *Proceedings of Korea Software Congress 2020*, pp.1385-1387, 2020.

[13] Y.-J. Lee and H.-J. Choi, "Joint Learning-based KoBERT for emotion recognition in Korean," in *Proceedings of Korea Software Congress 2020*, pp.568-570, 2020.

[14] K. H. Park and Y.-S. Jeong, "Korean daily conversation topics classification using KoBERT," in *Proceedings of Korea Software Congress 2021*, pp.1735-1737, 2021.

[15] A.-G. Kim and Y.-S. Jeong, "Topic classification of domestic music using KoBERT," in *Proceedings of Korea Software Congress 2021*, pp.1738-1740, 2021.

[16] H. Cho, Y. Yi, H. Im, J. Cha, and C. Lee, "Automatic score range classification of Korean essays using deep learning-based Korean language models -The case of KoBERT & KoGPT2-," *Journal of the International Network for Korean Language and Culture*, Vol.18, No.1, pp.217-241, 2021.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp.4171-4186, 2019.

[18] J. Lee, "KcBERT: Korean comments BERT," in *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pp.437-440, 2020.

[19] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, "KR-BERT: A small-scale Korean-specific language model," ArXiv, 2020. [Internet], https://arxiv.org/abs/2008.03979.

[20] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp.26-33, 2001.

조 희 련

https://orcid.org/0000-0001-9912-1002
e-mail : heeryon@cau.ac.kr
1995년 연세대학교 신문방송학과(학사)
2009년 교토대학 사회정보학과(석・박사)
2020년 ~ 현 재 중앙대학교 인문콘텐츠연구소 HK교수
관심분야 : 텍스트 분석, 기계학습 응용, 인공지능인문학

임 현 열

https://orcid.org/0000-0003-0628-0229
e-mail : languages@cau.ac.kr
2001년 중앙대학교 국어국문학과(학사)
2011년 중앙대학교 국어국문학과 (석・박사)
2011년 ~ 현 재 중앙대학교 다빈치교양대학 조교수
관심분야 : 국어음운론, 한국어교육학, 인공지능인문학

이 유 미

https://orcid.org/0000-0002-1230-041X
e-mail : joystu@cau.ac.kr
1998년 중앙대학교 국어국문학과(학사)
2001년 중앙대학교 국어국문학과(석사)
2006년 중앙대학교 국어국문학과(박사)
2019년 ~ 현 재 중앙대학교 인문콘텐츠연구소 부교수
관심분야 : 인공지능인문학, 휴먼커뮤니케이션

차 준 우

https://orcid.org/0000-0003-3117-5148
e-mail : yunu77@naver.com
2001년 중앙대학교 국어국문학과(학사)
2014년 중앙대학교 국어국문학과(석사)
2021년 중앙대학교 국어국문학과(박사수료)
2014년 ~ 현 재 중앙대학교 한국어교육원 강사
관심분야 : 한국어 교육, 한국어 화용 교육, 한국어 쓰기 평가