한국어 Sentence-BERT 임베딩을 활용한 자동 쓰기 평가 계층적 구조 모델

조민수O, 권오욱, 김영길

한국전자통신연구원
{mscho, ohwoog, kimyk}@etri.re.kr

Hierarchical Automated Essay Evaluation Model Using Korean Sentence-Bert Embedding

Minsoo ChoO, Oh Woog Kwon, Young Kil Kim
Electronics and Telecommunications Research Institute

요 약

자동 쓰기 평가 연구는 쓰기 답안지를 채점하는데 드는 시간과 비용을 절감할 수 있어, 교육 분야에서 큰 관심을 가지고 있다. 본 연구의 목적은 쓰기 답안지의 문서 구조를 효과적으로 학습하여 평가하고, 문장 단위의 피드백을 제공하는데 있다. 그 방법으로는 문장 레벨에서 한국어 Sentence-BERT 모델을 활용하여 각 문장을 임베딩하고, LSTM 어텐션 모델을 활용하여 문서 레벨에서 임베딩 문장을 모델링한다. '한국어 쓰기 텍스트-점수 구간 데이터'를 활용하여 해당 모델의 성능 평가를 진행하였으며, 다양한 KoBERT 기반 모델과 비교 평가를 통해 제안하는 모델의 방법론이 효과적임을 입증하였다.

주제어: 자동 쓰기 평가, 문장 임베딩, 계층 모델

1. 서론

자동 쓰기 평가(Automated Essay Scoring)란 특정 주제로 작성된 쓰기 답안지에 대해서 내용 전달성, 표현 적합성, 유창성 등을 고려하여 평가 항목별 또는 종합 점수를 자동으로 매기는 것이다. 실제 교육 분야에서는 평가자가 쓰기 답안지를 채점하는데 드는 시간과 비용을 절감하고, 평가자들 간 채점 결과에 대한 신뢰성을 검증하는 목적으로 자동 쓰기 평가 연구에 대한 관심을 가지고 있다.

최근 영어 자동 쓰기 평가 연구[1]에서는, 다양한 딥러닝 방법론을 통해 문서를 효과적으로 임베딩 및 학습하여 쓰기 평가 성능을 향상시켰다. 한국어 데이터를 활용한 연구[2,3] 중에는 한국어 버전의 BERT[4]인 KoBERT를 파인튜닝(Fine-tuning)하여, 적은 데이터로도 유용한 결과를 낸 연구가 있다. 이 연구는 쓰기 내용 전체를 문서 단위로 인코딩(Encoding)하는 방법으로, 문서 레벨에서 쓰기 내용을 학습한다. 본 연구에서는 문서를 보다 구조적으로 모델링하고, 문서 내 중요한 문장을 선택적으로 학습하도록 어텐션(Attention)을 적용하며, 피드백 측면에서는 학습자가 종합 점수 뿐만 아니라, 각 문장의 중요도에 대한 피드백을 제공받을 수 있도록 한다. 이를 위해, 문서를 문장 또는 문단 단위로 쪼개어 각각을 한국어 Sentence-BERT(SBERT)로 임베딩한 후, 임베딩된 각 문당을 LSTM 어텐션 레이어에 적용한다. 여러 비교 실험을 통해, 한국어 SBERT-LSTM-Attention 모델이 기본 KoBERT보다 적은 파라미터로 우수한 성능을 보임을 확인하였다.

2. 관련 연구

초기 자동 쓰기 평가 연구[5]는 문법 오류, 단어 개수 등의 중요 특징(Feature)을 수작업(Hand-crafted)으로 선택하고 추출하여 학습시키는 기계학습 방법을 적용했다. 대표적인 기계학습 기반 평가 시스템으로, 품사(POS) 정보, N-gram 모델 정보 등의 특징을 추출하여 회귀 모델에 적용한 오픈 소스 시스템 EASE(Enhanced AI Scoring Engine)가 있다.

