KoBERT, 나이브 베이즈, 로지스틱 회귀의 한국어 쓰기 답안지 점수 구간 예측 성능 비교1)

조희련*, 임현열**, 차준우*, 이유미*
*중앙대학교 인문콘텐츠연구소
**중앙대학교 다빈치교양대학
heeryon@cau.ac.kr, languages@cau.ac.kr, junu77@cau.ac.kr, joystu@cau.ac.kr

Comparison of Automatic Score Range Prediction of Korean Essays Using KoBERT, Naive Bayes & Logistic Regression

Heeryon Cho*, Hyeonyeol Im**, Junwoo Cha*, Yumi Yi*
*Humanities Research Institute, Chung-Ang University
**Da Vinci College of General Education, Chung-Ang University

요 약

한국어 심층학습 언어모델인 KoBERT와, 확률적 기계학습 분류기인 나이브 베이즈와 로지스틱 회귀를 이용하여 유학생이 작성한 한국어 쓰기 답안지의 점수 구간을 예측하는 실험을 진행하였다. 네 가지 주제('직업', '행복', '경제', '성공')를 다룬 답안지와 점수 레이블(A, B, C, D)로 쌍을 이룬 학습 데이터 총 304건으로 다양한 자동분류 모델을 구축하여 7-겹 교차검증을 시행한 결과 KoBERT가 나이브 베이즈나 로지스틱 회귀보다 약간 우세한 성능을 보였다.

1. 서론

심층학습(deep learning)은 충분한 데이터가 있을 때 유용한 기계학습 기법이지만, 상대적으로 적은 데이터로도 사전학습모델(pretrained model)을 미세조정(fine-tuning)하는 것으로 유용한 결과를 얻을 수도 있다. 이 논문에서는 학습 데이터가 적은 상황에서, 한국어 심층학습 언어모델(language model)인 KoBERT를 한국어 쓰기 답안지의 점수 구간 예측을 위해 미세조정하는 것이 어느 정도의 성능을 나타내는지를 확인하고, 확률적 기계학습 분류기(probablistic machine learning classifier)인 나이브 베이즈(naive Bayes, 이후 NB)와 로지스틱 회귀(logistic regression, 이후 LR)의 성능과 비교해 본다. 실험에서 다루는 데이터는 유학생이 작성한 한국어 쓰기 답안지로, 네 종류의 주제('직업', '행복', '경제', '성공')를 다루고 있으며, 이를 네 개의 점수 구간(A, B, C, D)으로 분류하는 실험을 진행한다.

1) 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A6A3A01078538).

2. 배경

채점자가 유학생이 쓴 한국어 쓰기 답안지를 하나하나 채점하는 것은 채점자의 주관성(subjectivity)과 피로도를 피할 수 없다는 점에서 보완의 여지가 있다. 만약 컴퓨터로 한국어 쓰기 답안지를 정확하게 자동채점할 수 있다면 단시간에 많은 답안지를 채점할 수 있는 것은 물론 일관된 채점 결과를 얻을 수 있어 채점자에게 도움을 줄 수 있을 것이다. 여기서 자동채점이란 컴퓨터가 정확한 점수(score)를 예측하는 것을 뜻할 수도 있으나, 이 논문에서는 주어진 답안지를 정해진 점수 구간(score range)이나 평어(grade)로 구분하는 것을 가리킨다. 우리는 이 논문에서 사전학습된 한국어 언어모델인 KoBERT를 이용하여 유학생의 한국어 쓰기 답안지를 네 개의 점수 구간으로 자동 분류하는 텍스트 분류문제를 다루고 그 성능을 확인한다.

