Journal of KIIT. Vol. 21, No. 4, pp. 31-43, Apr. 30, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.4.31

CLES-BERT: 에세이 점수 예측을 위한 대조학습 기반 BERT 모델

유대곤*, 김용연**[1], 한상우**[2], 온병원***

CLES-BERT: Contrastive Learning-based BERT Model for Automated Essay Scoring

Daegon Yu*, Yongyeon Kim**[1], Sangwoo Han**[2], and Byung-Won On***

이 논문은 2022년도 정부(과기정통부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2022R1A2C1011404)

요 약

4차 산업 혁명 시대에서 창의력은 중요하며, 글쓰기는 창의력 기르는 교육 방법 중 하나이다. 그러나 현재 학교에서 시행하는 글쓰기 평가 방식은 주관적으로 평가한다는 문제점이 있다. 에세이 자동 평가(AES)는 객관성을 확보할 수 있을 뿐만 아니라 평가자의 시간과 노력을 줄여주는 역할을 한다. 본 논문은 효과적인 AES 작업을 위해 BERT 모델에 대조 학습을 활용한 모델을 제안한다. 제안 모델은 대조 학습 손실 함수를 추가하여 효과적인 에세이 임베딩 표현을 구현한다. 또한, 점수별 에세이의 평균 임베딩을 대조 학습에서의 샘플로 사용하는 방식으로 양성과 음성 샘플을 선택한다. ASAP 에세이 데이터셋을 사용하여 실험한 결과, 제안 모델인 CLES-BERT 모델이 기존 BERT 모델보다 최대 3% 정확도 향상을 보였다.

Abstract

Creativity is an important ability in the 4th industrial revolution so writing is one of educational tools to improve creativity. However, student's essays have been mainly evaluated subjectively in schools. To address this problem, Automated Essay Evaluation(AES) plays an important role in objective evaluation in addition to reducing the time and effort of instructors. This paper presents a novel AES model in which contrastive learning-based loss function is added to BERT. Furthermore, for contrastive learning, positive and negative samples are selected based on mean embedding vectors per essay score. The experimental results show that the proposed Contrastive Learning Essay Scoring-Bidirectional Encoder Representations from Transformers(CLES-BERT) improved average accuracy up to 3%, compared to main AES models, in Automated Student Assessment Prize(ASAP) data set.

Keywords
automated essay scoring, BERT, multi-task loss function, contrastive learning, sampling

* ㈜애니파이브 연구원
- ORCID : https://orcid.org/0000-0003-4331-8302
** 군산대학교 소프트웨어학부 학사과정
- ORCID[1]: https://orcid.org/0000-0002-1502-1205
- ORCID[2]: https://orcid.org/0000-0002-7514-2041
*** 군산대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-6929-1388

· Received: Mar. 06, 2023, Revised: Apr. 03, 2023, Accepted: Apr. 06, 2023
· Corresponding Author: Byung-Won On
School of Software, Kunsan National University, 558, Daehak-ro, Gunsan, Jeollabuk-do, Korea
Tel.: +82-63-469-8913, Email: bwon@kunsan.ac.kr

Ⅰ. 서 론

4차 산업혁명 시대가 도래하면서 인간만이 가지고 있는 창의력의 중요성이 대두되고 있다[1]. 따라서 인간은 창의력을 계발해야 하며 이를 위한 교육 역시 중요하게 대두되고 있다[2]. 마찬가지로 고등교육영역에서도 창의융합교육을 통한 인재양성을 요구하고 있다. 이러한 교육방식 중 하나로 대상에 대한 깊은 성찰을 바탕으로 이루어지는 글쓰기는 우리의 창의력을 기르는데 도움을 주며 올바른 방향으로 창의력을 기르기 위해서는 글쓰기 평가가 제대로 이루어져야 한다.

학교에서 글쓰기를 평가하는 방법은 크게 3가지가 있다. 이 3가지 방법은 총체적 평가(Holistic scoring), 분석적 평가(Analytic scoring) 그리고 주요 특성 평가(Primary trait scoring)로 총체적 평가는 글 전체에 초점을 맞추어서 평가하는 방식이고 분석적 평가는 한 편의 글을 여러 구성 요소에 대하여 개별적으로 평가하는 방식이고 주요 특성 평가는 글의 특성에 따라 2개 이상의 주요 요소를 선정하여 평가하는 방식이다[3][4]. 제시된 3가지 글쓰기 평가 방법 모두 평가자의 주관적인 평가에 의존하기 때문에 평가자 간의 평가가 일치하는 경우가 드물다. 또한, 평가자 간의 협의 과정을 통해 개인의 채점 결과를 보완하여 평가자 간의 평가를 일치시킬 수 있지만, 시간과 전문가 초빙 등 큰 비용을 지불해야 한다는 단점이 있다. 추가로 평가해야 하는 학생 수가 많다면 평가하는 데 있어 평가자의 많은 시간과 노력을 요구한다는 문제점도 존재한다. 따라서 이러한 글쓰기 평가를 자동화 및 객관화시킨다면 글을 평가하는 데 있어 많은 비용, 시간과 노력을 줄이고 보다 객관적으로 글을 평가할 수 있을 것으로 기대된다.

글쓰기 평가를 자동화하는 작업인 에세이 자동평가(AES, Automated Essay Scoring) 작업은 사람이 평가하는 것이 아닌 컴퓨터로 작성된 모델이 자동으로 글을 읽고 평가하는 분야이다. 이러한 AES 모델의 입력 데이터는 텍스트 데이터이며 출력으로 입력 데이터에 대한 평가 점수가 나오게 된다. AES 모델은 보통 2가지 과정을 거치게 되는데 첫 번째 과정은 에세이의 임베딩을 표현하는 과정이고 두 번째 과정은 표현된 에세이 임베딩을 평가하는 과정이다.

초기 AES 연구는 전문가들이 수작업으로 지정한 자질(Feature)을 기반으로 AES 모델을 학습시켰다. 이 방식은 글의 길이, 문법, 어휘, 문장 구조 등을 분석하여 에세이의 점수를 계산하는 방식으로 동작한다. 하지만 전문가가 지정한 자질에 의존하므로 다양한 유형의 글을 처리하기에는 한계가 있다. 최근에는 이런 자질 선택 작업이 필요 없는 딥러닝 기술을 사용한 연구가 진행되고 있다[5]. 심층 신경망 기반의 AES 모델은 프롬프트(Prompt) 사용 여부와 평가 방법에 따라 4가지 유형으로 구분된다[6]. (1) Prompt-specific holistic scoring은 가장 보편적인 AES 유형으로 특정 프롬프트로 학습된 모델이 같은 프롬프트로 테스트를 진행하여 에세이에 대한 총점을 평가하는 방식이다. (2) Prompt-specific trait scoring은 (1)과 동일하게 모델에게 동일한 프롬프트로 학습과 테스트를 진행하지만 에세이에 대한 점수를 다양한 평가 항목에 대해 부여하는 방식이다. (3) Cross-prompt holistic scoring은 프롬프트의 사용 없이 에세이에 대한 총점을 평가하는 방식이다. (4) Cross-prompt trait scoring은 프롬프트의 사용 없이 에세이에 대한 점수를 다양한 평가 항목에 대해 부여하는 방식이다. 본 논문은 사전 학습된 기존의 BERT 모델과 비교하여 AES 작업에서 제안 방안이 효과적인지 입증하기 위해 (1)의 방식을 취한다.