딥러닝 기반의 연구에는 CNN(Convolutional Neural Network)과 RNN(Recurrent Neural Network)계열의 모델을 활용한 연구가 있다. 이후[6] 연구에서는 문장과 문서 레벨 모두에서 쓰기 내용을 학습하기 위해, 두 모델을 계층적으로 쌓아 결합하였다. SKIPFLOW[7]는 LSTM을 활용하여 문서 내 일관성(Coherence)을 명시적으로 학습하도록 한 연구이다.

사전 학습 언어 모델 등장 이후에는 Transformer기반의 BERT, Roberta, Albert 등의 모델에 분류기를 결합하여, 해당 학습 데이터로 파인튜닝하는 연구가 수행되었다[8]. [9]연구에서는 BERT 모델에 회귀 손실 함수(Regression loss)와 순위화 손실 함수(Ranking loss)를 두어 동시에 학습하도록 하여, 공개 당시 최고 성능을 달성하였다. 한국어 자동 쓰기 평가 [2]연구에서는 한국어 버전의 KoBERT와 KoGPT2를 활용하여 적용한 연구가 있다. 해당 연구는 나이브 베이즈(Naïve Bayes)와 로지스틱 (Logistic Regression) 방법론보다 우수한 성능을 보였다.

3. 모델

3.1 한국어 SBERT 임베딩

한국어 SBERT는 기존 BERT 모델에 NLI(Natural Language Inferencing)과 STS(Semantic Textual Similarity) 파인튜닝 태스크를 수행하여 문장 임베딩 성능을 개선시킨 모델로, 본 연구에서는 쓰기 답안지의 내용을 문장 단위로 나눈 후, 한국어 SBERT 활용하여 각 문장을 임베딩한다. 본 연구에서는 Huggingface에서 제공하는 'snunlp/KR-SBERT-V40K-klueNLI-augSTS' SBERT 모델을 사용한다.

추가적으로, 문장 단위가 아닌 문단 단위로 임베딩한 모델을 실험하기 위해 표 1의 쓰기 답안지 예시와 같이 데이터의 문서에 표시된 '[[문단]]' 을 기준으로 문단을 나누어 SBERT 모델로 각 문단을 임베딩한다. 단, 3.2의 SBERT-LSTM-Att 모델에 대한 설명은 문장 단위의 임베딩을 기준으로 한다.

표 1. 문단 단위로 표시된 한국어 쓰기 답안지 예시

[[문단]] 요새 중국에 대학교에서 학업그만두고 떠나간 학생은 많아지고 있다. 이런 학생들은 대학교에서 배운 지식이나 과정은 나중에 생활중에서 큰 도움을 될 수가 없다고 생각한다. '대학교에서 배운 것이 미친영향을 받을수있지만 이런지식을배워도 돈을 벌기가 어렵다.' **[[문단]]** 지금 생각해도 4년대학교 졸업하면 내가 돈 벌기가 어렵다. 그래서 여러분 직업을 선택할때 무슨 조건이 있을가? **[[문단]]** 내 생각은 직업 선택할때돈 취미 시간 세까지 차직한다. ...

