Journal of KIIT. Vol. 21, No. 9, pp. 51-63, Sep. 30, 2023. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2023.21.9.51

# 에세이 평가 자동화를 위한 대조 학습 기반 다중계층 BERT 모델 손실함수에 관한 연구

한상우*, 유대곤**, 온병원***, 이인규****

# Empirical Study on the Loss Functions of Contrastive Learning-based Multi-scale BERT model for Automated Essay Scoring

Sangwoo Han*, Daegon Yu**, Byung-Won On***, and Ingyu Lee****

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


## 요 약

에세이 작성은 학생들의 종합적 학습성과를 평가하기에 가장 유용한 방법으로 다양한 분야에서 사용되고 있다. 하지만 에세이 평가는 많은 시간이 소요되고 주관적인 경향이 있어서 사용이 제한적이다. 이를 극복하고자 에세이 평가 자동화에 관한 많은 연구가 진행되었지만, 현실적으로 사용할 수 있는 수준의 성능을 보이지 못하고 있다. 본 논문에서는 에세이 평가 자동화 시스템의 성능을 개선하기 위하여 현재 에세이 평가에서 가장 좋은 성능을 보이는 다중계층 BERT 모델의 평균제곱오차, 유사도, 순위 손실함수에 대조 학습 기반의 손실함수를 새롭게 추가하고, 손실함수 간의 성능을 비교한다. 에세이 평가에서 많이 사용되는 ASAP 데이터 세트를 이용하여 평가한 결과에 따르면, 대조 학습 기반 다중계층 BERT 모델은 기존의 BERT 모델보다 손실함수에 따라서 QWK는 3~4%, Pearson은 3~5% 향상되었다.

## Abstract

Essay writing is the most populary used method to evaluate students achievement in the class. However, the usage of essay writing is limited by the labor intensive and subjective nature of the essay scoring process. To overcome the latter, many researches have been done to automate essay scoring process. However, the performance of automated essay scoring is not suitable in practical usage. In this paper, we are proposing a contrastive learning-based Multi-scale BERT model to improve the performance of automated essay scoring. We applied many different loss functions to generate the positive and negative samples for contrastive learning, and experimented with ASAP benchmark dataset. According to our experiments, the constrastive learning based Multi-scale BERT model shows 3% improvement in QWK, and 4.5% improvement in Pearson according to the loss functions.



Keywords

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



* 군산대학교 소프트웨어학부 학사과정
- ORCID[1]: https://orcid.org/0000-0002-7514-2041
** 군산대학교 소프트웨어학부 학사
- ORCID[2]: https://orcid.org/0000-0003-4331-8302
*** 군산대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-6929-1388
**** 영남대학교 정보통신연구소 연구교수
- ORCID: https://orcid.org/0000-0001-9143-6384



· Received: Aug. 24, 2023, Revised: Sep. 18, 2023, Accepted: Sep. 21, 2023


· Corresponding Author: Byung-Won On
School of Computer Science and Engineering, Kunsan National University, 558, Daehak-ro, Gunsan, Jeollabuk-do, Korea
Tel.: +82-63-469-8913, Email: bwon@kunsan.ac.kr

# Ⅰ. 서 론

현시대에서는 인간과 AI가 공존을 고려해야 하는 시대가 열리고 있다. 인간의 노동시간 최소화와 생산성 극대화가 현시대의 흐름이다. 특히 교육 분야에서도 이러한 변화가 나타나고 있다. 교사는 AI 시스템을 활용하여 학생들의 교육과정을 지원할 수 있다. 교사의 에세이 평가는 학생들의 종합적인 학습 능력을 평가하는 중요한 방법의 하나다. 그러나 주관적이며 많은 시간을 요구하는 특징 때문에 한계가 존재한다.

최근에는 에세이 평가를 자동화하기 위한 연구가 활발하게 이루어지고 있지만, 아직은 실용적인 수준의 성능에 도달하지 못한 상황이다. 따라서 에세이 평가 분야에서 AI 기술을 개발하고 활용함으로써 교육과 평가의 효율성을 향상할 수 있을 것이다.

이 논문에서는 교육 분야에서의 노동집약적인 글쓰기 평가를 자동화하는 작업(AES, Automated Essay Scoring)을 다룰 것이다. AES 작업은 사람이 평가하는 것이 아닌 인공지능 모델이 자동으로 글을 읽고 평가하는 분야이다. 이러한 AES 모델의 입력 데이터는 글쓰기 데이터인 에세이 텍스트이고 출력으로 텍스트에 대한 평가 점수가 나오게 된다. 즉, 교사인 사람이 평가하는 것이 아닌 인공지능 모델이 자동으로 글쓰기를 평가한다. 이 과정은 두 단계를 거친다. 첫 번째 과정은 글쓰기의 임베딩을 표현하는 과정이고 두 번째 과정은 표현된 글쓰기 임베딩을 평가하는 과정이다.

초기 AES 연구는 전문가들이 수작업으로 지정한 특징들을 기반으로 AES 모델을 학습시켰다. 이 방법은 글의 길이, 문법, 어휘, 문장 구조 등을 분석하여 에세이의 점수를 계산하는 방식으로 작동했다. 그러나 이러한 방식은 전문가가 선택한 특징에 의존하기 때문에 다양한 유형의 글을 처리하는 데 한계가 있다.

표 1은 AES를 위한 벤치마크 데이터 세트인 ASAP[1]의 프롬프트 1~8을 사용하여 다양한 AES 모델들을 실험하고 얻은 정확도에 관한 결과를 보여준다. 초기에는 AES 문제를 회귀 분석(Regression analysis)으로 쉽게 모델링 가능하여 회귀 분석 기반의 AES  모델들이 제안되었다.

표 1. AES 모델 정확도 비교
Table 1. Accuracy comparison of AES models

| Type | Model | Accuracy |
|---|---|---|
| Machine learning | Regression analysis | 0.572 |
| | GBM | 0.617 |
| | RF | 0.625 |
| | SVR | 0.630 |
| Deep learning | Bi-LSTM | 0.607 |
| | BERT | 0.692 |
| | Multi-scale BERT | 0.712 |

하지만 정확도는 0.572로 성능 향상이 필요했다. 그래서 기계학습 모델인 GBM(Gradient Boost Machine), RF(Random Forest), SVR(Support Vector Regression) 등을 사용하며 정확도를 향상했다.