지금까지 KoBERT를 텍스트 분류에 사용한 기존 연구로는 ① 16만 건의 전자상거래 상품평을 긍정 상품평과 부정 상품평으로 자동 분류하여 90% 후반대의 정확도를 달성한 연구[1]와, ② 7,108건의 한국

어 기술문서를 33개의 국가 R&D 과제 중분류 코드로 분류하여 0.5 이상의 F-score를 달성한 연구[2]가 있다. 그런데 이 논문에서와같이 아주 적은 데이터(문서 500건 이하)로 KoBERT를 미세조정한 후 문서를 분류하여 성능을 살펴본 연구는 우리가 조사한 바로는 없었다. 실제로 여러 현장에서는 적은 양의 데이터밖에 구할 수 없는 경우가 많아, 이 연구는 그러한 현장에서 참고할 만한 결론을 제시하는 것을 목적으로 한다.

3. 비교 모델

- **KoBERT:** KoBERT는 SK텔레콤이 자체 개발한 한국어의 분석, 이해, 활용에 특화된 한국어 딥러닝 언어모델[2]로, 영어를 위해 개발된 BERT[3]의 오픈소스 인공지능 아키텍처를 활용하고 있다.
- **나이브 베이즈(NB):** NB는 주어진 클래스(class) 내에 있는 단어들의 조건부 독립을 가정하는 확률기반 텍스트 분류 모델로, 적은 양의 데이터로도 준수한 성능을 얻는 모델로 알려져 있다[4]. 이번 실험도 적은 양의 데이터로 분류기를 학습하는 경우여서 NB가 과연 이번에도 다른 모델에 비해 유리한지를 확인하려고 한다.
- **로지스틱 회귀(LR):** LR은 선형 회귀와 시그모이드 (또는 소프트맥스) 함수를 사용하여 데이터가 어떤 범주에 속할 확률을 계산하고, 확률이 더 높은 범주로 분류하는 모델이다. LR은 학습된 가중치로 중요한 특징 단어(feature word)를 발견할 수 있어, 이번 실험에서도 각 클래스(A, B, C, D)에서 중요한 단어가 무엇인지를 확인하려고 한다.

4. 실험

데이터: 실험에 사용한 한국어 쓰기 답안지는 주제별로 '직업의 조건' 100편, '행복의 조건' 95편, '경제와 행복' 61편, '성공의 기준' 44편과 채점자가 작성한 주제별 모범 답안 4편으로 총 304편이었다. 각 답안지가 가질 수 있는 점수의 범위는 0점부터 30점이었고, 답안지의 최저 점수는 6점, 최고 점수는 30점이었다. 이 답안지를 **<표 1>**과 같이 네 개의 점수 구간으로 나눠 실험에 사용하였다.

<표 1> 실험 데이터의 구성

평어	직업	행복	경제	성공	합계	점수 구간
A	24	28	8	10	70	24 - 30
B	45	53	26	13	137	18 - 23
C	17	6	13	5	41	15 - 17
D	15	9	15	17	56	0 - 14
합계	101	96	62	45	304	

점수 구간을 **<표 1>**과 같이 설정한 이유는 [5]에 자세히 나와 있다.

방법: 학습 데이터가 적기 때문에 자동 점수 구간 분류 성능의 통계적인 신뢰도를 높이기 위해 7-겹 교차검증(7-fold cross validation)을 시행했다. 이때 전체 데이터의 14%를 테스트 데이터로, 14%를 검증 데이터로, 나머지 72%를 훈련 데이터로 삼았다.

모델학습: KoBERT는 맨 끝에 768차원의 임베딩 벡터(embedding vector)를 출력하는데, 여기에 완전 연결 계층(fully-connected layer)을 연결한 후, 훈련 데이터에 과적합(overfit)되는 것을 완화하기 위해 dropout으로 0.5 비율을 적용한 후 전체 언어모델을 미세조정했다. 최적화에 AdamW optimizer를, 손실함수(loss function)로 cross entropy loss를 사용했고, 학습률(learning rate)은 0.00001였다. Batch size는 4였고, epoch는 50으로 정의했는데, epoch마다 검증(validation) 데이터로 성능을 확인한 후, 검증 데이터가 최고 성능을 나타내는 분류 모델의 테스트 데이터 예측 성능을 취합하여 평균을 냈다(7-겹 교차검증의 평균). NB와 LR은 scikit-learn[3]의 기본 설정값으로 모델을 구축하고 7-겹 교차검증을 했다. NB와 LR의 입력 데이터에는 KoNLPy[4]의 Komoran 형태소분석기로 먼저 단어를 취득하고 scikit-learn의 CountVectorizer로 단어 빈도를 계산한 문서 벡터를 사용했다. NB와 LR에서의 특징 단어 수는 적게는 1,100여 개, 많게는 2,600여 개였다.