심층 신경망 기반의 AES 모델 발전 동향은 다음과 같다. 순환 신경망인 RNN(Recurrent Neural Network)과 LSTM(Long Short-Term Memory)에서 출발하여 현재는 트랜스포머 기반의 사전 학습(Pre-training) 모델을 이용한 연구가 활발히 이루어지고 있다. 사전 학습 모델 중 대표적인 모델인 BERT(Bidirectional Encoder Representations from Transformers)[7] 모델을 활용한 연구가 현재 주를 이루고 있는데 사전 학습된 BERT를 그대로 사용할 경우 BERT에서 파생된 문장 표현은 작은 영역 안에 매핑되어 서로 의미가 다를지라도 높은 유사성을 보이게 된다는 문제점이 있다. 이는 적절한 에세이의 임베딩을 표현하는 데 있어 한계가 있음을 나타낸다. 이러한 문제를 해결하는 방법 중 하나로 대조 학습(CL, Contrastive Learning)이 있다[8].

대조 학습은 입력 데이터와 의미가 유사한 양성 샘플(Positive sample)과는 가깝게 하고 입력 데이터와 의미가 상이한 음성 샘플(Negative sample)과는 멀어지도록 학습하여 효과적인 이미지 또는 문장 표현을 학습하도록 모델을 훈련하는 학습 방법이다. 따라서 대조 학습에서는 양성 샘플과 음성 샘플을 어떻게 설정하느냐에 따라 성능이 크게 좌우될 수 있다. 이미지처리 분야에서의 양성 샘플은 입력 데이터의 회전이나 밝기 변경 등의 방법으로 생성하고 자연어처리 분야에서의 양성 샘플은 입력 데이터의 단어의 순서를 바꾸거나 임의의 단어를 삭제하는 방법으로 생성한다. 최근에는 사전 학습 시에 대조 학습을 사용하여 성능을 향상하거나 한국어의 경우 어미를 바꿔서 양성 샘플을 생성하는 연구가 진행되고 있다[9]. 음성 샘플은 이미지처리나 자연어처리 분야 둘 다 입력 데이터와 서로 다른 클래스의 데이터(또는 서로 다른 의미를 갖는 데이터)를 음성 샘플로 설정한다. 또한, 양성과 음성 샘플을 구성하는 데 있어 기존의 이미지처리 분야와 자연어처리 분야에서는 하나의 데이터만을 사용하여 양성과 음성 샘플을 구성하였는데 본 논문에서는 점수별 에세이의 평균 임베딩을 양성과 음성 샘플로 구성한다. 점수별 에세이의 평균 임베딩을 사용하는 이유는 같은 점수의 에세이일지라도 에세이에 담긴 내용은 다를 수 있다. 하지만 에세이 점수가 서로 같다면 두 에세이의 특징은 같을 것이다. 여기서 에세이의 특징은 “이해하기 쉬운 글인가?”, “명확한 표현을 사용하였는가?”, “주장에 대한 근거가 타당한가?”와 같이 글의 문법과 문체를 고려한 특징이라고 할 수 있다. 따라서 본 논문에서는 이러한 점수별 에세이의 공통되는 특징을 고려하기 위해 점수별 에세이의 평균 임베딩을 대조 학습에서의 양성과 음성 샘플로 사용한다.

또한, 양성과 음성 샘플을 선택하는 기준은 기존에는 입력 데이터와 유사한 의미를 갖는 데이터를 양성 샘플로 선택하고 입력 데이터와 상이한 의미를 갖는 데이터를 음성 샘플로 선택했다면 제안 방안의 양성과 음성 샘플의 선택 방식은 다음과 같다. 양성 샘플은 입력 데이터와 같은 점수의 평균 에세이 임베딩으로 선택되어 같은 점수의 문법과 문체가 입력 임베딩과 가까워지도록 유도한다. 음성 샘플은 에세이 점수 범위의 중앙값을 기준으로 양성과 음성 샘플을 분리하여 선택된다.

제안 모델인 CLES-BERT(Contrastive Learning Essay Scoring-BERT) 모델은 더 나은 에세이 임베딩 표현을 위해 제안 방안의 대조 학습을 BERT 모델에 적용하였다. 기존의 교차 엔트로피(Cross-entropy) 손실 함수에 대조 학습 손실 함수를 추가하여 손실 함수를 재조정하였으며 ASAP(Automated Student Assessment Prize)[10] 데이터셋을 이용하여 실험한 결과 기존의 BERT 모델보다 정확도가 최대 3% 향상되었다.

본 연구의 기여도는 다음과 같다.

AES 작업에서 효과적인 에세이 임베딩 표현을 위한 CLES-BERT 모델을 최초로 제안한다. 제안 모델의 손실 함수는 기존의 교차 엔트로피 손실 함수에 대조 학습 손실 함수를 추가하여 새롭게 정의되었다. 또한, AES 작업에서 적합한 대조 학습에서의 양성과 음성 샘플의 구성 방법과 선택 기준을 제안한다.

제안 방안의 대조 학습 적용 후 ASAP의 에세이 데이터셋에서 CLES-BERT 모델의 정확도가 최대 3% 향상되었다. 대조 학습 시 제안 방안의 양성과 음성 샘플의 분류 기준과 점수별 에세이의 평균 임베딩을 생성하여 점수별 에세이의 공통된 특징인 문법과 문체를 고려하는 것이 에세이 임베딩 표현에 효과적임을 알 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 관해 기술하고 3장에서는 본 논문에서 제안하는 방안에 관하여 기술한다. 4장에서는 본 논문의 실험 환경과 결과를 기술하고, 마지막으로 5장에서는 결론 및 향후 계획에 관해 기술한다.

II. 관련 연구

2.1 AES 모델

전통적인 AES 모델은 기계 학습(Machine learning)을 활용하여 단어 수, 단어 길이, 문장 길이와 문법과 같은 에세이를 표현하는 자질을 수작업(Hand-craft feature)으로 추출하는 방식으로 진행되었다.

[11][12]는 베이지안 모델과 k-NN(k-Nearest Neighbor) 모델을 사용한 연구이다. [13]는 베이지안 선형 회귀를 사용하여 학습한 도메인 외에 다른 도메인에서도 사용 가능한 모델을 제시하였다. [14]-[16]는 SVD(Singular Value Decomposition) 방법을 사용하여 에세이의 의미적 벡터를 생성하고 벡터간의 유사도를 측정하는 잠재 의미 분석(LSA, Latent Semantic Analysis) 방법을 사용한 연구이다. 이러한 기계 학습 접근 방법은 에세이를 평가하고 채점하기 위해 사람이 직접 자질을 추출해야 하고 모델의 성능이 추출된 자질에 의존한다는 문제점이 있다. 따라서 이러한 수작업 없이 자동으로 에세이와 레이블링 된 점수 사이의 관계를 학습하기 위해 심층 신경망을 이용한 연구가 이어졌다.