부스팅 모델인 GBM은 AES 문제를 다양한 특징을 고려하여 점수를 예측하는 데 활용했다. 부스팅 모델은 특히 단계별로 잘못 예측된 에세이를 학습하면서 AES 문제를 해결했다. 에세이에서 추출된 특징(단어, 문법 규칙, 구조 등)을 입력으로 사용하여 학습시켜 초기 모델을 만들고 잘못 예측된 에세이의 가중치를 높여서 다음 모델을 학습시켰다. 이렇게 반복하여 강력한 모델을 만들어 최종적으로 에세이의 특징들과 이에 대한 각 모델의 예측 결과를 결합하여 AES 문제를 해결했다[2]-[4].

RF는 AES에서 다양한 특징 간의 복잡한 상호작용을 모델링 하는 데 활용했다. 에세이의 다양한 특징들이 서로 영향을 미치며 점수에 기여하는 경우, RF는 이러한 복잡한 관계를 잘 파악할 수 있다. 에세이의 다양한 특징을 활용하여 의사결정 트리를 생성한다. RF는 부트스트랩 샘플링을 통해 여러 개의 트리를 학습한다. 각 트리가 예측한 결과를 평균하거나 다수결로 취합하여 최종 점수를 예측한다. RF는 변수의 중요도를 평가할 수 있어, 어떤 특성이 점수 예측에 영향을 미치는지 파악할 수 있는 장점이 있다.

SVR은 에세이의 특징을 사용하여 SVR 모델을 학습시켰다. SVR은 회귀 모델로, 주어진 특징들에 대한 실제 점수와 모델의 예측값 간의 차이를 최소화하는 초평면을 찾는다. 에세이의 다양한 특징들을 고려하여 회귀 모델을 구축하여, 예측값을 통해 에세이의 점수를 예측했다. 요약하면, GBM, RF, SVR 모델들은 AES 예측 모델의 정확도를 0.6 이상 향상했다.

최근에는 인공지능 분야의 발전과 딥러닝 기술의 부상으로 AES 분야에도 변화가 찾아왔다. 특히, 딥러닝 기술을 활용하여 특성 선택(Feature selection) 없이 AES 모델을 구축하는 연구가 진행되고 있다[5]. 이러한 딥러닝 기반의 AES 모델은 자동으로 텍스트의 패턴과 특징을 학습하므로 다양한 유형의 글을 처리하는 데 더 적합한 결과를 얻을 수 있다. 초기에는 순환 신경망인 RNN(Recurrent Neural Networks)과 LSTM(Long Short-Term Memory)에서 시작하여, 현재는 트랜스포머 기반의 사전학습 모델을 이용한 연구가 활발히 진행되고 있다. 트랜스포머 기반의 BERT(Bidirectional Encoder Representations from Transformers)라는 대표적인 사전학습 모델은 표 1에서 볼 수 있듯이 AES 예측 모델의 정확도를 70%까지 끌어올렸다. 하지만 아직도 30% 성능 향상이 필요하다[6]-[8].

그래서 기존의 BERT 모델을 개선한 모델이 최근 등장하였는데 [9]의 다중계층 BERT 모델(Multi-scale BERT)은 기존의 BERT 모델에 계층을 추가하고 여러 손실함수를 결합한 모델을 만들었다. 표 1을 보면 다중계층 BERT 모델이 기존의 BERT 모델보다 성능이 향상된 것을 알 수 있다. 기존의 BERT 모델은 토큰과 문서 레벨의 2단 구조를 활용하여 에세이의 특징 벡터를 얻었다면 다중계층 BERT 모델은 토큰, 세그먼트, 문서 레벨의 3단 구조를 이용해 에세이의 특징 벡터를 얻어서 학습을 시켰다. 또한, 기존 BERT 모델에서는 손실함수로 MSE(Mean Squared Error)만을 사용했다면 다중계층 BERT 모델은 MSE, MR(Margin Ranking), SIM(Cosine Similarity) 등 3가지 다른 손실함수를 결합하여 성능을 향상하였다. 본 논문은 다중계층 BERT 모델에서 사용된 MR과 SIM을 확장한 대조 학습 기반의 손실함수를 추가하여 에세이 점수 예측에 대한 정확도를 높이고자 한다.

본 연구의 기여도는 다음과 같다.

· AES에서 가장 높은 성능을 보이는 다중계층 BERT 모델에 대조 학습 기반 손실함수를 추가하여 정확도를 향상하고 다양한 손실함수가 정확도 향상에 어떤 영향을 끼치는지를 실험적으로 분석한다.

· 대조 학습 기반 손실함수가 추가된 다중계층 BERT 모델을 기존 BERT 모델과 비교했을 때, 피어슨 상관 계수(Pearson)는 3~5% 증가하였고 QWK(Quadratic Weighted Kappa) 는 3~4% 증가했다.

· 대조 학습 기반의 다중계층 BERT 모델에 배치 크기를 달리하여 비교했을 때, 피어슨 상관계수는 최대 8% 증가하였고 QWK는 최대 5% 증가했다.

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

과거에는 전통적인 AES(Automated Essay Scoring) 모델에서는 기계학습을 활용하여 에세이를 평가하기 위해 단어 수, 단어 길이, 문장 길이 및 문법과 같은 특징을 사람이 직접 추출하는 방식을 사용했다. 또한, AES는 회귀 분석이 이 문제를 해결하는 데 적합하여 베이지안 모델[10][11]과 k-NN 모델, 베이지안 선형 회귀[12], SVD[13]-[15] 방법 등이 성능을 향상했다. 그러나 이러한 기계학습 접근 방식은 수작업이 필요하고 모델의 성능은 추출된 특징에 의존하는 한계가 있다.

따라서 최근에는 딥러닝 기술이 발전하면서 자동으로 에세이와 레이블 된 점수 간의 관계를 학습하기 위해 딥러닝[16]-[22]을 사용한 연구가 진행되고 있다. 초기 연구에서는 RNN과 LSTM과 같은 순환 신경망 모델을 사용했으나 장기 기억의 어려움과 학습 속도 저하와 같은 문제점이 있다[23].

BERT는 트랜스포머의 인코더 구조를 활용한 언어 표현 모델로, AES에 적용하기 위해 여러 개의 BERT를 사용한 연구가 있다[9]. 기존의 BERT 모델은 문서와 토큰 단위를 사용하여 성능을 도출하였다면 [9]의 다중계층 BERT 모델은 문서 단위와 토큰 단위에 세그먼트 단위를 추가하고 다양한 손실함수를 통해 정확도를 10%가량 높였다. 다양한 단위로 에세이를 분할하여 정밀한 점수 산출이 가능하다는 장점이 있지만 각 단위의 에세이 임베딩 표현을 적절히 학습하기 어렵다는 단점이 있다.