실험 장비로는 RTX 3090 GPU와 32GB 메모리가 탑재된 컴퓨터를 사용하였다.

실험구성: 평가에는 크게 '직업' 테스트 데이터, '행복' 테스트 데이터, '통합' 테스트 데이터의 세 가지 테스트 데이터를 사용했다. 여기서 '통합'은 네 가지

2) https://github.com/SKTBrain/KoBERT

3) https://scikit-learn.org/stable/

4) https://konlpy.org/en/latest/

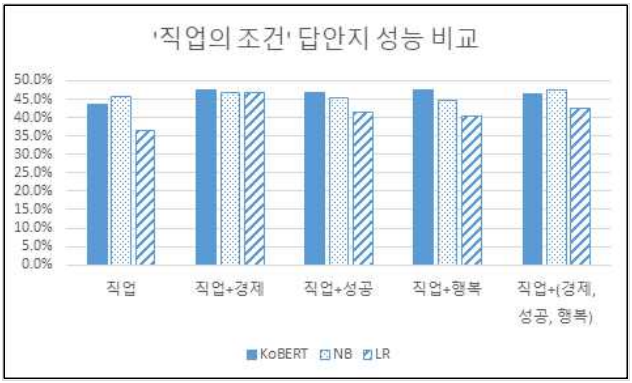
주제('직업', '행복', '경제', '성공')의 데이터를 모두 섞은 데이터를 가리킨다. 이때 다음과 같이 다양한 주제의 훈련 데이터를 혼합하여 분류 모델을 미세조정하고 자동 점수 구간 예측 성능을 알아보았다.

① '직업'(훈련) / '직업'(테스트)
② '직업+경제'(훈련) / '직업'(테스트)
③ '직업+성공'(훈련) / '직업'(테스트)
④ '직업+행복'(훈련) / '직업'(테스트)
⑤ '직업+경제/성공/행복'(훈련)/ '직업'(테스트)
⑥ '행복'(훈련) / '행복'(테스트)
⑦ '행복+경제'(훈련) / '행복'(테스트)
⑧ '행복+성공'(훈련) / '행복'(테스트)
⑨ '행복+직업'(훈련) / '행복'(테스트)
⑩ '행복+경제/성공/직업'(훈련) / '행복'(테스트)
⑪ '통합'(훈련)/ '통합'(테스트)

이렇게 훈련 데이터를 섞어서 실험한 이유는, 주제가 서로 다른 한국어 쓰기 데이터를 혼합하여 언어 모델을 미세조정했을 때에 분류 성능이 얼마나 향상하는지를 확인하기 위해서이다. 데이터가 적은 상황에서 이러한 방법이 도움이 되는지를 아는 것은 훗날 데이터가 부족한 상황에서 유용할 수 있다.

평가척도: 평가척도로는 7-겹 교차검증 결과의 평균 정확도(average accuracy)를 사용했다.