심층 신경망을 이용한 초기 연구는 [17]-[19]로 순환 신경망 모델인 RNN과 LSTM을 사용한 연구이다. 순환 신경망 모델은 입력 데이터를 순차적으로 처리한다는 특징이 있지만 시퀀스-투-시퀀스(Sequence-to-sequence) 모델로 BERT에 비해 장기 기억에 취약하고 학습 속도가 느린 문제가 있다[20].

BERT는 트랜스포머의 인코더 구조를 활용한 언어 표현 모델이다. [21]는 여러 개의 BERT를 활용하여 AES에 접근한 연구로 토큰 단위(Token-scale), 세그먼트 단위(Segment-scale)와 문서 단위(Document-scale)로 다양한 단위(Scale)로 나눠서 에세이를 평가한다는 특징이 있다. 에세이의 임베딩 표현을 다양한 단위로 나눠서 함으로써 보다 정밀하게 에세이 점수를 산출할 수 있다는 장점이 있지만 각 단위마다의 에세이 임베딩 표현이 적절히 학습되기에는 어렵다는 단점이 있고 제안 방안은 하나의 BERT 모델을 사용하여 에세이 임베딩 표현이 AES 작업에서 영향을 끼치는 정도를 알아보고자 하기에 해당 연구와는 성격이 다르다.

2.2 CL 모델

최근 대조 학습은 좋은 성능을 보이며 이미지처리 분야나 자연어처리 분야 등 다양한 분야에서 활발히 연구되고 있는 기술이다. 대조 학습은 이미지처리 분야에서 처음으로 사용되었으며[22] 이미지처리 분야에서 쓰인 대조 학습 연구는 다음과 같다. [23]-[25]은 각 이미지에 대해 잘라내기, 회전등의 이미지 변환 방식으로 두 가지의 이미지를 생성하고 잠재 공간(Latent space)에서 서로 가깝게 만든다. [23]은 InfoNCE[26]라 불리는 정규화된 템퍼레이쳐 스케일(Temperature scale) 기반의 교차 엔트로피 손실(NT-Xent)을 손실 함수로 사용하여 정규화된 임베딩에서 더 나은 임베딩 표현을 보였다.

최근 대조 학습은 자연어처리(NLP, Natural Language Process) 분야에서도 활발히 사용되고 있다. [27]은 BERT 모델 위에 CNN(Convolutional Neural Network) 레이어를 추가하고 글로벌 문장 임베딩과 해당 로컬 문맥(Context) 간의 상호 정보(MI, Mutual Information)를 최대화하는 학습 방안을 제안한다. [28]는 [24]와 유사한 구조를 채택하고 데이터 증강을 위해 역번역을 사용한다. 하지만 역번역은 거짓 정보를 생성할 수 있다는 단점이 존재한다. [29]는 [23]의 아키텍처를 활용하여 대조 학습과 마스크 언어 모델(Masked language model)을 함께 학습한다. 그러나 의미의 유사도를 최대화하는 범위(Span) 내에서만 대조 학습을 사용한다는 단점이 존재한다.

제안 모델은 에세이 임베딩의 더 나은 표현을 위해 대조 학습을 활용하여 에세이를 평가한다. 또한, 대조 학습에서 양성과 음성 샘플을 구성하기 위해 점수별로 에세이 벡터들을 평균 내어 점수별 에세이 특징 벡터를 추출했다는 점에서 기존 대조 학습 연구와는 다르다. 또한, 중앙값을 기준으로 양성과 음성 샘플을 분리시킨 점에서도 기존 대조 학습 연구와는 다르다.

III. 제안 방안

이번 절에서는 AES 작업에서 효과적인 에세이 임베딩 표현을 위한 CLES-BERT의 제안 방안에 대해 자세히 설명한다.

본 논문의 제안 방안은 크게 3단계로 구성된다. 첫 번째는 제안 방안의 대조 학습에서 샘플을 구성하는 방법, 두 번째는 양성 샘플과 음성 샘플의 선택 기준, 마지막으로는 제안 모델의 손실 함수에 대해 설명한다.

대조 학습이란 그림 1과 같이 의미가 유사한 쌍의 벡터는 서로 가깝게 유도하고 의미가 다른 쌍의 벡터는 멀리 떨어지게 유도하여 효과적인 이미지 또는 문장 표현을 학습하도록 모델을 훈련하는 학습 방법이다. 그림 1은 대조 학습 방법의 예시이다. Dog 1과 Dog 2는 같은 레이블인 Dog로 레이블링 된 임베딩이다. Cat은 Dog와는 다른 레이블로 Dog와는 다른 의미를 지닌다. 따라서 유사한 의미를 지닌 Dog 1과 Dog 2의 임베딩은 서로 가까워지고 Dog 레이블과는 다른 의미를 지닌 Cat과는 멀어지게 학습을 진행하게 된다.

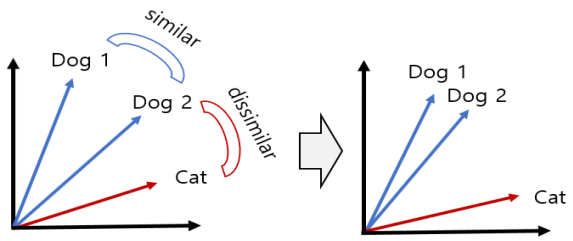


그림 1. 대조 학습 개념
Fig. 1. Contrastive learning concept

3.1 제안 방안의 샘플을 구성하는 방법

본 논문의 대조 학습을 활용할 때의 양성과 음성 샘플의 구성은 에세이의 개별 임베딩이 아닌 점수별 평균 임베딩으로 구성된다. 점수별 평균 임베딩을 사용하는 이유는 다음과 같다. 같은 점수의 에세이일지라도 에세이에 담긴 내용은 다를 수 있다. 하지만 에세이 점수가 서로 같다면 두 에세이의 특징은 유사할 것이다. 즉, 에세이의 내용이 아닌 점수별 에세이의 공통적인 특징을 고려하는 벡터를 샘플로 구성한다. 여기서 점수별 에세이의 공통적인 특징은 “이해하기 쉬운 글인가?”, “명확한 표현을 사용하였는가?”, “주장에 대한 근거가 타당한가?”와 같이 문법과 문체를 고려한 특징이라고 할 수 있다. 점수별 평균 벡터는 이러한 점수별 에세이의 공통적인 특징인 문법과 문체를 내포한다. 점수별 에세이의 공통적인 특징 벡터를 추출하기 위해서는 해당하는 작업에 맞게 학습된 모델이 필요하다. 따라서 교차 엔트로피 손실 함수만을 사용하여 첫 번째 미세조정 단계(Fine-tuning)를 통해 모델을 일차적으로 학습시킨다. 그리고 이렇게 학습된 모델을 이용하여 Algorithm1을 통해 점수별 평균 벡터를 생성한다. 점수별 평균 벡터를 구한 방식은 다음과 같다. 첫 번째 미세조정 단계를 마친 BERT 모델을 사용하여 추출된 단어별 임베딩 값 중 [CLS](Classification) 벡터만을 추출한다. 추출된 [CLS] 벡터를 사용하여 Algorithm1을 통해 점수별 평균 벡터를 그림 2와 같이 생성한다. 여기서 추출된 [CLS] 벡터는 다른 모든 단어 벡터를 모두 참고한 문맥 벡터의 역할을 한다.