### 2.2 CL 모델

최근에는 대조 학습(Contrastive learning)이 이미지 처리 및 자연어처리 분야를 포함하여 다양한 분야에서 활발하게 연구되고 있다. 이미지 처리 분야에서는 대조 학습이 처음으로 도입되었으며, [24]에 따르면 이미지 처리 분야에서 사용된 대조 학습 연구는 [25]-[27]와 같다. 이 연구들은 각 이미지에 대해 자르기, 회전 등의 이미지 변환을 통해 두 가지 이미지를 생성하고 잠재 공간에서 서로 가깝게 만든다. [24]에서는 InfoNCE(정규화된 템퍼레이쳐 스케일 기반의 교차 엔트로피 손실)라는 손실함수를 사용하여 정규화된 임베딩에서 더 좋은 임베딩 표현을 보인다.

최근에는 대조 학습이 자연어처리 분야에서도 활발히 사용되고 있다. [28]에서는 BERT 모델 위에 CNN(Convolutional Neural Network) 계층을 추가하고 글로벌 문장 임베딩과 로컬 문맥 간의 상호 정보(MI, Mutual Information)를 최대화하는 학습 방법을 제안한다. [29]은 [26]와 유사한 구조를 사용하고 데이터 증강을 위해 역번역을 사용하지만, 역번역은 가짜 정보를 생성할 수 있는 단점이 있다. [29]에서는 [25]의 아키텍처를 사용하여 대조 학습과 마스크 언어 모델(Masked language model)을 함께 학습하지만, 의미 유사도를 최대화하는 범위 내에서만 대조 학습을 사용하는 단점이 있다.

제안된 모델은 에세이 임베딩의 향상된 표현을 위해 대조 학습을 사용하여 에세이를 평가한다. 또한, 기존의 대조 학습 연구와는 다르게 양성과 음성 샘플을 구성하기 위해 점수에 따라 에세이 벡터를 평균 내어 점수별 에세이 특징 벡터를 추출하고 중앙값을 기준으로 양성과 음성 샘플을 분리한다. 이는 기존의 대조 학습 연구와는 차별화된 점이다.

## III. 제안방안

제안방안의 3.1 에서는 기존의 BERT 모델과 다중계층 BERT 모델을 비교하며 설명하고 3.2 에서는 4가지 손실함수를 비교하며 자세히 설명한다.

### 3.1 다중계층 BERT 모델의 형태

이 절에서는 먼저 기존의 BERT 모델을 설명하고 [9]의 다중계층 BERT 모델을 비교하며 설명하겠다.

그림 1은 기존의 BERT 모델의 계층이다. 그림 1를 보면 기존의 BERT 모델은 에세이가 입력으로 들어오면 전체 문서를 학습한 전체 문서 벡터와 토큰별로 학습한 토큰 벡터를 이용해 에세이의 특징 벡터를 얻었다.

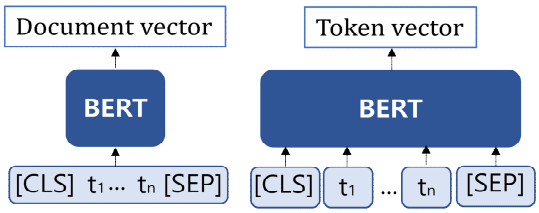



그림 1. 기존의 BERT 모델
Fig 1. Traditional BERT model

그림 2는 다중계층 BERT 모델의 계층이다. 그림 2를 보면 다중계층 BERT 모델은 기존의 전체 문서 벡터와 토큰 벡터뿐만 아니라 세그먼트 벡터의 정보가 추가된 것을 볼 수 있다. 세그먼트 벡터는 에세이 데이터가 있다면 주어진 세그먼트 값에 따라 전체 토큰을 세그먼트 값씩 잘라 여러 세그먼트로 나눈 후 BERT의 입력 데이터로 학습한 특징 벡터이다. 예를 들어 150개의 토큰이 있다면 세그먼트 값이 30이라면 150개의 토큰을 30개씩 자른다. 30개씩 잘린 세그먼트 데이터들을 BERT 모델의 입력으로 들어가면 세그먼트 벡터가 생성된다. 그래서 다중계층 BERT 모델은 토큰, 문서, 그리고 세그먼트 레벨의 3단 구조를 통해 학습한다.

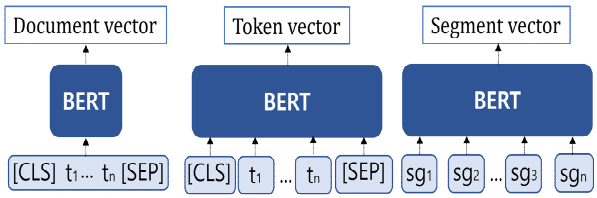



그림 2. 다중계층 BERT 모델
Fig 2. Multi-scale BERT model

### 3.2 손실함수

이 절에서는 다른 목적으로 사용되는 4가지 손실함수를 소개한다. 기존의 BERT 모델은 손실함수로 평균제곱오차(MSE)만을 사용하여 손실함수를 계산했다. 다중계층 BERT에서는 손실함수를 MSE뿐만 아니라 코사인 유사도(SIM)와 순위 손실함수(MR)를 결합한 손실함수를 이용하여 학습을 진행한다.

먼저, MSE는 예측 점수와 실제 점수 사이의 오차를 제곱하여 평균값을 측정한 것으로 식 (1)과 같이 정의한다.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(P(x) - Q(x))^2 \quad (1)$$

P(x)와 Q(x)는 각각 $i$번째 에세이의 예측 점수와 실제 점수이고, N은 에세이의 개수이다.

두 번째 손실함수인 SIM은 두 벡터 간의 코사인 각도를 이용하여 두 벡터의 유사도를 측정하는 것으로 식 (2)와 같이 정의한다.

$$\sim = 1 - \cos(P(x), Q(x)) \quad (2)$$

SIM은 유사한 벡터 쌍에 큰 값을 부여하여 모델이 에세이 배치 간의 상관관계에 대해 더 많이 고려되도록 한다.

세 번째 손실함수인 MR은 배치의 각 에세이 쌍에 대한 순위를 측정한다. 에세이의 정렬 속성이 채점의 핵심 요소이기 때문에 MR 손실을 도입하여 식 (3)과 같이 정의한다.

$$MR = \frac{1}{\hat{N}}\sum \max(0, -r_{i,j}(y_i - y_j) + b) \quad (3)$$

$$r_{i,j} = \begin{cases} 1 & \hat{y_i} > \hat{y_j} \\ -1 & \hat{y_i} < \hat{y_j} \\ -sgn(y_i - y_j) & \hat{y_i} = \hat{y_j} \end{cases}.$$

$y_i$와 $\hat{y_i}$는 각각 $i$번째 에세이의 예측 점수와 실제 점수이다. $\hat{N}$은 에세이 쌍의 수이다. $b$는 하이퍼 매개변수이다. 실험에서는 0으로 설정했다.

