에세이 자동 평가 모델 성능 향상을 위한 데이터 증강과 전처리

고강희○, 김도국
인하대학교
ghko99@gmail.com○, dgkim@inha.ac.kr

Data Augmentation and Preprocessing to Improve Automated Essay Scoring Model

Kanghee Go○, Doguk Kim
Inha University

요 약

데이터의 품질과 다양성은 모델 성능에 지대한 영향을 끼친다. 본 연구에서는 Topic을 활용한 데이터 전처리와 BERT 기반 MLM, T5, Random Masking을 이용한 증강으로 데이터의 품질과 다양성을 높이고자 했으며, 이를 KoBERT 기반 에세이 자동 평가 모델에 적용했다. 데이터 전처리만 진행했을 때, Quadratic Weighted Kappa Score(QWK)를 기준으로 모델이 에세이의 모든 평가 항목에 대해 베이스라인보다 더욱 높은 일치도를 보였으며 평가항목별 일치도의 평균을 기준으로 0.5368029에서 0.5483064(+0.0115035)로 상승했다. 여기에 제안하는 증강 방식을 추가 할 경우 MLM, T5, Random Masking 모두 성능 향상 효과를 보였다. 특히, MLM 데이터 증강 방식을 추가로 적용하였을 때 최종적으로 0.5483064에서 0.55151645(+0.00321005)으로 상승해 가장 높은 일치도를 보였으며, 에세이 총점의 QWK를 기준으로 성능을 평가하면 베이스라인 대비 0.4110809에서 0.4380132(+0.0269323)로의 성능 개선이 있었다.

주제어: 에세이 자동 평가, 데이터 증강, 데이터 전처리

1. 서론

최근 NLP 연구에서는 GPT, BERT, BART, T5와 같은 Transformer 기반 언어 모델의 등장으로 텍스트 자동화 기술에 관한 연구가 활발하게 이루어지고 있다 [1,2,3,8]. 그리고 한국어에 대해서는 KoBERT와 같이 한국어 텍스트에 대해 사전 훈련된 언어 모델이 자동화된 NLP 모델을 구축하는데 새로운 기회를 제공하였다 [1]. 한편 국내의 교육 평가 시스템과 관련하여 서·논술형 평가 수요가 늘어남에 따라 에세이 자동 평가(Automated Essay Scoring; AES) 기술의 중요성이 계속 증가하는 추세이다 [2]. AES는 이미 해외 NLP 분야에 있어서 매우 중요한 연구 중 하나이며 국내에서도 다양한 AES 모델들이 제시되어 점점 많은 연구가 이루어지는 추세이다 [1,2,3,6].

NLP 모델의 성능은 모델 자체의 구조 뿐 아니라 데이터의 다양성과 품질에 매우 지대한 영향을 받는다. 한국어는 영어에 비해 NLP 분야에 이용할 만한 텍스트 데이터가 상대적으로 부족한 상황이다 [4]. 이에 따라 국내에서는 다양한 한국어 텍스트 데이터 증강 방식이 제시된 바 있다 [4,5]. 한국어 텍스트 데이터 증강은 전통적으로 동의어 교체(SR), 단어삽입/삭제 (RI, RD), 단어 위치 바꾸기(RS), 노이즈 추가 등이 있고, 최근에는 다양한 언어 모델을 활용한 데이터 증강 방식이 제시되곤 한다 [4,5]. 또한 해외에서는 에세이의 Topic을 활용한 데이터 전처리를 통해 AES 모델의 성능을 높이려는 연구가 이루어진 바 있다 [6].

하지만 국내 AES 연구에서는 이러한 데이터 증강 방식과 전처리를 AES 모델에 적용하는 연구가 활발하게 이루어지지 않았다. 따라서 본 연구에서는 영어 에세이에 적용했던 데이터 전처리 방식을 한국어 에세이에 적용해 데이터의 품질을 높이고자 했으며, 데이터 증강을 추가로 적용해 AES 모델의 성능을 개선해 보고자 한다. 제시하는 전처리 방식은 에세이의 매 문장 앞부분에 Topic 정보를 라벨링(Topic Labelling 전처리) 하는 것이며, 데이터 증강 방식은 총 세 가지로 제시한다. 첫 번째는 문장 내 영향력이 적은 토큰을 MASK 토큰으로 교체하는 Random Masking 방식, 두 번째는 KcELECTRA의 MLM 기반 유의어 교체, 마지막은 T5의 paraphrase-generation 기능을 이용한 유의 문장 생성이다.