Algorithm 1의 입력 데이터는 첫 번째 미세조정 단계을 통해 학습된 BERT 모델에서 학습 데이터를 1 에폭(epoch) 만큼 진행하여 출력된 [CLS] 벡터, 학습 데이터의 라벨 값과 평균 벡터를 저장할 리스트이다. Line 1은 학습 데이터의 라벨 값 중 중복된 것을 제거한 점수를 라벨 리스트로 설정한다. 예를 들어 학습 데이터의 라벨이 [1,1,2,2,3]이라면 라벨 리스트는 [1,2,3]이다. Line 2는 for 문에서 라벨 리스트의 원소를 한 개씩 반복하며 실행한다. Line 3은 하나의 점수에 해당하는 [CLS] 벡터들을 하나의 변수에 저장한다. Line 4는 저장된 하나의 점수에 해당하는 [CLS] 벡터들의 평균 벡터를 구한다. 이때 히든 사이즈는 유지한 채 각 [CLS] 벡터들에 대한 평균을 계산한다. Line 5는 생성한 평균 벡터를 리스트에 추가한다. 반환 값은 평균 벡터들을 점수별로 모은 리스트이다.

그림 2의 $e^{n}(m\text{p점})$은 학습 데이터에서 n번째 학습 데이터의 에세이에 레이블링 된 점수 m를 의미한다. $\bar{e}_{l}$은 l점의 평균 벡터를 의미한다.

Algorithm 1 : Mean vector generation

```
INPUT : training_cls_vectors, training_labels,
                mean_vectors
1. label_list = set(training_labels)
2. for label in label_list:
3. cls_vectors_per_score
   = training_cls_vectors[training_labels == label]
4. mean_vector = mean(cls_vectors_per_score)
5. mean_vectors.append(mean_vector)
   return mean_vectors
```

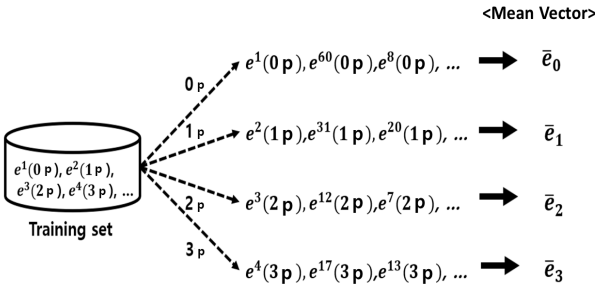


그림 2. 에세이 점수별 평균 벡터 생성
Fig. 2. Generating the mean embedding vector per essay score

3.2 양성 샘플과 음성 샘플의 선택 기준

제안 방안의 샘플링 방식은 중앙값 기준으로 점수 범위를 나누어 나눈 두 개의 점수 범위에서 각각 하나씩 점수를 선택한다. 그리고 선택된 두 개의 점수의 평균 벡터를 각각 양성 샘플과 음성 샘플로 설정한다. 예를 들어 입력 데이터의 점수가 1점이고 그림 3과 같이 점수 범위가 1~6점으로 점수 범위의 값이 짝수 개라면 중앙값 3.5점을 기준으로 3.5점보다 큰 값과 3.5점 이하의 값으로 점수 범위를 나눈다. 양성 샘플은 입력 데이터와 같은 점수의 평균 벡터로 설정한다. 즉, 입력 데이터의 점수가 1점이므로 1점의 에세이 평균 벡터를 양성 샘플로 설정한다. 음성 샘플은 양성 샘플과 다른 점수 범위에서 무작위로 하나의 점수를 선택하여 해당 점수의 평균 벡터로 설정한다. 반면 그림 4와 같이 점수 범위가 2~12점으로 점수 범위의 값이 홀수 개라면 마찬가지로 점수 범위를 중앙값 7점을 기준으로 7점보다 큰 값과 7점 이하의 값으로 나눈다. 양성 샘플은 입력 데이터와 같은 점수의 평균 벡터로 설정하고 음성 샘플은 양성 샘플과 다른 점수 범위에서 무작위로 하나의 점수를 선택하여 해당 점수의 평균 벡터로 선택한다.

그림 3과 같이 에세이의 점수 범위에 해당하는 값의 개수가 짝수 개라면 값을 크기순으로 나열했을 때 가장 중앙에 위치하는 값이 유일하지 않고 두 개가 될 수 있다. 이 경우 그 두 값의 평균을 중앙값으로 설정하여 중앙값보다 큰 값과 중앙값 이하의 값으로 점수 범위를 나눈다.

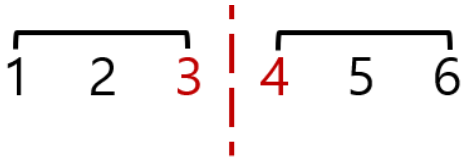


그림 3. 양성과 음성 샘플을 구분하는 중앙값 찾기 (점수 개수가 짝수일 때)
Fig. 3. Boundary decision for positive and negative samples(# of score bins is even)

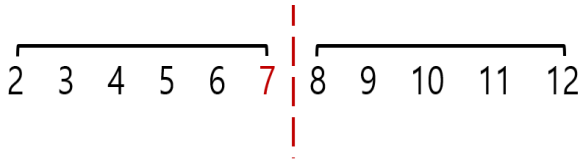


그림 4. 양성과 음성 샘플을 구분하는 중앙값 찾기 (점수 개수가 홀수일 때)
Fig. 4. Boundary decision for positive and negative samples(# of score bins is odd)

그림 4와 같이 에세이의 점수 범위에 해당하는 값의 개수가 홀수 개라면 값을 크기순으로 나열했을 때 가장 중앙에 위치하는 값이 유일하게 되고 그 값을 중앙값으로 설정하여 중앙값보다 큰 값과 중앙값 이하의 값으로 점수 범위를 나눈다.

중앙값보다 큰 값과 중앙값 이하의 값으로 점수 범위를 나누는 이유는 에세이를 평가할 때 점수가 높은 글인 잘 쓴 글과 점수가 낮은 글인 잘 쓰지 못한 글로 구분하여 대조 학습에서 입력 데이터와 다른 성격을 띄는 쌍과는 서로 멀리 떨어지게 학습하기 위함이다.

Algorithm 2의 입력 데이터는 Algorithm 1에서 생성한 평균 벡터 리스트, 중앙값과 미니 배치의 라벨 값이다. Line 1은 중앙값을 정수형으로 변환한다. 예를 들어 중앙값이 1.5라면 소수점 값을 내려서 1로 정수 값을 반환한다. Line 2~3은 배치 데이터의 에세이 점수를 하나씩 불러오며 for문을 반복한다. Line 4는 에세이 점수와 같은 점수의 평균 벡터를 양성 샘플로 선택한다. Line 5~9는 그림 5와 같이 중앙값을 기준으로 점수 범위를 나누고 에세이 점수와는 다른 점수 범위에서 무작위로 점수 하나를 뽑아 해당 점수의 평균 벡터를 음성 샘플로 구성한다.