마지막 손실함수로는 대조 학습을 사용한다. 대조 학습은 이미지 처리나 자연어처리 등에 연구되고 있는 기술로 의미가 유사한 쌍의 벡터는 서로 가깝게 유도되고 의미가 다른 쌍의 벡터는 멀어지도록 유도하여 훈련하는 모델이다.

[30]에서 대조 학습의 양성과 음성 샘플을 구하기 위해 점수별 평균 임베딩 방법을 사용했다. 점수별 평균 임베딩을 사용한 이유는 에세이의 내용이 아닌 점수별 에세이의 공통적인 문법과 문체를 고려한 특징 벡터를 추출하기 위해서다.

첫 번째 미세조정 단계에서 BERT 모델을 사용하여 추출된 벡터를 점수별로 리스트를 만든 후 해당 리스트를 이용해 점수별 평균벡터를 생성한다. 여기서 평균벡터는 모든 단어 벡터를 참고한 문맥 벡터의 역할을 한다. 대조 학습의 양성 샘플과 음성 샘플을 선택하는 방법은 중앙값을 기준으로 점수 범위를 나눈다. 그 후 입력 데이터가 들어오면 양성 샘플은 입력 데이터와 같은 점수의 평균 벡터를 선택하고 음성 샘플은 다른 점수 범위에서 무작위로 하나를 점수의 평균벡터로 선택한다. 하지만 표 2의 ASAP 데이터 세트의 점수의 범위를 보면 광범위한 점수 범위를 가지고 있는 프롬프트 1, 7, 8은 점수별 평균 임베딩 방법을 구하기 위해 아래와 같은 추가적인 방법을 사용한다. 프롬프트 1은 2~12까지의 점수 범위를 가지고 있고 프롬프트 7은 0~30까지의 점수 범위를 가지고 있다. 또한, 프롬프트 8은 0~60까지의 점수 범위를 가지고 있다. 그래서 광범위한 점수를 범위를 갖는 프롬프트는 그림 3과 같이 점수 범위를 6등분하여 대조 학습 시에 1~6의 점수로 변환한다.

표 2. ASAP 데이터 세트
Table 2. ASAP data set

| Prompt | Essays | Score range |
|---|---|---|
| 1 | 1783 | 2~12 |
| 2 | 1800 | 1~6 |
| 3 | 1726 | 0~3 |
| 4 | 1772 | 0~3 |
| 5 | 1805 | 0~4 |
| 6 | 1800 | 0~4 |
| 7 | 1569 | 0~30 |
| 8 | 723 | 0~60 |

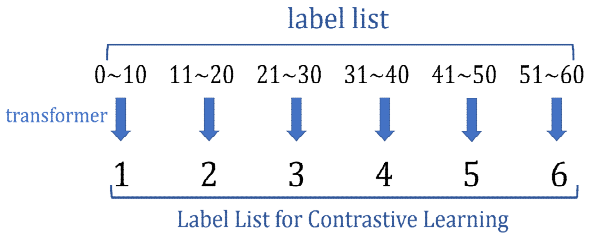



그림 3. 광범위한 점수 변환 예시
Fig 3. Extensive example of score conversion

대조 학습 기반의 손실함수를 적용하기 위해 두 단계의 미세조정을 시행한다. 1단계 미세조정에서는 식 (4)과 같이 교차 엔트로피 손실함수만을 사용한다. 이때 $P(x)$는 실제 라벨값을 나타내고 $Q(x)$는 모델이 추정한 확률값을 나타낸다.

$$Loss_{CE} = -\sum P(x)\log Q(x) \tag{4}$$

교차 엔트로피 손실함수는 각 클래스에 대한 추정 확률값과 실제 정답 클래스에 대한 확률값 간의 차이를 최소화한다. 이 손실함수를 사용하여 에세이와 레이블링 된 점수 사이의 거리를 줄여, 입력값으로 들어간 에세이가 적절한 레이블로 분류되도록 학습을 진행하게 된다.

2단계 미세조정에서 식 (6)과 같이 기존의 교차 엔트로피 손실함수에 대조 학습 손실함수를 추가한다. $e_i$는 입력 데이터의 임베딩, $e_p$와 $e_n$은 각각 양성 샘플과 음성 샘플을 의미한다. $dist$는 유클리드 거리와 $sim$은 코사인 유사도를 의미한다. $\lambda_1$과 $\lambda_2$는 초매개변수이며 모든 실험에서 각각 5와 10으로 설정하였다.

$$Loss_{CL} = \lambda_1 * p + \lambda_2 * q \tag{5}$$
$$p = \frac{dist(e_i, e_p)}{\{dist(e_i, e_p) + dist(e_i, e_n)\}}$$
$$q = |sim(e_i, e_n)|$$

식 (5)에서 대조 학습 손실함수는 $p$와 $q$로 이루어져 있다. 입력 데이터의 임베딩이 양성 샘플과 유클리드 거리가 작고 음성 샘플과는 클수록 $p$의 값이 작아지게 된다. 한편 입력 데이터의 임베딩이 음성 샘플과 코사인 유사도가 작을수록 $q$의 값이 작아지게 된다. 즉, 입력 데이터 임베딩이 양성 샘플과는 유사하고 음성 샘플과는 상이할수록 손실 값이 작아진다.

$$Loss_{TOTAL} = \alpha * Loss_{CE} + (1-\alpha) * Loss_{CL} \tag{6}$$

전체 손실함수($Loss_{TOTAL}$)는 식 (6)과 같이 교차 엔트로피 손실함수($Loss_{CE}$)에 대조 학습 손실함수($Loss_{CL}$)를 더한 값이 된다. $\alpha$는 초매개변수이며 모든 실험에서 0.5로 설정하였다.

제안한 모델의 손실함수는 식 (7)와 같다. $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$는 데이터 세트의 성능에 따라 조정되는 하이퍼 파라미터다.

$$Loss = \beta_1 * MSE + \beta_2 * MR + \beta_3 * \sim + \beta_4 * CON \tag{7}$$

## Ⅳ. 실 험

### 4.1 실험 환경

표 3는 실험에 사용된 컴퓨터 환경과 BERT 딥러닝 모델 학습에 관한 내용이다. 실험에서는 Python 3.8, PyTorch 1.10 및 CUDA 11.1을 이용하여 BERT 모델을 학습했다. 사전 학습 모델로는 BERT-base를 활용했다.