본 연구에서는 Topic Labelling 데이터 전처리와 더불어 다양한 데이터 증강 방식을 추가할 때 AES 모델의 성능이 개선됨을 보였다. Topic Labelling 전처리 방식만을 사용할 경우, 평가자와 모델 간의 점수 일치도를 Quadratic Weighted Kappa Score(QWK)를 기준으로 측정했을 때 베이스라인 대비 에세이의 모든 평가 항목에 대해 더욱 높은 일치도를 보임을 확인했다. 이와 더불어 다양한 데이터 증강 방식 (T5, MLM, Random Masking)과 데이터 전처리를 같이 적용했을 때 QWK를 비교, 분석했으며, 모든 증강 방식이 전처리만 했을 때 보다 평가 항목별 QWK의 평균과 총점의 QWK을 기준으로 더욱 높은 성능을 보이는 것을 확인했다. 또한 증강 방식에 따라 성능이 크게 오르는 평가 항목도 모두 다른 것으로 확인되었는데, 이는 각 데이터 증강 방식이 어떤 항목에 대해 모델 성능 개선에 도움을 줬는지 명확하게 보여주며, 다양한 증강 방식을 엮으면 특정 평가 항목의 일치도만 개선되는 것을 넘어 모든 평가 항목의 일치도 개선의 가능성이 있음을 보여준다.