Algorithm 2 : Sampling for contrastive learning

```
INPUT : mean_vectors, median, batch_labels
median = int(median)
for idx in range(len(batch_labels)):
essay_score = batch_labels[idx]
positive_sample = mean_vectors[essay_score]
if essay_score > median :
negative_samples = mean_vectors[: median+1]
else :
negative_samples = mean_vectors[median+1 :]
negative_sample = random.choice(negative_
samples)
```

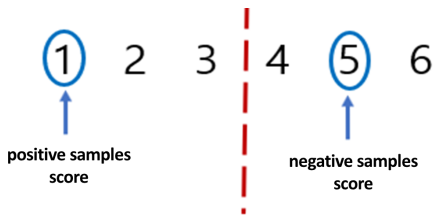


그림 5. 중앙값 기준 양성 샘플과 음성 샘플의 선택 예시
Fig. 5. Example of sampling positive and negative samples

그림 5과 같이 에세이 점수 범위가 1~6점이고 에세이 점수가 1점이라면 중앙값 3.5점을 기준으로 점수를 나누게 되고 양성 샘플의 점수는 입력 데이터의 점수와 같은 1점의 평균 벡터를 양성 샘플로 분류하고 양성 샘플이 속한 점수 범위와는 다른 점수 범위에서 무작위로 하나의 점수를 선택하여 해당 점수의 평균 벡터를 음성 샘플로 구성한다.

3.3 CLES-BERT 손실 함수

CLES-BERT의 미세조정 단계의 입력값은 한 에세이 씩 들어간다. ASAP 데이터셋에서 프롬프트 1의 경우 에세이당 평균 단어 수는 350개이다. 에세이 당 하나의 레이블링 된 점수를 라벨값으로 사용한다. 제안 모델은 미세조정을 2번 하는 학습 과정을 거치는데 1단계 미세조정의 손실 함수는 식 (1)과 같이 교차 엔트로피 손실 함수만을 사용한다. 이때 $P(x)$는 실제 라벨값을 나타내고 $Q(x)$는 모델이 추정한 확률값을 나타낸다.

$$Loss_{CE} = -\sum P(x)\log Q(x) \quad (1)$$

교차 엔트로피 손실 함수는 클래스마다 추정된 확률값에 대하여 정답 클래스에 해당하는 추정 확률값이 1에 가까워질수록 손실 값은 작아지는 특성이 있다. 따라서 입력값으로 에세이가 들어가면 레이블링 된 점수 값으로 분류되도록 학습을 진행하게 된다.

2단계 미세조정의 손실 함수는 식 (3)과 같이 기존의 교차 엔트로피 손실 함수에 본 논문에서 제안한 대조 학습 손실 함수를 추가한다. e_i는 입력 데이터의 임베딩, e_p와 e_n은 각각 양성 샘플과 음성 샘플을 의미한다. $dist$는 유클리드 거리와 sim은 코사인 유사도를 의미한다. λ_1과 λ_2는 초매개변수이며 모든 실험에서 각각 5와 10으로 설정하였다.

$$Loss_{CL} = \lambda_1 * p + \lambda_2 * q \quad (2)$$
$$p = \frac{dist(e_i, e_p)}{\{dist(e_i, e_p) + dist(e_i, e_n)\}}$$
$$q = \left|sim(e_i, e_n)\right|$$

식 (2)에서 대조 학습 손실 함수는 p와 q로 이루어져 있다. 입력 데이터의 임베딩이 양성 샘플과 유클리드 거리가 작고 음성 샘플과는 클수록 p의 값이 작아지게 된다. 한편 입력 데이터의 임베딩이 음성 샘플과 코사인 유사도가 작을수록 q의 값이 작아지게 된다. 즉, 입력 데이터 임베딩이 양성 샘플과는 유사하고 음성 샘플과는 상이할수록 손실 값이 작아져 그림 6와 같이 학습이 진행되게 된다.

$$Loss_{TOTAL} = \alpha * Loss_{CE} + (1-\alpha) * Loss_{CL} \quad (3)$$

전체 손실 함수($Loss_{TOTAL}$)는 식 (3)과 같이 교차 엔트로피 손실 함수($Loss_{CE}$)에 대조 학습 손실 함수($Loss_{CL}$)를 더한 값이 된다. α는 초매개변수이며 모든 실험에서 0.5로 설정하였다.

그림 6에서 $\vec{e_p}$와 $\vec{e_n}$은 양성과 음성 샘플에 해당하는 벡터를 의미하고 $\vec{e_i}$는 입력 데이터의 임베딩을 의미한다. 따라서 대조 학습 시 입력 데이터의 임베딩은 양성 샘플과는 가까워지고 음성 샘플과는 멀어지게 학습을 진행하게 된다.

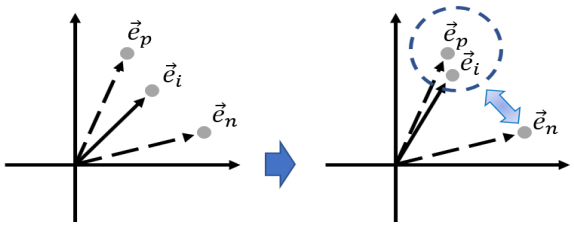


그림 6. 대조 학습에 의한 입력 임베딩의 학습 과정
Fig. 6. Contrastive learning based on mean embedding vectors

IV. 실 험

4.1 실험 환경

ASAP 데이터셋은 8개의 서로 다른 프롬프트로 구성되어 있다. 본 실험은 1~8의 8개의 프롬프트 중 프롬프트 7~8은 점수 범위가 0~30과 0~60으로 넓으므로 분류 문제를 해결하기에는 적절하지 못하기 때문에 제외했다. 따라서 1~6의 에세이 프롬프트만 사용하여 실험을 진행하였다. 학습과 테스트셋은 8:2 비율로 분할한다. 표 1은 실험한 ASAP 데이터셋의 6개의 프롬프트에 대한 에세이 수, 에세이당 평균 단어 수, 에세이 점수 범위, 에세이 주제를 나타낸다. 에세이 점수가 높을수록 명확하고 설득력 있는 에세이고 낮을수록 모호하거나 이해하기 어려운 에세이다.

표 2는 실험에 사용한 컴퓨터 사양을 나타낸다. 딥러닝 모델인 BERT를 학습하기 위해 Python 3.7, PyTorch 1.8 그리고 CUDA 11.1을 사용하였다. 또한, 사전 학습 모델은 BERT-base 모델을 사용하였다.

표 3는 BERT 학습 시 미세조정 단계에 사용한 초매개변수이다. 표에 제시된 파라미터 값을 적용했을 때 가장 좋은 결과를 보인다. 해당 에폭을 넘어가면 과적합으로 모델의 정확도가 떨어지게 된다.