표 3. 실험 환경
Table 3. Experimental set-up

| Hardware | Specification |
|---|---|
| CPU | Intel(R) Core™ i9-10940X (3.30GHz, core: 14) |
| GPU | NVIDIA GeForce A 6000 |
| SSD | 500GB |
| OS | Window 10 |

또한, 표 4는 BERT 모델의 미세조정 단계에 적용된 초매개변수(Hyperparameters)를 보여준다. 표4에 제시된 파라미터 값을 사용하였을 때, 가장 우수한 결과를 얻을 수 있다. 다만, 해당 에폭(epoch)을 초과하게 되면 과적합으로 인해 모델의 정확도가 감소하는 현상이 발생한다. 따라서 적절한 에폭을 선택하여 모델을 학습하는 것이 중요하다.

표 4. 미세조정 단계 초매개변수
Table 4. Fine-tuning hyperparameters

| Hyperparameter | Fine-tuning step |
|---|---|
| Dropout rate | 0.1 |
| Batch size | 24 |
| Learning rate | 6e-5 |
| Max length | 512 |
| Epochs | 80 |
| $\lambda_1$ | 5 |
| $\lambda_2$ | 10 |
| $\beta_1$ | 2 |
| $\beta_2$ | 1 |
| $\beta_3$ | 1 |
| $\beta_4$ | 1 |

모델의 성능을 측정하기 위해 QWK(Quadratic Weighted Kappa)을 사용한다. QWK는 두 명 이상의 평가자들이 같은 대상을 평가하는 경우의 일치도를 측정하기 위한 통계적 지표이다. QWK는 일치도를 -1부터 1 사이의 값으로 나타내며, 1에 가까울수록 평가자들 간의 높은 일치도를 의미한다. 0 또는 음수의 값은 불일치 정도를 의미한다.

QWK 계산 방법으로는 NxN 히스토그램 행렬 $O$를 구성한 후 $O_{i,j}$에 $i$(실제 값)과 $j$(예측된 값)의 차이를 기반으로 NxN 무게 행렬 $w_{i,j}$를 식 (8)과 같이 계산한다.

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{8}$$

그 후 NxN 히스토그램 행렬 $E$를 계산하고, 마지막으로 세 개의 행렬을 식 (9)를 대입해 QWK를 계산한다.

$$qwk = 1 - \frac{\sum w_{i,j} O_{i,j}}{\sum w_{i,j} E_{i,j}} \tag{9}$$

또한, 피어슨 상관 계수(PCC, Pearson Correlation Coefficient)는 두 변수 간의 선형 관계의 강도와 방향을 측정하기 위해 사용하는 지표이다. 상관 계수의 값도 - 1부터 1까지의 범위를 가지고 있다. 1에 가까울수록 두 변수는 양의 선형 관계를 갖는다. 하나의 변수가 증가하면 다른 변수도 증가하는 경향이 있다는 것이다. -1에 가까워진다면 두 변수는 음의 선형 상관관계를 갖는다.

하나의 변수가 증가할 때 다른 변수도 감소하는 경향이 있다는 것을 의미한다. 0에 가깝다면 두 변수 산의 선형 상관관계는 약한 것으로 간주한다. 피어슨 상관 계수는 두 변수 간의 선형 관계만을 측정한다. 만약 두 변수 간의 관계가 비선형적이거나 다른 형태의 관계를 갖는다면 상관 계수는 정확한 결과를 제공하지 못한다. 상관 계수의 공식은 식 (10)으로 계산한다.

$$pea = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{10}$$

$x_i$와 $y_i$는 각각 두 변수의 관측값을 나타내며, $\bar{x}$와 $\bar{y}$ 는 각각 두 변수의 평균값을 나타낸다.

## 4.2 실험 결과

본 실험에서는 세 개의 실험을 진행했다. 첫 번째 실험은 기존의 BERT 모델의 성능과 다중계층 BERT 모델의 성능을 비교하여 어떤 영향을 미치는지 알기 위해 비교 실험했다. 또한, BERT 모델과 다중계층 BERT 모델의 손실함수를 변경하거나 본 연구에서 제안한 대조 학습을 추가하며 성능에 어떤 영향을 주는지 비교 실험하였고 두 번째 실험은 프롬프트별 실험을 진행하여 다중계층 BERT 모델과 제안한 모델을 성능을 비교하였다. 세 번째 실험은 제안한 모델에 배치 크기를 변경하였을 때 성능에 어떤 영향을 주는지 비교 실험했다.

그림 4는 프롬프트 2를 활용하여 기존의 BERT 모델과 다중계층 BERT 모델의 성능을 손실함수를 변경하며 비교한 결과이다. 그림 4에서는 피어슨 상관 계수와 QWK 사용하여 모델의 성능 평가하였다.

그림 4를 보면 'MSE'는 손실함수인 평균제곱오차만을 활용하여 성능을 측정한 것이다. 'MSE+SIM'는 MSE와 SIM 손실함수를 결합하여 성능을 측정한 것이고 'MSE+MR'은 MSE와 MR 손실함수를 결합하여 측정한 것이다.

‘MSE+SIM+MR’은 MSE, SIM, 그리고 MR인 세 가지 손실함수를 모두 결합하여 측정한 것이다. ‘MSE+SIM+MR+CL’은 기존의 세 가지 손실함수인 MSE, SIM, MR에 대조 학습을 추가하여 측정한 것이다.

그림 4의 손실함수로 MSE만을 사용하여 성능을 비교한 결과를 보면 기존의 BERT 모델의 피어슨 상관 계수는 0.664를 보여준다. 그런데 대조 학습 기반의 다중계층 BERT 모델은 0.695로 기존의 BERT 모델보다 3~4%가 상승한 것을 볼 수 있다. 또한 QWK를 이용해 비교한 결과를 봐도 기존의 BERT 모델은 0.652를 보여주지만 대조 학습 기반의 BERT 모델은 0.681로 3~4%가 상승한 것을 볼 수 있다. 이로써 에세이를 평가할 때 단어 규모와 문서 규모의 척도만을 이용하여 예측한 것보다 세그먼트 규모의 척도를 포함한 것이 성능이 향상된 것을 볼 수 있다. 이는 문서 규모와 단어 규모의 정보만을 이용하여 점수를 매기는 것보다 세그먼트 규모의 정보를 이용하여 점수를 매겼을 때 성능 향상에 효과적인 것을 보여준다.