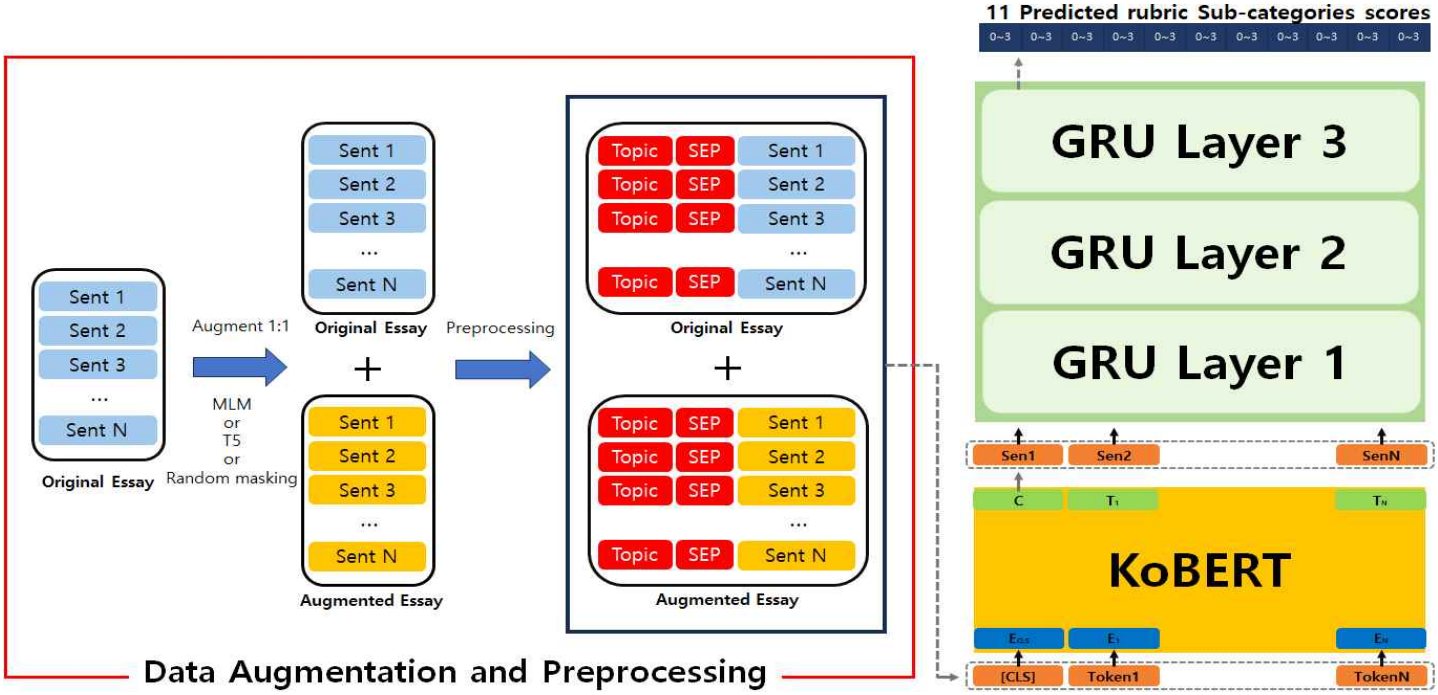

그림 1. 제안하는 데이터 증강 및 전처리를 활용한 AES 모델 개요

2. 관련 연구

2.1 AES 모델

국내에서는 사전 학습 언어 모델과 LSTM, GRU, RNN 등 다양한 모델을 활용한 AES 모델들이 제시된 바 있다. [1]에서는 KoBERT와 KoGPT2 와 같은 사전 학습된 한국어 언어 모델을 활용해 '직업', '행복'을 주제로 한 304개의 한국어 서술형 답안지를 4개의 점수 구간으로 자동 분류하는 모델을 제안했다. 결과적으로 KoBERT와 KoGPT2가 각각 '직업' 주제로 한 글에서 43.5%, 50.6%의 정답률을 보였고, '행복'을 주제로 한 글에서 각각 60.7%와 50.6%의 정답률을 보였다. [3]에서는 Argument Mining과 KLUE RoBERTa-base를 활용해 에세이의 논증 구조가 반영된 에세이의 표현 벡터를 만들고 평가 항목별로 표현을 학습하는 모델을 제안했다. AI-Hub의 '에세이 글 평가 데이터'를 활용했으며, 최종 제안 모델의 성능은 QWK 기준으로 0.543에서 0.627까지 향상되었다. [2]에서는 사전 학습된 언어 모델이 아닌 LSTM, GRU, RNN을 이용한 AES 모델의 성능을 비교한다. EBS <당신의 문해력>에서 대학생 530명을 대상으로 수집한 한국어 에세이 문항 답안지를 활용했으며, LSTM과 GRU 기반 모델이 가장 우수한 것으로 확인되었다. LSTM과 GRU는 성능에 큰 차이가 없었으나 학습 소요 시간 측면에서의 효율은 GRU가 더욱 우세한 것으로 나타났다. 해외 연구 중 [6]에서는 LSTM과 Transformer 기반 AES 모델의 성능을 Accuracy Score를 기준으로 비교했다. 실험 결과 전체적으로 Transformer 기반 모델들이 AES에서 더욱 우수한 성능을 보였다.

2.2 한국어 텍스트 데이터 증강

NLP 연구에 있어서 한국어는 영어에 비해 상대적으로 부족한 데이터로 인해 많은 어려움을 겪는다. [4]에서 적은 양의 한국어 상호 참조 데이터는 모델 성능 저하의 원인이 되었다. 이 문제는 데이터 증강을 통해 해결했으며 실험 결과 규칙 기반 데이터 증강보다 BERT의 MLM을 활용한 유의어 교체 데이터 증강이 더욱 높은 성능을 보였다. [5]에서는 AI-Hub에서 제공하는 웰니스 데이터셋을 EDA, BERT의 MLM, GPT-2, T5, 노이즈 추가 등으로 각각 증강했을 때 모델의 분류 성능을 비교했다. 결과적으로 원본 데이터와 증강 데이터의 비가 1:1일 때 가장 높은 성능을 보였으며, T5로 증강했을 때 가장 높은 성능을 보이는 것이 확인되었다.