표 2. 실험 환경
Table 2. Experimental set-up

Classification	Specification
CPU	Intel(R) Core™ i9-10940X (3.30GHz, core: 14)
GPU	NVIDIA GeForce RTX 3090
SSD	250GB
OS	Ubuntu 18.04 LTS
Software	Python, Pytorch, BERT(Base Model)

표 3. 미세조정 단계 초매개변수
Table 3. Fine-tuning hyperparameters

Hyperparameter	1st Fine-tuning step	2nd Fine-tuning step
Dropout rate	0.1	0.1
Batch size	24	24
Learning rate	6e-5	6e-5
max length	512	512
Epochs	4	1
λ_1	-	5
λ_2	-	10

표 1. ASAP 데이터셋 통계
Table 1. Characteristics of ASAP data set

Prompt	# of essays	Average # of word tokens	Score range	Essay topic
1	1,785	350	2~12	Write a letter to the local newspaper giving an opinion about the effect computers are having on people
2	1,800	350	1~6	Write a persuasive essay for the newspaper that reflects your views on library censorship
3	1,726	150	0~3	Write a response explaining how environmental influences affect cyclists
4	1,772	150	0~3	Write a response explaining the reason for the author's last phrase
5	1,805	150	0~4	Describing the atmosphere created by the author in the memoir
6	1,800	150	0~4	To describe the obstacles the builders of the Empire State Building faced when attempting to dock airships there.

모델의 정확도는 식 (4)와 같이 모델이 에세이 점수 분류를 올바르게 했는지 평가하기 위해 평가 에세이 수와 점수 분류를 맞힌 에세이 수의 비율을 통해 계산한다.

$$\text{정확도} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

표 4는 혼동 행렬에서의 분류 모형 평가 지표로 각 요소의 의미는 True Positive(TP)는 맞는 것을 올바르게 예측한 것, True Negative(TN)은 아닌 것을 올바르게 예측한 것, False Positive(FP)는 아닌 것을 올바르지 않게 예측한 것과 False Negative(FN)은 맞는 것을 올바르지 않게 예측한 것을 의미한다.

표 4. 혼동 행렬
Table 4. Confusion matrix

		Prediction	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

4.2 실험 결과

제안 방안의 성능 평가로 표 2의 하드웨어 사양과 표 3의 초매개변수 값을 사용하여 각 프롬프트에 대해 기존 모델과 제안 모델의 에세이 점수 분류의 정확도를 비교 실험하였다. AES 문제는 선형 회귀문제로 선형회귀, 로지스틱 회귀, SVM(Support Vector Machine)의 분류 모델을 사용하여 에세이 점수 분류 문제를 해결하였다. 하지만 최근에 앙상블 모델인 RF(Random Forest)와 GBM(Gradient Boosting Machine) 모델이 선형 회귀, 로지스틱 회귀, SVM의 분류 모델보다 더 높은 성능을 내는 것으로 밝혀졌다[30,31,32]. 따라서 본 실험에서는 선형 회귀, 로지스틱 회귀, SVM의 분류 모델은 비교 실험하지 않았다. 비교 실험으로 사용한 모델은 RF, GBM, Bi-LSTM, BERT, 제안 모델인 CLES-BERT로 총 5개의 모델을 사용하였다.

그림 7은 실험한 모든 프롬프트에 대해 기존 모델과 제안 모델의 평균 분류 정확도를 비교한 결과이다. RF의 평균 정확도는 62.5%로 61.7%인 GBM과 60.7%인 Bi-LSTM보다 평균 정확도가 높다. 이는 에세이의 다양한 자질을 기준으로 에세이 점수를 분류하는 것이 AES 작업에 효과적임을 알 수 있다. 또한, GBM과 RF 모델 모두 텍스트 데이터와 같은 고차원 데이터에서도 잘 동작하고 입력 데이터의 차원이 높아지더라도 과적합 문제를 잘 다루며 각각의 결정 트리에서 분류 기준을 임의로 선택하므로 데이터의 노이즈에 대한 강건성이 높다. 또한, 텍스트 데이터와 같은 데이터에서 중요한 특징을 찾아내는 능력이 뛰어나므로 분류 성능이 높다.

딥러닝을 사용한 점수 분류의 성능 비교는 다음과 같다. Bi-LSTM의 평균 정확도는 60.7%이고 제안 모델의 평균 정확도는 70.2%로 정확도가 약 10% 차이가 난다. 같은 딥러닝 모델이라도 AES 작업에서 성능 차이가 크게 나는 이유는 Bi-LSTM은 입력 데이터를 각각 정방향과 역방향으로 순차적으로 처리하여 그 결과를 병합하는 방식으로 문맥을 이해하는 모델인 반면 BERT는 양방향 트랜스포머를 기반으로 입력 데이터를 동시에 양방향으로 학습하므로 입력 데이터의 문맥을 더 잘 이해할 수 있기 때문이다. 또한, BERT는 대량의 텍스트 데이터를 사용하여 사전학습된 언어 모델이기 때문에 텍스트 데이터에 대한 사전 지식을 효과적으로 활용하여 텍스트의 의미와 문맥을 이해한다.

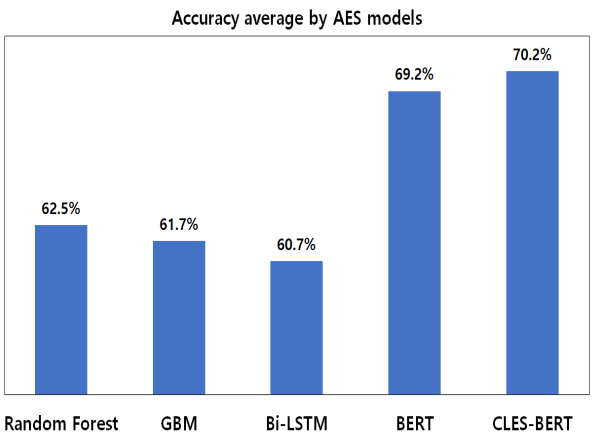


그림 7. 기존 모델과 제안 모델의 평균 정확도 비교 결과
Fig. 7. Results of main AES models

따라서 사전 학습된 모델 사용 유무가 AES 작업에서의 임베딩 표현에 영향을 주고 AES 작업에서 에세이 임베딩을 잘 표현하는가 그렇지 못한가에 따라 성능이 크게 달라질 수 있음을 확인할 수 있다. 제안 모델의 평균 정확도가 70.2%로 69.2%인 BERT의 평균 정확도보다 높은 것을 보아 AES 작업에서 제안 방안의 대조 학습 방법이 BERT 모델만 이용한 것보다 에세이의 임베딩을 보다 잘 표현하는 데 효과적임을 알 수 있다. BERT는 에세이의 의미를 내포하는 벡터를 표현하여 에세이 점수 분류를 하게 된다. 하지만 AES 작업에서 에세이의 점수가 같을지라도 에세이의 담긴 내용은 서로 다를 수 있다. 에세이의 점수가 같다는 것은 에세이의 내용이 아닌 글의 전달력, 명확한 표현 사용, 근거의 타당성과 같은 에세이의 특징이 비슷하다는 것을 의미한다. 따라서 좀 더 정확한 에세이 점수 분류를 위해서는 에세이의 특징을 고려해야 한다. 제안 모델은 대조 학습 방식을 이용하여 이러한 에세이의 특징을 고려하였기 때문에 기존 모델보다 높은 성능을 보인다.