그림 4의 여러 손실함수를 결합하며 비교한 결과를 보면 손실함수에 대조 학습을 추가했을 때 가장 성능이 좋았다. 대조 학습을 사용한 기존의 BERT 모델의 피어슨 상관 계수는 0.707를 보여주지만 대조 학습 기반의 다중계층 BERT 모델은 0.744로 피어슨 상관 계수를 사용해 비교한 결과 중 가장 좋은 결과를 보여준다. 또한 QWK를 이용해 비교한 결과를 보면 기존의 BERT 모델은 0.697이지만 대조 학습 기반의 다중계층 BERT 모델은 0.714로 QWK도 대조 학습을 사용한 후 가장 높은 성능을 보여줬다. 이는 기존의 문제점인 서로 의미가 다르더라도 높은 유사성을 보여주는 문제점을 해결한 것을 알 수 있다.

표 5는 다중계층 BERT 모델과 제안한 모델의 프롬프트 1~8까지에 대한 성능을 비교한 결과이다. 손실함수인 MSE와 SIM, MR을 결합한 ‘MSE+SIM+MR’와 결합한 손실함수에 대조 학습을 추가한 ‘MSE+SIM+MR+CL’를 프롬프트별 피어슨 상관 계수와 QWK를 사용해 성능을 비교한 것이다.

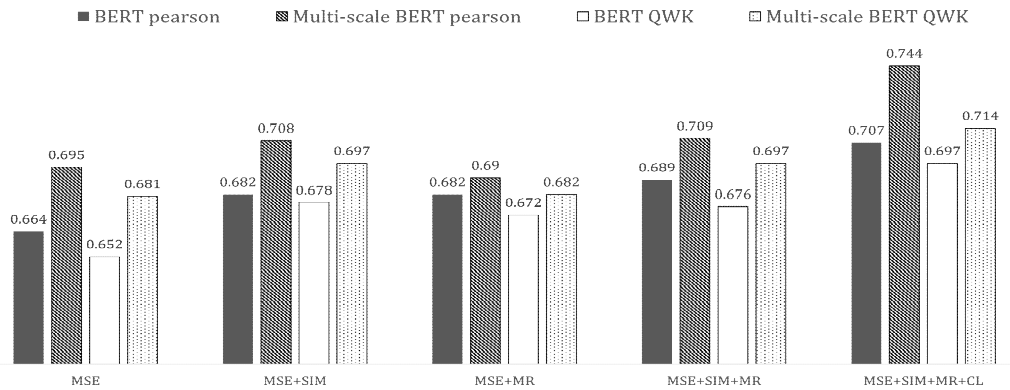



그림 4. BERT 모델과 다중계층 BERT 모델 성능 비교
Fig 4. Comparison of BERT model and multi-scale BERT model performance

표 5. 프롬프트별 실험 결과
Table 5. Experiment results of CL-based multi-scale BERT per pPrompt

| Evaluation metrics | Loss function | Prompt | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| pearson | MSE+SIM+MR | 0.845 | 0.707 | 0.704 | 0.818 | 0.812 | 0.822 | 0.843 | 0.745 |
| | MSE+SIM+MR+CL | **0.867** | **0.744** | **0.746** | **0.844** | **0.837** | **0.843** | **0.856** | **0.786** |
| QWK | MSE+SIM+MR | 0.822 | 0.697 | 0.700 | 0.809 | 0.795 | 0.805 | 0.827 | 0.737 |
| | MSE+SIM+MR+CL | **0.846** | **0.714** | **0.730** | **0.819** | **0.818** | **0.829** | **0.836** | **0.739** |

표 5의 프롬프트 1에서 피어슨 상관 계수를 이용해 비교한 결과를 보면 손실함수에 대조 학습을 추가하지 않은 'MSE+SIM+MR'은 0.845를 보여준다. 하지만 대조 학습을 추가한 'MSE+SIM+MR+CL'은 0.867로 약 3% 향상된 것을 볼 수 있다. QWK를 이용해 비교한 결과를 보면 'MSE+SIM+MR'은 0.822를 보여주지만, 대조 학습을 추가한 후 0.846으로 약 3%의 성능 향상을 보여준다. 특히 점수 범위가 적은 프롬프트 1~6은 피어슨 상관 계수는 약 3~5%의 성능이 향상된 것을 보여준다. 그리고 QWK는 3~4%의 성능이 향상된 것을 보여준다.

하지만 점수 범위가 넓었던 프롬프트 7, 8은 피어슨 상관 계수는 약 1~4%를 보여주고 QWK는 약 1~2%의 성능 향상을 보여준다. 이는 손실함수로 대조 학습을 사용할 시 성능이 향상된 것을 보여주지만, 점수 범위가 좁은 형태를 구분할 때 훨씬 더 효과적인 것을 알 수 있었다.

표 6은 프롬프트 2를 활용하여 대조 학습 기반의 다중계층 BERT 모델에 배치 크기를 달리했을 때의 QWK과 피어슨 상관관계로 성능을 비교한 것이다. 표 6을 보면 배치 크기가 4였을 때는 피어슨 상관 계수는 0.658, QWK는 0.646으로 가장 낮은 성능을 보여준다. 배치 크기를 10으로 늘린 후 성능을 비교한 결과 피어슨 상관 계수는 0.696, QWK는 0.653으로 전 배치 크기와 비교하면 성능이 각각 4%, 1% 정도 향상된 것을 볼 수 있다. 또한, 배치 크기를 20으로 늘린 후 피어슨 상관 계수는 0.728, QWK는 0.700으로 성능은 3%, 5%로 대폭 성능이 향상되었다. 최대 배치 크기인 24로 늘린 후 피어슨 상관 계수는 0.744, QWK는 0.714으로 성능이 2%, 1%로 미묘하게 향상되었다. 결국, 최소 배치 크기인 4와 최대 배치 크기인 24의 성능을 비교하면 피어슨 상관 계수는 8% 향상되었고 QWK는 5% 향상되었다. 배치 크기가 작으면 노이즈와 변동성이 크게 영향을 주지만 배치 크기가 커지면 이러한 노이즈와 변동성의 영향을 상쇄시킬 수 있고 대조 학습 시 더 많은 데이터를 고려하여 더 정확한 패턴을 학습할 수 있다. 하지만 배치 크기가 20이 넘어가면 성능 향상의 차이가 급격히 줄어드는 것을 볼 수 있다.