2.3 AES 모델에서의 데이터 전처리

[6]에서는 AES 모델의 성능 향상을 위한 데이터 전처리 방식을 제안한다. ASAP 데이터를 활용했으며, 에세이의 10문장마다 Summary of Topic을 삽입했을 경우, 모델과 상관없이 항상 더욱 높은 성능을 보임을 확인하였다. [13]에서는 KoBERT기반 AES 모델 성능 개선을 위한 데이터 증강법을 소개한다. 에세이의 매 문장마다 Topic을 삽입하는 방식을 사용하며, 해당 논문에서는 데이터 증강으로 표시했으나 사실상 데이터 전처리로 보아야 한다.

3. 제안 방법

본 논문에서 제안하는 데이터 증강과 전처리를 활용한 AES 모델 개요는 그림 1과 같다. 원본 에세이로부터 증

강된 에세이를 1:1 비율로 생성하고, 이를 합쳐 데이터셋을 두 배로 만든다. 이와 더불어 에세이의 매 문장 앞에 Topic 정보를 덧붙여 전처리하고, AES 모델은 해당 데이터를 학습한다. 상기 과정을 MLM, T5, Random Masking 방식으로 각각 진행해 어떤 데이터 증강 방식이 가장 높은 성능을 보이는지 확인했다.

3.1 모델

모델은 AI-Hub의 '에세이 글 평가 데이터'의 베이스라인 모델을 사용한다 [7]. KoBERT 기반 구조로, 한국어 데이터에 대해 사전 학습된 'monologg/kobert'와 GRU를 사용했다. KoBERT에서 에세이 문장별 임베딩 벡터를 출력하고, GRU 모델에서 해당 임베딩 벡터를 입력으로 받아 11개 소분류별 평가 지표(rubric)에 대한 점수를 예측하도록 학습한다.

3.2 MLM 기반 데이터 증강

BERT는 특정 토큰을 Masking 한 후 해당 토큰을 맞추는 MLM(Masked Language Modeling) 태스크로 사전 학습을 진행한다 [4,9,10]. 해당 태스크를 이용해 Masking된 토큰을 교체하면 문장의 문맥을 고려한 유의어 교체가 구현된다 [4,9]. BERT뿐만 아니라 ELECTRA 모델도 Generator에서 MLM 태스크 수행이 가능하다 [10]. 본 연구에서는 학생들이 작성한 에세이라는 데이터 특성을 고려해서 뉴스 기사, 백과사전과 같이 잘 정제된 데이터로만 검증된 KoBERT 대신 신조어와 오탈자가 많은 데이터로도 사전 학습 성능이 검증된 KcELECTRA의 MLM으로 데이터 증강을 진행했다. 'KcELECTRA-base-v2022' 모델에 에세이의 훈련 데이터에 있는 약 46만 개의 문장을 RTD(Replaced Token Detection) 방식으로 사전 학습시킨 다음 해당 모델로 데이터 증강을 진행했다. MLM은 문서 단위가 아닌 문장 단위로 진행했으며, 문장마다 Masking 비율은 0.25로 고정해 증강했다. 표 1은 KcELECTRA 모델의 MLM을 적용해 증강된 에세이 문장의 예시를 보여준다.

표1. MLM 기반 증강 문장 예시

원문	증강 문장
특히 K-POP으로 인한 가수들이 널리널리 알려져 다들 한국에 대해 관심을 가지게 되고 한국을 좋아하는 것 같습니다.	최근 K-POP으로 인한 가수들이 더 널리 알려져 다들 한국에 대해 관심을 가지게 되고 한국을 사랑하는 것 같습니다.
우리나라에는 프랑스 마을, 스위스 마을 등 다양한 나라들의 모습이 담겨있는 마을들이 관광지로 생겨나고 있다.	우리나라에는 프랑스 마을, 스위스 마을 등 다양한 종류의 문화가 담겨 있는 건물들이 관광지처럼 생겨나고 있다.
나라도 라는 생각으로 열심히 절약하고 절약하는 방법을 주변 사람들에게 알려야 한다.	나라도 라는 마음으로 열심히 실천하고 절약하는 방법을 주변 사람들에게 알려야 한다.