3.2 한국어 SBERT-LSTM-Att 모델

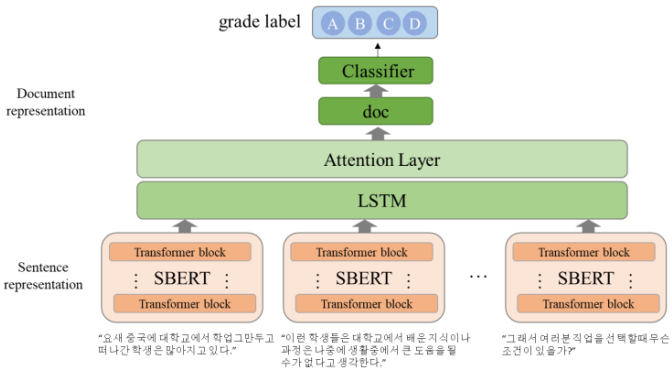

그림 1. SBERT-LSTM-Att 모델

그림 1은 SBERT-LSTM-Att을 나타낸 모델이다. 3.1에서 문장 단위로 임베딩한 각 벡터를 문서 레벨로 임베딩하기 위해 LSTM(Long Short-Term Memory)에 순차적으로 입력한다. 수식 1은 문장 임베딩 $(s_1, s_2, \dots s_t)$을 순차적으로 인코딩한 LSTM의 은닉 상태(Hidden states)을 나타낸다.

$$h_i = LSTM([s_1, s_2, \dots s_t]) \quad (1)$$

LSTM으로 임베딩된 벡터는 어텐션 레이어에 입력되어 문서 내 중요도가 높은 문장을 학습한다. 수식 2는 수식 3에서 계산된 어텐션 가중치 a_i와 h_i의 가중합으로 문서 벡터 doc을 나타낸다. 수식 3의 W_a, V_a와 b_a는 각각 학습 가능한 가중치 행렬, 가중치 벡터, 편향 벡터이다.

$$doc = \sum_i a_i h_i \quad (2)$$

$$a_i = softmax(V_a \tanh(W_a h_i + b_a)) \quad (3)$$

수식 3의 doc 벡터는 수식 4의 선형 분류기(Linear classifier)에 의해 최종 점수-구간 레이블을 예측한다.

$$FCNN(r) = Wr + b \quad (4)$$

$$\text{grade} = FCNN(doc) \quad (5)$$

4. 실험 및 결과

4.1 실험 데이터

실험에 사용한 데이터는 '한국어 쓰기 텍스트-점수 구간 데이터 세트' [3]로, 표 2의 네 개의 주제에 대해 작성된 304건의 한국어 쓰기 시험 답안지로 구성된다. 기존 답안지의 평가 점수는 0점부터 30점 사이에 분포되어 있으나, 공개된 데이터 셋은 전체 점수 구간을 네 개의 구간으로 나누어, 각 점수를 구간별 레이블(A, B, C, D)로 나타낸다.

표 2. 시험 답안지의 주제별 데이터 수

주제	데이터 수
① 직업의 조건	101건
② 행복의 조건	96건
③ 경제와 행복	62건
④ 성공의 기준	45건

모델 간 비교 평가를 위해 실험 방법은 [2]연구와 동일하게 진행하였다. 학습(Train) 데이터는 전체 데이터의 72%, 검증(Validation)과 테스트(Test) 데이터는 전체 학습 데이터의 14%로 각각 설정하였으며, 실험은 7겹 교차 검증(7-cross validation)으로 수행하였다. 모델 학습 시에는 (1)~(4) 주제 데이터를 모두 활용하였으나, 주제별 테스트 시에는 데이터의 수가 적어 과적합이 발생하는 90건 미만의 (3)과 (4)주제 데이터는 제외하였으며, 통합 테스트시에는 (1)~(4) 주제 데이터를 모두 활용하였다.

4.2 실험 환경

표 3. 모델 하이퍼파라미터

Hyperparameter	Value
Epoch	50
Batch size	4
SBert hidden size	768

LSTM hidden size	768
Max sentence length	25
Max paragraph length	10
Learning rate	1e-5
Dropout	0.1

본 모델은 Python기반의 Pytorch 프레임워크로 구현되었으며, 모델의 하이퍼파라미터(Hyperparameter)는 표 3과 같다. 문장 단위의 SBERT-LSTM-Att모델에서는 최대 문장의 개수를 25개로 설정하였으며, 문단 단위에서는 최대 문단의 개수를 10개로 설정하여 실험을 진행하였다.