<그림 1> (위), <표 2> (아래) '직업' 테스트 데이터에서의 KoBERT, NB, LR 예측 성능 비

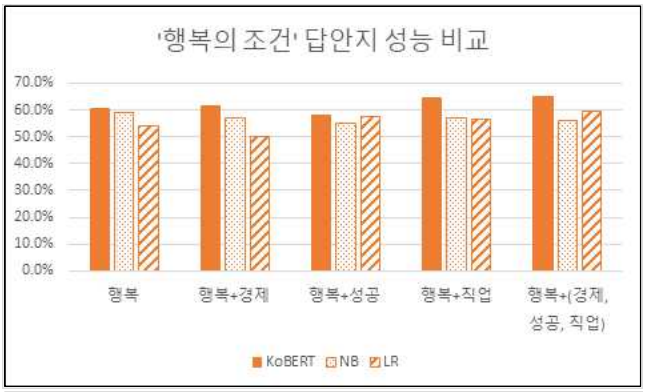


실험 구성	KoBERT	NB	LR
직업	43.5%	45.6%	36.6%
직업+경제	47.3%	46.8%	46.7%
직업+성공	46.7%	45.5%	41.6%
직업+행복	47.3%	44.7%	40.5%
직업+(경제, 성공, 행복)	46.4%	47.6%	42.7%

4. 결과

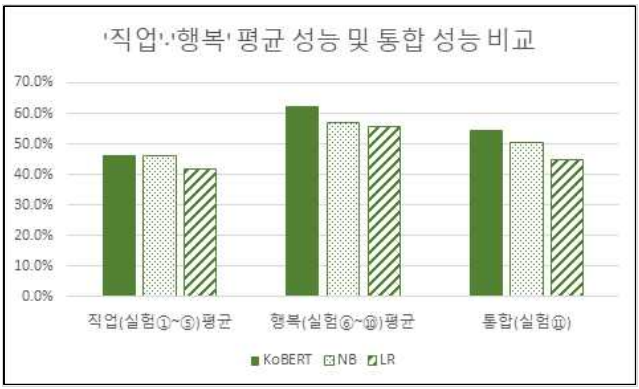
'직업', '행복', '통합' 테스트 데이터에 대한 실험 결과를 각각 (**<그림 1>**, **<표 2>**), (**<그림 2>**, **<표 3>**), (**<그림 3>**, **<표 4>**)에 제시한다. '직업' 테스트 데이터 분류에서는 '직업' 훈련 데이터만 사용한 경우(①)와 ⑤의 경우를 제외하고 나머지 모든 경우에서 KoBERT가 NB와 LR보다 조금 나은 성능을 보였다. LR이 셋 중에서 가장 안 좋은 성능을 보였다.

<그림 2> (위), <표 3> (아래) '행복' 테스트 데이터에서의 KoBERT, NB, LR 예측 성능 비교



실험 구성	KoBERT	NB	LR
행복	60.7%	59.3%	54.2%
행복+경제	61.6%	57.2%	50.0%
행복+성공	58.0%	55.3%	57.3%
행복+직업	64.3%	57.2%	56.4%
행복+(경제, 성공, 직업)	65.2%	56.2%	59.4%

<그림 3> (위), <표 4> (아래) '직업'·'행복'의 평균 성능과 '통합' 테스트 데이터의 성능



실험 구성	KoBERT	NB	LR
직업(실험①~⑤)평균	46.2%	46.0%	41.6%
행복(실험⑥~⑩)평균	62.0%	57.0%	55.5%
통합(실험⑪)	54.5%	50.7%	44.7%

‘행복’ 테스트 데이터 분류에서는 모든 경우에서 KoBERT가 NB나 LR보다 나은 성능을 보였다. ‘통합’ 테스트 데이터 분류에서도 KoBERT가 NB나 LR보다 높은 성능을 보였고, LR이 가장 낮은 성능을 보였다. 또 한국어 쓰기의 주제에 따라 성능에 차이가 나는 것을 확인할 수 있었다. ‘행복’ 실험(⑥~⑩)이 상대적으로 높은 분류 정확도를 보였다.

한편, 실험 결과 중 KoBERT, NB, LR 세 모델의 성능 차이가 가장 작았던 ‘직업+경제’ 훈련 데이터로 학습한(②) 로지스틱 회귀의 특징 단어를 살펴보았다. 가중치가 큰 순으로 각 클래스(A, B, C, D)에서 상위 10개 단어를 **<표 5>**에 제시한다.