3.3 T5 기반 데이터 증강

T5(Text-To-Text Transfer Transformer) 모델은 텍스트로 입력받아 텍스트를 출력하는 Transformer 기반의 모델이다[5,11]. T5 모델의 paraphrase task를 사용하면 입력 문장과 비슷한 문장을 만드는 것이 가능하므로 해당 태스크로 데이터 증강을 진행했다 [5]. 본 연구에서는 'psyche/KoT5-paraphrase-generation' 모델을 사용했으며, MLM 방식과 같이 문장 단위로 나눠 증강한 뒤 다시 합쳐 문서로 복구했다. 표 2는 T5 기반으로 증강한 에세이 문장의 예시를 보여준다.

표2. T5 기반 증강 문장 예시

원문	증강 문장
특히 K-POP으로 인한 가수들이 널리널리 알려져 다들 한국에 대해 관심을 가지게 되고 한국을 좋아하는 것 같습니다.	특히 K-POP이 널리 퍼지면서 많은 사람들이 한국에 관심을 갖고 한국을 좋아하는 것 같습니다
우리나라에는 프랑스 마을, 스위스 마을 등 다양한 나라들의 모습이 담겨있는 마을들이 관광지로 생겨나고 있다.	프랑스 마을, 스위스 마을 등 여러 나라들의 모습을 담은 마을들이 우리나라에도 관광지로 생겨나고 있다
나라도 라는 생각으로 열심히 절약하고 절약하는 방법을 주변 사람들에게 알려야한다.	주변 사람들에게도 나라도 라는 생각으로 절약하고 절약하는 방법을 알려야 한다.

3.4 Random Masking 기반 데이터 증강

노이즈 추가 기법은 전통적인 텍스트 데이터 증강 방식 중 하나이다[5]. 본 연구에서는 BART의 노이즈 추가 방식 중 Token Masking을 약간 변형한 Random Masking 증강 방식을 제안한다. BART의 Token Masking은 무작위 토큰을 선정해 [MASK] 토큰으로 바꾸는 방식이다 [12]. 하지만 완전히 무작위로 토큰을 교체하면 AES 점수에 영향을 크게 주는 고급 어휘가 가려질 위험이 크다. 따라서 mecab 형태소 분석기를 활용해, 최대한 고급 어휘가 존재할 가능성이 낮은 형태소(각종 기호나 숫자, 감탄사, 부사)만 선택해 Masking 하도록 구현했다. 표 3은 Random Masking 기반으로 증강한 문장의 예시이다.

표3. Random Masking 기반 증강 문장 예시

원문	증강 문장
특히 K-POP으로 인한 가수들이 널리널리 알려져 다들 한국에 대해 관심을 가지게 되고 한국을 좋아하는 것 같습니다.	[MASK] K [MASK] POP으로 인한 가수들이 [MASK] [MASK] 알려져 [MASK]들 한국에 대해 관심을 가지게 되고 한국을 좋아하는 것 같습니다.
우리나라에는 프랑스 마을, 스위스 마을 등 다양한 나라들의 모습이 담겨있는 마을들이 관광지로 생겨나고 있다.	우리나라에는 프랑스 마을 [MASK] 스위스 마을 등 다양한 나라들의 모습이 담겨 있는 마을들이 관광지로 생겨나고 있다.
나라도 라는 생각으로 열심히 절약하고 절약하는 방법을 주변 사람들에게 알려야한다.	나라도 라는 생각으로 [MASK] 절약하고 절약하는 방법을 주변 사람들에게 알려야 한다.

표 5. 증강 방식에 따른 평가 지표별 모델의 예측점수와 평가자의 실제 점수간 일치도(QWK)