그림 8은 프롬프트 1부터 프롬프트 6에 대해 기존 모델과 제안 모델의 에세이 점수 분류 정확도를를 비교한 결과이다. 비교 결과 프롬프트 1에서 베이스라인 모델인 BERT는 57.7% 정확도를 보였고 제안 모델의 정확도는 60.7%의 정확도를 보여 약 3% 향상으로 BERT와 비교 했을 때 가장 높은 성능향상을 보인다. 프롬프트 1의 점수 범위는 2~12로 잘 써진 글과 그렇지 않은 글을 구분하기에 충분하다. 따라서 프롬프트 1과 같이 적당히 넓은 점수 범위의 경우 제안 방안의 대조 학습 방식이 유용하게 동작함을 알 수 있다. 또한, 실험을 진행한 모든 프롬프트에서 제안 모델의 정확도가 각각 60.7%, 75%, 73.1%, 72.6%, 68.7%, 71.1%로 가장 높다. 이는 에세이에 사용된 구성이나 표현인 문법과 문체를 점수별 공통적인 특징으로 고려하는 것이 AES 작업에서 효과적임을 보여준다.

점수 범위에 따른 실험 결과는 프롬프트 1을 제외한 프롬프트 중 좁은 점수 범위는 0~3이고 넓은 점수 범위는 1~6에 해당하기 때문에 에세이 점수 범위가 2~12로 다른 프롬프트보다 점수 분류 범위가 넓은 프롬프트 1이 상대적으로 낮은 정확도를 보인다. 또한, 분류 범위가 좁은 프롬프트 2~6의 정확도를 비교하면 분류해야 할 에세이 점수의 개수가 4개에서 6개까지는 분류 정확도에 크게 영향을 끼치지 않는 것으로 보인다.

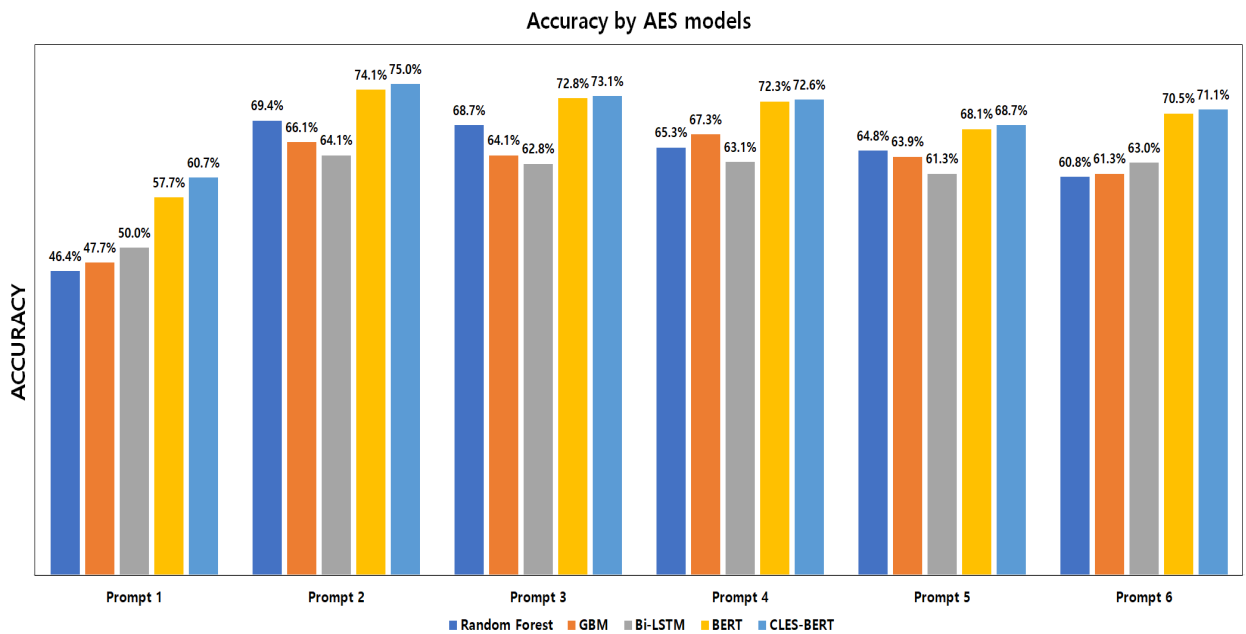


그림 8. 프롬프트별 기존 모델과 제안 모델의 정확도 비교 결과
Fig. 8. Results of main AES models per prompt

에세이 길이가 분류 성능에 미치는 영향은 프롬프트 2에서 대부분 모델의 정확도가 다른 프롬프트보다 높은 것으로 보아 길이가 짧은 에세이의 점수 분류보다 길이가 긴 에세이의 점수 분류가 더 쉽다는 것을 알 수 있다. 에세이 길이가 길수록 좀 더 많은 정보를 싣게 되므로 이러한 정보들이 에세이 분류 모델의 성능에 영향을 주기 때문이다. 프롬프트 1의 에세이 길이도 긴 편이나 분류해야 할 점수의 개수가 많아 다른 프롬프트보다 성능이 낮다. 따라서 에세이 점수 분류 시 에세이의 길이보다 분류해야 할 에세이의 점수 개수가 AES 작업에서 더 많은 영향을 끼침을 알 수 있다.

Ⅴ. 결론 및 향후 계획

본 논문의 결론은 다음과 같다. AES 연구에 이전에 시도된 바 없던 대조 학습을 AES 작업에 맞게 BERT 모델에 적용했고 실험 결과로 제안 모델인 CLES-BERT 모델이 BERT 모델보다 정확도가 최대 3% 향상되었다는 것을 확인함으로써 에세이 점수 평가의 정확도를 높일 수 있음을 입증하였다.

향후 계획은 베이스라인 모델로 사용한 BERT 모델은 최대 입력 데이터 길이가 512 토큰으로 제한된다는 한계가 있다. 본 한계점을 극복하기 위해 Longformer를 사용하여 입력 데이터 길이가 512가 넘는 데이터에 대해서도 에세이 자동 평가를 수행할 수 있게 하고 본 논문에서의 휴리스틱한 대조 학습의 샘플링 방식의 개선을 위해 MGRC(Mixed Gaussian Recurrent Chain) 알고리즘을 적용하여 대조 학습 샘플링을 진행할 예정이다.

References

[1] THE KOREA INDUSTRY DAILY, "Talent in the era of the 4th industrial revolution, creativity and communication skills are important", https://kidd.co.kr/news/208235 [Accessed: Feb. 18, 2023]

[2] ChosunMedia, "Dr. Jungkyu Lee, Creativity is the answer in the era of the 4th industrial revolution", https://www.chosun.com/site/data/html_dir/2018/02/05/2018020501828.html [Accessed: Feb. 18, 2023]

[3] V. Spandel and C. Marten, "Creating writers thro-ugh 6-trait writing assessment and instruction", Pe-arson, Montreal, 2005.

[4] T. H. Park, "Teaching essay rewriting using 6 ma-jor characteristic evaluation methods", The Interna-tional Association of Korean Language Education, pp. 143-162, 2005.

[5] Z. Ke and V. Ng, "Automated Essay Scoring: A Survey of the State of the Art", IJCAI, Vol. 19, pp. 6300-6308, 2019. https://doi.org/10.24963/ijcai.2019/879.