<표 5> ‘직업+경제’ 로지스틱 회귀의 특징 단어

클래스	A	B	C	D
특징 단어	직성	위하	회사	사회
	가능	마다	가지	어떤
	발전	때문	취미	ㅂ니다
	아무리	고려	다고	영향
	에서	도덕	왜냐하면	문제
	여유	힘들	집안	연봉
	조건	는데	생각	전문
	ㄴ다면	직업	문단	우리
	지만	전공	기분	그러나
	그리고	조사	스트레스	미치

본 문서 분류의 목적이 한국어 쓰기 수준을 가늠하는 것이라는 사실을 고려할 때, 오탈자, 표현 오류, 문법 오류 등의 오류 표현들이 특징 단어로 취급돼야 함이 마땅하나, 이 실험에서는 형태소분석기를 이용한 기본적인 단어 취득만을 시행했기 때문에 두드러지는 오류 표현들이 특징 단어에 포함되지 않았다. 이러한 점은 향후 한국어 쓰기 답안지의 점수 구간 분류기를 만들 때 유의해야 할 사항이다. 같은 맥락에서 KoBERT의 분류 성능이 예상보다 낮았던 이유를 이 같은 불충분한 오류 표현에서 찾을 수 있을 것 같다. 사전학습된 언어모델은 오류가 거의 없는 위키피디아와 같은 텍스트로부터 구축된 것이어서, 이러한 언어모델을 가지고 오류 표현을 발견하려는 훨씬 더 많은 오류 표현 데이터가 필요할 것으로 보인다.

그런데도 **<표 5>**와 같이 특정한 특징 단어에 높은 가중치가 부여된 것은 답안지의 어휘 사용 수준(고급·중급 등의 어휘 사용)이 암암리에 반영된 결과로 추정된다.

실험 당시 훈련 데이터로 언어모델을 미세조정할 때마다 KoBERT 분류 모델이 훈련 데이터에 과적합 하는 양상을 보였는데, 그런데도 확률기반 분류기보다 조금이기는 하나 더 나은 예측 성능을 나타내, 앞으로의 활용 가능성을 엿볼 수 있었다.

5. 결론

데이터가 매우 적은 상황에서 심층 언어모델을 이용하는 것은 그다지 좋은 방법은 아니지만, 우리는 이 논문에서 한국어 사전학습모델을 활용함으로써 적은 양의 데이터로도 NB나 LR보다 더 나은 심층학습 기반 텍스트 분류 모델을 구축할 수 있다는 사실을 확인했다. 앞으로 우리는 한국어 쓰기 답안지 데이터를 확충하면서, 적은 양의 데이터로도 더 정확하게 점수 구간을 분류할 수 있는 사전학습모델 미세조정 기법에 관해 연구하려고 한다.

본 연구에서 사용된 데이터는 중앙대학교 인문콘텐츠연구소 홈페이지[5]에서 내려받을 수 있으며, 실험에 사용한 코드는 연구자의 GitHub 저장소에서 확인할 수 있다.[6]

참고문헌

[1] 최준영 · 임희석, 자연어처리 모델을 이용한 이커머스 데이터 기반 감성 분석 모델 구축, 한국융합학회논문지, 11권, 11호, pp. 33-39, 2020.

[2] 황상흠 · 김도현, 한국어 기술문서 분석을 위한 BERT 기반의 분류 모델, 한국전자거래학회지, 25권, 1호, pp. 203-214, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT, 2019, pp. 4171 - 4186.

[4] Michele Banko, Eric Brill, “Scaling to Very Very Large Corpora for Natural Language Disambiguation,” ACL, 2001, pp. 26-33. (Fig. 1)

[5] 조희련 · 이유미 · 임현열 · 차준우 · 이찬규, 딥러닝 기반 언어모델을 이용한 한국어 학습자 쓰기 평가의 자동 점수 구간 분류 -KoBERT와 KoGPT2를 중심으로-, 한국언어문화학, 18권, 1호, 2021.

5) http://aihumanities.org/ (상단 메뉴의 ‘아카이브’ > ‘데이터’)
6) https://github.com/heeryoncho/korean_essay_grade_prediction