평가 지표		baseline	only topic labelling	topic labelling+mlm	topic labelling+t5	topic labelling+ random masking
표현	문법	0.2622596	0.26930579	0.27715189	0.26810265	**0.27733506**
	단어	0.31521567	0.33054024	0.33295944	**0.3407643**	0.32937973
	문장표현의 적절성	0.94882755	0.96459496	**0.96494575**	0.96397149	0.96377661
구성	문단 내 구조의 적절성	0.30914716	0.3140109	0.31557667	**0.32264054**	0.32144692
	문단 간 구조의 적절성	0.95852761	0.97029699	**0.9711687**	0.97029894	0.97107676
	구조의 일관성	0.93373588	0.94756419	0.94766025	0.94720578	**0.94791928**
	분량	0.45040848	0.46493739	**0.4759848**	0.4624171	0.47497088
내용	주제의 명료성	0.25975582	0.27844139	0.27842683	**0.2898706**	0.28078415
	참신성	0.16967992	0.18105092	**0.19837977**	0.17461686	0.18336668
	프롬프트 독해력	0.95475131	0.96046529	0.96039533	0.96044773	**0.96092315**
	서술력	0.34252284	0.35016203	0.34403149	**0.35335624**	0.34666475
평균		0.5368029	0.5483064	**0.55151645**	0.5503357	0.5506949

3.5 Topic을 활용한 데이터 전처리

표4. Topic을 활용한 데이터 전처리 예시

원문	전처리 문장
저는 학습시간이 5시간 여가시간이 2시간을 쓰고 있습니다.	학습과 여가활동 [SEP] 저는 학습시간이 5시간 여가시간이 2시간을 쓰고 있습니다.
기억력이 좋지 않다는 것이 나의 약점이다.	본인의 성격 [SEP] 기억력이 좋지 않다는 것이 나의 약점이다.

본 연구에서 제안하는 전처리 방식은 표4와 같이 에세이의 매 문장 앞에 해당 에세이의 Topic과 SEP 토큰을 라벨링 하는 것이다. 이는 [13]에서 에세이의 매 문장마다 Topic을 삽입하는 방식을 참고하였다. 실제로 에세이의 점수는 Topic에 따라서도 달라지기 때문에, 만약 Topic에 대한 정보를 모델이 추가로 학습하게 된다면 더욱 높은 성능을 기대해 볼 수 있다 [6]. Topic은 '에세이 글 평가 데이터'에서 에세이 원문과 함께 제공되기 때문에 이를 그대로 활용해 전처리를 진행했다.

4. 실험

4.1 데이터 구성

AI-Hub의 '에세이 글 평가 데이터'는 초등 4학년 ~ 고등 3학년 학생들을 대상으로 수집된 에세이 글 데이터이다 [7]. 총 50,400개의 라벨링된 에세이로 구성되어 있으며, 대분류 Type에 따라 논술형(51.5%), 수필형(48.5%)으로 나뉘며 여기에 세부 Type에 따라 논술형-(주장, 찬성/반대, 대안 제시), 수필형-(설명글,글짓기)으로 나뉜다. 더욱 세분화하면 Topic(prompt)에 따라 50개로 분류할 수 있다. 본 연구에서는 50,400개의 에세이를 주제별 8:2 비율로 분리해, 총 39,591개의 데이터를 훈련 데이터로, 4,955개의 데이터를 평가 데이터로 활용했으며, 나머지는 검증 데이터로 활용했다. 에세이의 label은 대분류로 표현, 구성, 내용으로 나뉘고 소분류로는 11개로 나뉜다. 소분류별 평가지표(rubric)점수는 항상 0~3점 사이의 정수이다. 또한 11개의 평가 지표는 각 rubric 가중치를 통해 합쳐져 에세이의 총점으로 계산될 수 있는데, 이때 rubric 가중치는 에세이의 주제별로 다르게 책정된다. 본 연구에서는 소분류별 평가 지표에서 모델의 예측점수와 평가자의 실제 점수 간의 일치도를 QWK로 측정했고, 총점에 대한 모델과 평가자 간의 일치도 또한 측정했다.