4.3 실험 결과

제안하는 분류 모델의 성능 평가를 위해, 평가 척도는 정확도(Accuracy)를 사용하였으며, 데이터 수가 적은 환경에서 보다 안정적으로 성능을 확인하고자 7겹 교차 검증을 세 번 수행한 평균 수치로 비교 평가를 진행한다.

표 5와 6은 각각 '직업'과 '행복' 테스트 데이터에 대한 KoBERT, KoBERT-CNN, KoBERT-LSTM, 제안하는 모델(Kor-SBERT-LSTM-Att)의 성능 결과로, KoBERT-CNN, KoBERT-LSTM은 KoBERT의 마지막 층의 은닉 상태(hidden state) 벡터에 각각 CNN, LSTM을 적용한 모델이다. 각 표의 AVG(평균)는 (1)~(5), (6)~(10) 데이터로 학습한 테스트 데이터에 대한 평균 정확도로, KoBERT-CNN, KoBERT-LSTM이 KoBERT보다 전반적으로 우수한 성능을 보인다. KoBERT로 학습된 벡터가 CNN과 LSTM을 통해 문서 특징에 따라, 지역적인 정보 또는 전역적인 정보를 학습하는 것으로 분석할 수 있다. AVG 정확도 기준, Kor-SBERT-LSTM-Att 모델이 KoBERT보다 '직업' 테스트 데이터에 대해서는 1.89%p, '행복' 테스트 데이터에 대해서는 4.76%p 성능 향상을 보였다. 이를 통해, 계층 구조를 활용하여 문서를 구조적으로 모델링하는 것이 문서 레벨의 KoBERT 모델보다 효과적임을 알 수 있다.

표 7은 문장과 문단 단위 임베딩의 성능 비교 표로, 문장 단위로 나누어 임베딩한 모델이 문단 단위보다 네 개의 주제에 대한 통합 데이터에 대해서 떨어지는 성능을 보였으나, '직업', 행복' 테스트 데이터에 대해서는 우수한 성능을 보였다. 따라서, 데이터 수가 상대적으로 보장된 주제의 데이터에서는 문장 단위로 임베딩한 LSTM-Att 신경망 모델이 한국어 쓰기 평가 모델로 적합함을 확인할 수 있다.

표 4. 주제별 데이터 조합

(1) 직업
(2) 직업 + 행복
(3) 직업 + 경제
(4) 직업 + 성공
(5) 직업 + 행복 + 경제 + 성공
(6) 행복
(7) 행복 + 경제
(8) 행복 + 성공
(9) 행복 + 직업
(10)행복 + 경제 + 성공 + 직업

표 5. '직업' 테스트 데이터에 대한 성능 비교

	KoBERT (base)	KoBERT-CNN	KoBERT-LSTM	Kor-SBERT-LSTM-att
(1)	39.58 %	**46.83 %**	43.35 %	44.05 %
(2)	**50.20 %**	42.06 %	49.80 %	49.90 %
(3)	45.83 %	46.73 %	46.03 %	**48.12 %**
(4)	44.35 %	46.13 %	45.04 %	**47.02 %**
(5)	49.01 %	47.02 %	**49.70 %**	49.31 %
AVG	45.79 %	45.75 %	46.79 %	**47.68 %**

** (1)~(5)에 대한 학습 데이터는 표 4를 참고

표 6. '행복' 테스트 데이터에 대한 성능 비교

	KoBERT (base)	KoBERT-CNN	KoBERT-LSTM	Kor-SBERT-LSTM-Att
(6)	59.82 %	62.20 %	61.61 %	**65.77 %**
(7)	58.63 %	61.61 %	56.55 %	**69.05 %**
(8)	58.04 %	58.04 %	60.12 %	**61.31 %**
(9)	62.50 %	**62.80 %**	60.71 %	62.20 %
(10)	62.20 %	65.48 %	64.29 %	**66.67 %**
AVG	60.24 %	62.02 %	60.66 %	**65.00 %**