표 6. 배치 크기별 실험 결과
Table 6. Experiment results of CL-based multi-scale BERT per batch size

| Evaluation metrics | Batch size | | | |
|---|---|---|---|---|
| | 4 | 10 | 20 | 24 |
| pearson | 0.658 | 0.696 | 0.728 | 0.744 |
| QWK | 0.646 | 0.653 | 0.700 | 0.714 |

표 7은 대조 학습 기반 손실함수에 필요한 양성과 음성 샘플 예시이다. 입력 에세이, 양성 및 음성 샘플의 점수는 각각 5점, 5점, 2점이다. 입력 에세이와 양성 샘플은 에세이 품질면에서 유사하지만 음성 샘플은 큰 차이가 있다는 것을 알 수 있다.

## Ⅴ. 결론 및 향후 과제

본 논문은 최근 AES 작업에 가장 높은 성능을 보이는 다중계층 BERT 모델의 손실함수에 대조 학습 기반 손실함수를 추가하여 에세이 점수 예측에 대한 정확도를 더욱 향상했다. 실험 결과에 따르면, 기존 BERT 모델보다 제안방안은 QWK는 3~4%, 피어슨 상관 계수는 3~5% 향상되었다. 특히 점수 범위가 좁은 에세이에 효과적인 것을 알 수 있다. 또한, 다양한 손실함수 중에서 대조 학습 기반의 손실함수가 AES 모델의 정확성을 향상하는데 가장 크게 기여한다는 사실을 보였다.

현재 AES 문제는 에세이 텍스트가 입력으로 주어지면 그 에세이에 대한 전체적인 점수를 예측한다. 향후 연구로는 ASAP++ 벤치마크 데이터 세트를 사용하여 채점표에 있는 각 채점 항목에 대한 개별 점수를 예측하고 이를 바탕으로 전체적인 점수를 예측할 수 있도록 제안방안을 확장하고 음성 샘플과 양성 샘플 선택하는 기준을 사용자가 직접 정해줬다. 하지만 점수 범위가 넓을 시 대조 학습을 진행할 때는 성능 개선이 필요하다. 넓은 점수 범위를 갖는 에세이에 대조 학습을 진행할 경우 해당 점수 범위의 최적 파라미터를 찾을 수 있도록 확장할 것이다.

표 7. ASAP 데이터 세트 양성 샘플과 음성 샘플 예시
Table 7. Positive and negative samples of the ASAP dataset

| Type | Essay | Score |
|---|---|---|
| Input essay | Censorship, or the removing of material deemed inappropriate or offensive, has been used as a means of controlling people's thoughts and feelings for many centuries. The @CAPS1 used it in the @CAPS2 colonies during the @CAPS2 @CAPS4 to keep the colonists loyal to the crown. The @CAPS5 used censorship not just during the @CAPS6 @CAPS7 @CAPS8 as a brand of propaganda, but also during @CAPS7 @CAPS8 @CAPS11 under @CAPS12 guidance. Censorship is still in use in the @LOCATION1 today, as a way to 'protect the children'. The @ORGANIZATION1 tells the @CAPS2 public what we can and cannot watch on television and what we can and cannot listen to on the radio. Other organizations tell us what books we can and cannot read, what video games we can and cannot play, and what items we can and cannot buy. Such organizations, I believe, are pushing the @LOCATION1 @LOCATION1 and other freedom-loving nations to the edge of totalitarianism. As Katherine Paterson stated, 'If I have the right to remove that book from the shelf-that work I abhor-then you also have exactly the same right and so does everyone else.' If one person has the right to control what we buy, what we listen to, and what we see, then how come we don't have that right to control the same things for them? Where do we draw the line? Eventually, all books, movies, magazines, music, and video games would be deemed offensive and inappropriate. Culture would slowly wither and die, and our lives would become drab and essentially useless, for lack of a better term. I recently read a book titled '@CAPS14', by @PERSON2, about two periods in a boy's life: his childhood with his family in tribal @LOCATION2, and the life he had after his mother's death in the @CAPS15 capital of @LOCATION3. @PERSON2's work includes some passages that are less than angelic, and, had strict censorship been in place, these passages would have caused the book to go unpublished. @PERSON2's work gave not only myself but a multitude of others insight into the life that @CAPS16 face in the changing landscape of their @CAPS7, and it certainly made me realize once again that I have a life rich in family, friends, and comfort. All in all, censorship is a horrid disease that slowly infects and consumes every part of our lives. If you get to censor something you find offensive, then why don't I do the same thing? Some censorship is good, and a fact of life. But we have to ask ourselves, where do we draw the line? Where's the boundary between safety and and outright police state? Censorship is just a fancy term for blatant disregard of the @CAPS17, as freedom of speech is guaranteed for all @CAPS2 citizens. As a famous colonial revolutionary once said, '@CAPS19 me liberty, or @CAPS19 me death!' If censorship continues to grow and our libraries grow emptier and emptier, then I shall take death over liberty, for liberty with censorship is no liberty at all. | 5 |
| Positive sample | Censorship in the world we live in today is a very interesting concept. With the advent of social media, we now have thriving communities of people that want their voice to be heard, and people are listening. We now are able to tap into the huge amount of knowledge from people all around the world, allowing us to see what is really happening in the countries like @LOCATION2, where media is considered treason. What do we do now that governments are starting to censor the internet? Lets start by looking an arguement that has been long talked about: library censorship. Today, it is almost a requirement for cities to have public libraries to be considered as real cities. Public libaries, and the funding they require, are not anywhere within the @ORGANIZATION1. Even so, they have taken a place in our neighborhoods that has made them indefinate sources of knowledge and culture. One public library @MONTH1 be a source of; historical town records, children's books, novels, research papers, non-fiction books, fiction books, magazines, and many other types of entertainment and knowledge. Public libraries are, even with the widespread usage of the internet, important places for us, as residents of a city, to socialize and learn. Libraries today @MONTH1 carry books that were in the past on the 'banned books' list, but that does not mean books are not being banned from libraries. Libaries are generally family institutions, and as being funded by the public, required to adhear by the moral and ethical standards of those who live in the area. This means that there is not one, 'national banned books list', but there many discrepencies througout @LOCATION1 in banning books. As culture moves more and more into more and more questionable materials, libaries are faced with a tough decision; keep up with the times and have more questionable and controversial reading material, or get lost in the past and try to kling onto the senior generation to stay alive. Bringing new literary materials into the library that is fresh and new means that more people will return to the library again, be it young-adults or older adults trying to get in touch with the current world. One major controversy of books in public libaries are the inclusion of sexual intimacy. As a new generation sprouts, fed with music lyrics by such names as; '@ORGANIZATION2', or '@ORGANIZATION3', we enter a new age where younger and younger kids are learning about sex. The authors of today have recognised this and have started to include references to sexual intimacy, but not detailed encoutners. I believe this is fine, because as we move on in generations we are going to find out that education about sex; safe sex, birth control, @CAPS1 prevention and testing, is much more important than trying to keep children ignorant. Keeping these young adults involved in books and trying to continue their learning outside of school is a much larger issue we should be trying to tackle instead of trying to create silly arguments against new reading material in a public libary. Public libaries are the centerpiece for learning outside of school in many communities, small and large. Knowing this, how could anyone truly believe that it would be smart to stop including new books that would entice new patrons? Even as a commercial buisness, this would make no sense, simply turning people away because you have a single view of what the future is and you refuse to listen to anyone else's opinion of the future. So why is there a question on including new knowledge? Isn't it obivous that we, as a nation, should show the world that we aren't living in the past. That we are indeed the future of the world, and that our young-adults will recieve the best education in the world by not only learning in schools, but by being the most active readers, and by extension, learners in the world. | 5 |
| Negative sample | I believe that they need to keep every thing and gain more. Their is so much to learn about. The more information you have the better. I for one think that they need to put more magazines on the stands so you can have more information. How you can not trust every thing in one, You can still find many facts about what your wanting to learn about. On the other hand, Children should not be able to reach or see some of this. I could teach our children stuff that they are not yet ready to learn about. The fact is they need to have people watching the stuff so kids wont get in it and we can still lurean and find out new things. Not everything should be hidden. The more you know could just save the world | 2 |