4.2 실험 환경

데이터 증강 방식과 전처리 여부에 따라 총 5가지 Case에 대해 실험을 진행했다. 5가지 실험이 진행됨에 따라 모델과 실험 환경은 항상 일정하게 유지했다. 모든 실험에는 Nvidia RTX 3090Ti GPU 1개가 사용되었으며, 모델은 항상 사전 학습된 'monologg/kobert'와 GRU 레이어 3개를 사용했다. 하이퍼 파라미터는 항상 epoch 30, batch size 64, max sentence length 70, max sequence length 128로 고정했다. 데이터 형태에 따른 5가지 Case는 각각 baseline = 베이스라인, only topic labelling = 데이터 증강 없이 전처리, topic labelling + mlm = (전처리 + MLM 증강), topic labelling + t5 = (전처리 + T5 증강), topic labelling + random masking = (전처리 + Random Masking 증강)이다. 실험을 진행하면서 Topic Labelling 전처리 방식이 성능 향상 효과가 있는지 검사했고, 여기에 추가로 원본 데이터뿐만 아니라 증강된 데이터로 학습시킬 때 성능 향상 여부와 증강 방식에 따른 성능 비교도 진행했다. 실험의 결과는 모두 각각 5회씩 진행해, 평균으로 계산해 낸 값이다.

4.3 실험 결과

표 5를 통해 알 수 있듯이 실험 결과 Topic Labelling으로 데이터를 전처리할 경우, 모든 평가 지표에서 베이스라인 대비 성능 향상이 있었다. 여기에 데이터 증강을 추가로 적용할 때 특정 평가 항목에서 전처리 된 원본 데이터로만 학습할 때보다 더욱 높은 일치도를 보였다. 각 평가 지표별 가장 높은 일치도를 보이는 방식은 모두 데이터 증강과 전처리를 함께 진행했을 때 나타났다. 문법, 구조의 일관성, 프롬프트 독해력 항목에서 가장 높

은 일치도를 보이는 방식은 (전처리 + Random Masking 증강) 방식이었고, 단어, 문단 내 구조의 적절성, 주제의 명료성, 서술력 항목에서 가장 높은 일치도를 보이는 방식은 (전처리 + T5 증강) 방식이었다. 문장표현의 적절성과 문단 간 구조의 적절성, 분량, 참신성에서 가장 높은 성능을 보이는 방식은 (전처리 + MLM 증강) 방식이었다. 11개의 일치도의 평균으로 성능 평가를 해봤을 때 전처리만 할 때 베이스라인 대비 QWK가 약 0.0115035 증가했으며 여기에 증강을 추가했을 때 모두 더욱 상승했다. 증강 방식 중 가장 높은 성능을 보인 것은 (전처리 + MLM 증강) 방식이었다.

해당 11개의 소분류별 평가 지표와 각 평가 지표별 가중치를 통해 에세이의 총점을 계산할 수 있다. 계산된 총점들을 바탕으로 모델의 예측점수와 평가자의 실제 점수 간의 일치도를 비교한 결과는 표 6과 같다. 이전 실험 결과와 마찬가지로 QWK, Pearson 상관계수 모두 Topic Labelling 전처리를 하면 베이스라인 대비 높은 성능을 보였고, 여기에 데이터 증강을 추가하면 추가적인 성능 향상을 보인다. QWK는 평가지표별 결과와 비슷하게 (전처리 + MLM 증강) 방식이 가장 높았고, Pearson 상관계수는 (전처리 + T5 증강) 방식이 가장 높았다. 결과적으로 제안하는 방식이 베이스라인 대비 QWK와 Pearson 상관계수가 0.0269323, 0.0190172 만큼 개선되었다.

표6. 총점 기준 각 Case별 QWK와 Pearson 상관계수

모델	QWK	Pearson 상관계수
baseline	0.4110809	0.5065425
only topic labelling	0.4255103	0.5202876
topic labelling+mlm	0.4380132	0.5254088
topic labelling+t5	0.433079	0.5255597
topic labelling+ random masking	0.4338197	0.5242532