** (6)~(10)에 대한 학습 데이터는 표 4를 참고

표 7. 문장과 문단 단위 임베딩 성능 비교

	Kor-SBERT-LSTM-Att (문장)	Kor-SBERT-LSTM-Att (문단)
직업 평균	**47.68 %**	46.96 %
행복 평균	**65.00 %**	64.11 %
통합 평균	49.82 %	**50.47 %**

표 8은 '행복의 조건' 주제에 대한 쓰기 답안지 문장의 어텐션 스코어 예시이다. 각 어텐션 스코어는 LSTM-Att 레이어에서 계산된 각 문장의 중요도로, 스코어가 높을수록 중요도가 높은 문장을 나타낸다. 분석 결과, 스코어가 비교적 높은 2번, 13~15번 문장의 경우, '행복의 조건' 주제와 관련성이 높고, 문법 오류도 상대적으로 적다. 반면, 스코어 낮은 문장 4번의 경우 문법적 오류가 존재하며, 주요 문장보다는 보충 설명 또는 근거 문장에 가까운 것을 확인 할 수 있다. 이와 같이 어텐션 방법론을 통해 평가 정확도를 높일 수 있을 뿐 아니라, 어텐션 스코어를 통해 평가 결과에 대한 피드백을 제공할 수 있다. 그림 2와 같이 어텐션 스코어를 시각화하여 학습자에게 제공할 경우, 학습자는 평과 결과를 보다 직관적으로 이해할 수 있을 것이다. 특히나, 교육 분야에서 피드백은 학습자에게 중요한 학습 과정 중 하나이므로, 어텐션 스코어 및 시각화 제공을 통해 학습자의 쓰기 교육에 도움을 줄 수 있다.

표 8. 쓰기 답안지 문장의 어텐션 스코어 예시

번호	문장	어텐션 스코어

1	[[문단]] 행복한 삶을 사느냐는 자신이 원하는 삶을 사느냐에 달려있다.	0.036
2	명확한 삶의 이미와 분명한 행복관을 가지고 자신이 원하는 삶을 사는 것이야말로 행복한 삶이다고 할 수 있다.	0.072
3	[[문단]] 행복하게 살기 위해 충족되어야 할 조건은 객관전인 측면과 주관적인 측면에서 살펴볼 수 있다.	0.049
4	무선 객관적인 측면에서 살펴보자.	0.021
5	우리가 사회에 속해 있는 이상 사회 제도하에 놓이는 건 당연하다.	0.023
6	사회제도가 없으면 정확한 행복관을 갖기는 어렵다.	0.024
7	아무리 꿈이 많은 사람이라도 뜻을 펴지 못하면 안타까울 뿐이다.	0.032
8	그래서 꿈을 실현 할 수 있는 객관적인 환경도 필요하다.	0.046
9	또한 사회 복지도 필요하다.	0.03
10	사회 복지를 통해서 우리는 자신이 원하는 삶에 한발짝 다가갈 수 있기때문이다.	0.042
11	[[문단]] 주관적인 측면에서 필요한 행복한 삶의 조건은 도덕성을 갖는 것이다.	0.038
12	아무리 목적을 성취했기로서니 도덕성이 결여된다면 행복한 삶을 누릴 수 없다.	0.033
13	행복한 삶을 누리기 위해서는 꿈과 희망도 꼭 필요하다.	0.076
14	또한 적극적인 태도와 만족감을 느끼는 것도 행복한 삶에 필요한 조건이다.	0.085
15	웃음을 잃지 않고 꾸준히 긍적적인 태도를 견지하노라면 세상이 아름답게 느껴질 것이다.	0.079
16	[[문단]] 또한 가족들이 행복하면 우리는 행복한다고 느낀다.	0.065
17	한 마디로 주변 환경이 얼마나 평화롭고 안정적이냐에 따라 행복을 느낄 수 있다.	0.043