[6] R. Ridley, L. He, X. Dai, S. Huang, and J. Chen, "Automated cross-prompt scoring of essay traits", Proceedings of the AAAI conference on artificial Intelligence, Vol. 35, No. 15, May 2021. https://doi.org/10.1609/aaai.v35i15.17620.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanov-a, "Bert: Pre-training of deep bidirectional transfor-mers for language understanding", arXiv preprint arXiv:1810.04805, Oct. 2018. https://doi.org/10.48550/arXiv.1810.04805.

[8] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "Consert: A contrastive framework for self-supervised sentence representation transfer", arXiv preprint arXiv:2105.11741, May 2021. https://doi.org/10.48550/arXiv.2105.11741.

[9] D. G. Yu, S. W. Han, and B. Y. On, "A Survey of Contrastive Learning-based Deep Learning Models for Natural Language Processing", Proceedings of KIIT Conference, Jeju, Korea, pp. 475-478, Dec. 2022.

[10] Kaggle, "The Hewlett Foundation: Automated Es-say Scoring", https://www.kaggle.com/competitions/asap-aes/data [Accessed: Feb. 18, 2023]

[11] L. S. Larkey, "Automatic essay grading using text categorization techniques", Proceedings of the 21st annual international ACM SIGIR conference on R-esearch and development in information retrieval, pp. 90-95, Aug. 1998. https://doi.org/10.1145/290941.290965.

[12] L. M. Rudner and T. Liang, "Automated essay sc-oring using Bayes' theorem", The Journal of Tech-nology, Learning and Assessment, Vol. 1, No. 2, Jun. 2002.

[13] P. Phandi, K. M. A. Chai, and H. T. Ng, "Flexible d-omain adaptation for automated essay scoring usin-g correlated linear regression", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 431-439, Sep. 2015. https://doi.org/10.18653/v1/d15-1049.

[14] A. A. P. Ratna, P. D. Purnamasari, and B. A. Adhi, "SIMPLEO, the Essay grading system for Indones-ian Language using LSA method with multi-level keywords", The Asian Conference on Society, Ed-ucation & Technology, pp. 155-164, Oct. 2015.

[15] A. Amalia, D. Gunawan, Y. Fithri, and I. Aulia, "Automated Bahasa Indonesia essay evaluation wit-h latent semantic analysis", Journal of Physics: C-onference Series, Medan, Sumatera Utara, Indonesia, Vol. 1235, No. 1, Sep. 2018. https://doi.org/10.1088/1742-6596/1235/1/012100.

[16] A. A. P. Ratna, H. Khairunissa, A. Kaltsum, I. Ibra-him, and P. D. Purnamasari, "Automatic essay grad-ing for Bahasa Indonesia with support vector mac-hine and latent semantic analysis", 2019 Internatio-nal Conference on Electrical Engineering and Co-mputer Science(ICECOS), Batam, Indonesia, pp. 363-367, Oct. 2019. https://doi.org/10.1109/ICECOS47637.2019.8984528.

[17] H. Nguyen and L. Dery, "Neural networks for a-utomated essay grading", CS224d Stanford Report-s, pp. 1-11, 2016.

[18] K. Taghipour and H. T. Ng, "A neural approach t-o automated essay scoring", Proceedings of the 2016 conference on empirical methods in natural la-nguage processing, Austin, Texas, pp. 1882-1891, Nov. 2016. http://dx.doi.org/10.18653/v1/D16-1193.

[19] J. Yang, Y. Zhang, and S. Liang, "Subword enc-oding in lattice LSTM for Chinese word segment-ation", arXiv preprint arXiv:1810.12594, Oct. 2018. https://doi.org/10.48550/arXiv.1810.12594.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukh-in, "Attention is all you need", Advances in neur-al information processing systems, 2017.

[21] Y. Wang, C. Wang, R. Li, and H. Lin, "On the use of bert for automated essay scoring: Joint lea-rning of multi-scale essay representation", Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, pp. 3416-3425, Jul. 2022. http://dx.doi.org/10.18653/v1/2022.naacl-main.249.

[22] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensi-onality reduction by learning an invariant mappin-g", 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'06), New York, NY, USA, Jun. 2006. https://doi.org/10.1109/CVPR.2006.100.

[23] T. Chen, S. Kornblith, and M. Norouzi, "A sim-ple framework for contrastive learning of visual r-epresentations", International conference on machi-ne learning, pp. 1597-1607, Jul. 2020.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual repre-sentation learning", Proc. of the IEEE/CVF conference on computer vision and pattern recogn-ition, pp. 9729-9738, Nov. 2019. https://doi.org/10.48550/arXiv.1911.05722.

[25] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners", Advances in neu-ral information processing systems, Vol. 33, No. 1865, pp. 22243-22255, Dec. 2020.

[26] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Be-ngio, "Learning deep representations by mutual

in-formation estimation and maximization", preprint arXiv:1808.06670, Aug. 2018. https://doi.org/10.48550/arXiv.1808.06670.

[27] Y. Zhang, R. He, Z. Liu, K. H. Lim, and L. Bin-g, "An unsupervised sentence embedding method by mutual information maximization", arXiv prepr-int arXiv:2009.12061, Sep. 2020. https://doi.org/10.48550/arXiv.2009.12061.

[28] H. Fang, S. Wang, M. Zhou, J. Ding, and P. Xi-e, "Cert: Contrastive self-supervised learning for l-anguage understanding", arXiv preprint arXiv:2005.12766, May 2020. https://doi.org/10.48550/arXiv.2005.12766.

[29] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "Declutr: Deep contrastive learning for unsupervis-ed textual representations", arXiv preprint arXiv:2006.03659, Jun. 2020. https://doi.org/10.48550/arXiv.2006.03659.

[30] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning", The Stata Journ-al, Vol. 20, No. 1, pp. 3-29, 2020. https://doi.org/10.1177/1536867X20909688.

[31] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest", Information Computing and Applications: Third International Conference, pp. 246-252, 2012. https://doi.org/10.1007/978-3-642-34062-8_32.

[32] H. Ghanta, "Automated essay evaluation using n-atural language processing and machine learning", 2019.

저자소개

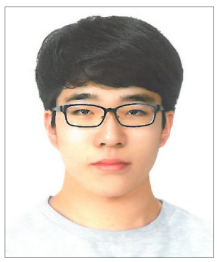
유 대 곤 (Daegon Yu)

2023년 2월 : 군산대학교 소프트웨어학과(학사)
2023년 4월 ~ 현재 : ㈜애니파이브 연구원
관심분야 : 자연어처리, 인공지능

김 용 연 (Yongyeon Kim)

2017년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심분야 : 자연어처리, 인공지능

한 상 우 (Sangwoo Han)



2018년 3월 ~ 현재 : 군산대학교 소프트웨어학부 학사과정
관심분야 : 자연어처리, 인공지능

온 병 원 (Byung-Won On)

2007년 : 펜실베이니아주립대학교 컴퓨터공학과 박사
2008년 ~ 2009년 : 브리티시컬럼비아대학교 컴퓨터과학과 박사후연구원
2010년 : 일리노이대학교 차세대디지털과학센터 선임연구원
2011년 ~ 2014년 : 서울대학교 차세대융합기술연구원 선임연구원
2014년 ~ 현재 : 군산대학교 소프트웨어학부 교수
관심 분야 : 데이터 마이닝, 자연어처리, 빅데이터, 인공지능, 강화학습