5. 결론

본 논문에서는 '에세이 글 평가 데이터'의 KoBERT 기반 AES 모델 성능 향상을 위한 Topic Labelling 데이터 전처리와 MLM, T5, Random Masking 데이터 증강 기법을 제안한다. 에세이 데이터에 Topic Labelling 전처리를 할 때 모든 평가 항목과 총점에 대해 더욱 높은 일치도를 보였고, 여기에 데이터 증강법을 추가하면 데이터 증강 방식에 따라 특정 항목에 대해 추가적인 일치도 향상을 보였다. (전처리 + 데이터 증강)을 할 때 평가항목별 QWK의 평균과 총점의 QWK, Pearson 상관계수에서 모든 증강 방식이 더욱 높은 성능을 보였고, (전처리 + MLM 증강) 방식으로 AES 모델을 학습하면 가장 높은 성능을 보였다. 11개의 평가항목별 QWK의 평균은 0.5368029에서 0.55151645(+0.01471355)로 증가 했고, 총점에 대한 QWK, Pearson 상관 계수는 각각 0.4110809에서 0.4380132(+0.0269323)로, 0.5065425에서 0.5254088(+0.0188663)로 증가했다.

또한 실험 결과, 데이터 증강을 추가하면 특정 평가 항목에 대해서만 일치도가 상승하는 것을 보였고, 증강 방식에 따라서 상승하는 평가 항목이 미묘하게 달랐다. 따라서 향후 연구에서는 더 다양한 데이터 증강 방식을 엮어 특정 항목에 대해서만 성능 향상 효과가 있는 증강 방식이 아닌 모든 평가 항목에 대해 성능 향상 효과가 있는 증강 방식을 구축해볼 예정이다.

감사의 글

이 논문은 2023년도 인하대학교의 지원, 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.RS-2022-00155915, 인공지능융합혁신인재양성 (인하대학교)) 및 2023년 대한민국 교육부와 한국연구재단의 지원(NRF-2023S1A5A2A21085373)을 받아 수행된 연구임

참고문헌

[1] 조희련, 이유미, 임현열, 차준우, 이찬규, "딥러닝 기반 언어모델을 이용한 한국어 학습자 쓰기 평가의 자동 점수 구간 분류 - KoBERT와 KoGPT2를 중심으로 -" 한국 언어 문화학, 2021, vol.18, no.1, pp. 217-241(25 pages) 국제 한국 언어 문하 학회

[2] 박강윤, 이용상, "한국어 에세이 문항 자동 채점을 위한 딥러닝 알고리듬 탐색" vol.35, no3, pp.465-488(24 pages), 2022

[3] 이예진, 장영진, 김태일, 최성원, 김학수, "논증 구조 정보를 반영한 심층 신경망 기반 에세이 자동 평가 파이프라인 모델",제 34회 한글 및 한국어 정보처리 학술대회 논문집, pp.354-359, 2022.10.19.

[4] 김기훈, 이창기, 류지희, 임준호, "한국어 상호 참조 해결을 위한 BERT 기반 데이터 증강 기법", 제 32회 한글 및 한국어 정보처리 학술대회 논문집, pp.249-253, 2020

[5] 서가은, 오하영, "한국어 데이터를 활용한 data augmentation", 한국 정보통신학회 논문지 제 27권 제4호, pp.483-488 (6 pages), 2023.4

[6] GUPTA, Kshitij, "Data Augmentation for Automated Essay Scoring using Transformer Models." In: 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, p.853-857, 2023

[7] 데이터 분야 - AI 데이터 찾기 - AI-Hub, 에세이 글 평가 데이터, 갱신년월: 2022-10, 구축년도: 2021 https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=545

[8] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. "On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation". In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3416-3425, Seattle, United States. Association for Computational

Linguistics, 2022
[9] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data Augmentation using Pre-trained Transformer Models. In Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems, pages 18-26, Suzhou, China. Association for Computational Linguistics, 2020
[10] Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning, “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”,ICLR 2020,arXiv:2003.10555, 2020
[11] Raffel Colin, Shazeer Noam, Roberts Adam, Lee Katherine, Narang Sharan, Matena Michael, Zhou Yanqi, Li Wei, and Liu Peter J, “Exploring the limits of transfer learning with a unified text-to-text transformer”. Journal of Machine Learning Research 21, 140 (2020), 5485-5551, 2020
[12] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”, In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871-7880, Online. Association for Computational Linguistics, 2020
[13] 고강희 and 김도국, "데이터 증강을 이용한 KoBERT 기반 에세이 자동 평가 성능 향상," in 한국정보과학회 학술발표논문집, 개최지, pp.1764-1766, 2023.