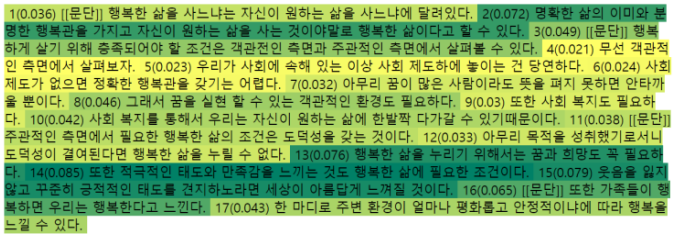

그림 2. 쓰기 답안지 문장의 어텐션 스코어 시각화 예시

5. 결론

본 연구에서는, 한국어 자동 쓰기 평가 연구를 위한 한국어 SBERT-LSTM-Att 모델을 제안한다. 각 쓰기 답안지를 문장과 문서 레벨에서 구조적으로 모델링하기 위해 SBERT 임베딩과 LSTM 모델을 계층적으로 구성한다. 또한 LSTM 모델에 어텐션을 적용하여 문서 내 중요한 문장 정보를 효율적으로 학습하도록 하며, 해당 가중치 정보를 활용하여 학습자에게 부여되는 점수 구간에 대한 피드백을 제공한다. 제안하는 모델을 검증하고자, '한국어 쓰기 텍스트-점수 구간 데이터 세트'를 활용하여, Kor-Bert기반의 모델 간 비교 평가를 진행하였다. 비교 실험을 통해, 한국어 SBERT-LSTM-Att 모델이 우수한 성능을 보임을 확인하였다.

후속 연구로는, 한국어 쓰기 평가 데이터가 적은 조건에서 주제별 전이가 가능한 자동 쓰기 평가 연구를 진행할 것이다. 실제 교육 현장에서, 자동 쓰기 평가 모델을 활용하기 위해 모든 주제에 대한 충분한 데이터를 제공할 수 없기 때문에 주제별 전이가 가능한 형태의 시스템이 필요하다. 향후에는 구조적으로 모델링 가능하며, 적은 데이터로도 다양한 주제에 대해서 처리 가능한 형태로 연구를 확장할 계획이다.

사사문구

* 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

* This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners)

참고문헌

[1] Uto, Masaki., "A review of deep-neural automated essay scoring models", Behaviormetrika 48.2 pp.459-484, 2021.

[2] 조희련, 임현열, 이유미, & 차준우, "한국어 학습 모델별 한국어 쓰기 답안지점수 구간 예측 성능 비교", 정보처리학회논문지, 소프트웨어 및 데이터 공학, 11, pp.133-140, 2022.

[3] 조희련, 임현열, 차준우, & 이유미, "KoBERT, 나이브 베이즈, 로지스틱 회귀의한국어 쓰기 답안지 점수 구간 예측 성능 비교", 한국정보처리학회 학술대회논문집, 28(1), pp.501-504, 2021.

[4] Devlin, Jacob, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding.", NAACL, Volume 1, pp.4171-4186, 2018.

[5] Hussein, M. A., Hassan, H., & Nassef, M., "Automated language essay scoring systems: A literature review", PeerJ Computer Science, 5, e208, 2019.

[6] Taghipour, K., & Ng, H. T., "A neural approach to automated essay scoring", In Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 1882-1891, 2016.

[7] Tay, Y., Phan, M., Tuan, L. A., & Hui, S. C., "Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring", In Proceedings of the AAAI conference on artificial intelligence, Vol. 32, No. 1, 2018.

[8] Ormerod, C. M., Malhotra, A., & Jafari, A., "Automated essay scoring using efficient transformer-

based language models”, arXiv preprint arXiv:2102.13136, 2021.
[9] Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X., “Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking”, In Findings of the Association for Computational Linguistics: EMNLP, pp. 1560-1569, 2020.