## References


[1] Kaggle, "The Hewlett Foundation: Automated Es-say Scoring", https://www.kaggle.com/competiti-ons/asap-aes/data [accessed: Feb. 18, 2023]

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

[3] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest", Information Computing and Applications: Third International Conference, Vol. 7473, pp. 246-252, Sep. 2012. https://doi.org/10.1007/978-3-642-34062-8_32.

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

[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] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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.

[7] E. Mayfield and A. W. Black, "Should you fine-tune BERT for automated essay scoring?", Proc. of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 151-162, Jul. 2020. http://dx.doi.org/10.18653/v1/2020.bea-1.15.

[8] R. Yang, et al., "Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking", Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1560-1569, Nov. 2020. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.141.

[9] 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", arXiv p-reprint arXiv:2205.03835, May 2022. https://doi.org/10.48550/arXiv.2205.03835.

[10] L. S. Larkey, "Automatic essay grading using text categorization techniques", Proc. of the 21st annual international ACM SIGIR conference on R-esearch and development in information retrieval, pp. 90-95, 1998.

[11] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' theorem", The Journal of Technology, Learning and Assessment, Vol. 1, No. 2, Jun. 2002.

[12] P. Phandi, K. M. A. Chai, and H. T. Ng, "Flexible d-omain adaptation for automated essay scoring usin-g correlated linear regression", Proc. 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.

[13] 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, 2015.

[14] 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.

[15] A. A. P. Ratna, H. Khairunissa, A. Kaltsum, I. Ibrahim, 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/ICECOS4763 7.2019.8984528.

[16] H. Nguyen and L. Dery, "Neural networks for automated essay grading", CS224d Stanford Reports, pp. 1-11, 2016.

[17] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring", Proc. of the 2016 conference on empirical methods in natural language processing, Austin, Texas, USA, pp. 1882-1891, Nov. 2016.

[18] J. Yang, Y. Zhang, and S. Liang, "Subword encoding in lattice LSTM for Chinese word segmentation", arXiv preprint arXiv:1810.12594, Oct. 2018. https://doi.org/10.48550/arXiv.1810.12594.

[19] J.-U. Heu, "News Recommendation Exploiting Document Summarization based on Deep Learning", The Journal of the Institute of Internet, Broadcasting and Communication, Vol. 22. No. 4, Aug. 2022. https://doi.org/10.7236/JIIBC.2022.22.4.23.

[20] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks", arXiv preprint arXiv:1606.04289, Jun. 2016. https://doi.org/10.48550/arXiv.1606.04289.

[21] F. Dong and Y. Zhang, "Automatic features for essay scoring–an empirical study", Proc. of the 2016 conference on empirical methods in natural language processing, Austin, Texas, pp. 1072-1077, Nov. 2016.

[22] U. Mushtaq and J. Cabessa, "Argument Mining with Modular BERT and Transfer Learning", 2023 International Joint Conference on Neural Networks IJCNN, Gold Coast, Australia, pp. 1-8. Jun. 2023. https://doi.org/10.1109/IJCNN54540.2023.10191968.

[23] A. Vaswani, et al., "Attention is all you need", Advances in neur-al information processing systems 30, 2017.

[24] 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, Vol. 2, pp. 1735-1742, Jun. 2006. https://doi.org/10.1109/CVPR.2006.100.

[25] T. Chen, S. Kornblith, and M. Norouzi, "A sim-ple framework for contrastive learning of visual r-epresentations", International conference on machine learning, Vol. 119, pp. 1597-1607, Nov. 2020.

[26] 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.

[27] 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, pp. 22243-22255, 2020.

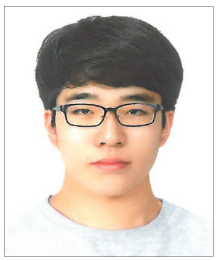
[28] Y. Zhang, R. He, Z. Liu, K. H. Lim, and L. Bing, "An unsupervised sentence embedding method by mutual information maximization", arXiv prepr-int arXiv:2009.12061, Sep. 2020. https://doi.org/10.48550/arXiv.2009.12061.

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

[30] D. Yu, Y. Kim, Sangwoo Han, and B.-W. On, "CLES-BERT: Contrastive Learning-based BERT Model for Automated Essay Scoring", The Journal of Korean Institute of Information Technology, Vol. 21, pp. 31-43, Apr. 2023. http://dx.doi.org/10.14801/jkiit.2023.21.4.31.

## 저자소개


### 한 상 우 (Sangwoo Han)



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

### 유 대 곤 (Daegon Yu)

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

### 이 인 규 (Ingyu Lee)

2007년 : 펜실베이니아주립대학교 컴퓨터공학과 박사
2007년 ~ 2013년 : 앨라바마 트로이대학교 조교수
2013년 ~ 2014년 : 차세대융합기술원 책임연구원
2015년 ~ 2020년 : 앨라바마 트로이대학교 부교수
2021년 ~ 현재 : 영남대학교 정보통신연구소 연구교원
관심 분야 : 선형대수학, 수치해석, 소셜미디어분석, 빅데이터, 머신러닝