영어 교육을 위한 거대 언어 모델 활용 말뭉치 확장 프레임워크

정진우 , *정상근
충남대학교 컴퓨터융합학부
{wlsdn2749, hugmanskj}@gmail.com

Data Augmentation using Large Language Model for English Education

Jinwoo Jung, Sangkeun Jung†
Department of Computer Science and Engineering, Chungnam National University

요 약

최근 ChatGPT와 같은 사전학습 생성모델은 자연어 이해 (natural language understanding)에서 좋은 성능을 보이고 있다. 또한 코드 작업을 도와주고 대학수학능력시험, 중고등학교 수준의 문제를 풀거나 도와주는 다양한 분야에서 활용되고 있다. 본 논문은 사전학습 생성모델을 이용하여 영어 교육을 위해 말뭉치를 확장하는 프레임 워크를 제시한다. 이를 위해 ChatGPT를 사용해 말뭉치를 확장 한 후 의미 유사도, 상황 유사도, 문장 교육 난이도를 사용해 생성된 문장의 교육적 효과를 검증한다.

주제어: 데이터 증강, 평가지표, 자연어 생성

1. 서론

ChatGPT의 출시 이후로 대중들에게 생성형 언어모델이 많이 알려지게 되면서 일반 사용자들은 간단한 프롬프트를 생성형 언어모델에 입력으로 넣어주는 것으로 결과를 얻어낼 수 있게 되었다.

말뭉치 생성 같은 경우에는 기존에 확률 기반 언어모델 [1]부터 시작해서 인터넷의 웹이나 문서들을 크롤링하여 얻은 데이터를 기반으로 생성하거나 [2]사진 데이터에서 문자 인식 (Optical Character Recognition)이나 음성 인식 (Automatic Speech Recognition) [3, 4] 같은 콘텐츠에서도 데이터를 생성하여 말뭉치 생성을 진행하고 있고 현재는 기계학습 기반의 시계열 회귀 [5], Transformer 구조를 사용하여 다음 단어를 예측하는 방식으로 말뭉치 생성 방향이 진행되고 있다 [6]. 그런데 이러한 학습방법에는 모두 대량의 데이터가 필수적으로 필요하다는 점에 있어서 대량의 데이터를 획득할 수 없는 일반 사용자가 교육 등의 이유로 말뭉치 확장이 필요하다고 하더라도 어려운 점이 있다.

이 연구는 일반 사용자를 대상으로 이미지나 말뭉치 같은 콘텐츠가 필요할 경우에 생성형 언어 모델을 사용하여 손쉽게 정확도 있고 신뢰도 높은 데이터를 생성 사용할 수 있게 프레임워크를 제작하려고 한다. 따라서 본 연구는 82개의 상황을 가지는 독자적인 데이터 셋을 구축하고. 이를 ChatGPT 프롬프트를 통해 각 문장당 10개씩 확장하였다. 확장된 문장에 대해서 의미 유사도, 상황 유사도, 문장 교육 난이도 총 3가지의 지표를 구하고 이를 가중합하여 가지는 점수 S를 낸다. 이를 통해 생성 문장 D_i^G에서 점수를 가장 높게 가지는 D_i^*를 추출해낸다.

2. 관련연구

2.1 데이터 증강

데이터 증강 (Data Augmentation)은 여러 변형을 통해 데이터의 인공적인 생성 야기하여 여러 다운스트림 테스크 (Downstream task)에서 모델의 성능을 높히기 위해 사용되고 있다. 컴퓨터 비전 (Computer Vision) 에서는 기하학적 변환, 색상 공간 증강, 커널 필터, 이미지 혼합, 특성 공간 증강 등의 방법을 통해 증강을 시도하고 있고 [7] 자연어 처리에서는 규칙 기반, 예제 보간, 모델 기반 방법을 통해 증강을 시도 하고 있다 [8].

규칙 기반 텍스트 데이터 증강 방법 중 Easy Data Augmentation (EDA) [9]는 동의어 대체, 삽입, 단어 교환, 제거 같은 방법을 사용하여 데이터를 증강하는 방법을 제안했다. Seq2seq 기반 가장 유명한 방법중 하나인 Back Translation [10] 방법은 문장을 다른 언어로 변경하고 다시 원래의 언어로 변경하는 방법으로 모델 기반의 데이터 증강 방법을 제안했고, masked language models 기반의 유명한 방법 중 하나인 BERT [6]는 ⟨**mask**⟩ token을 사용하여 ⟨**mask**⟩ token을 다른 단어로 대체하는 방법을 사용했다.

최근 연구에서는 거대 언어 모델인 ChatGPT를 이용하여 분류와 같은 다운스트림 테스크에서 높은 성능을 보여주는 것이 실험을 통해 증명 되었고, ChatGPT를 통해 생성된 문장으로 Finetuning을 하여 분류 문제의 성능을 더욱 높힌 사례도 존재한다 [11, 12]. 이외에도 ChatGPT를 사용한 교육적 데이터 증강 또한 시도하고 있다 [13].

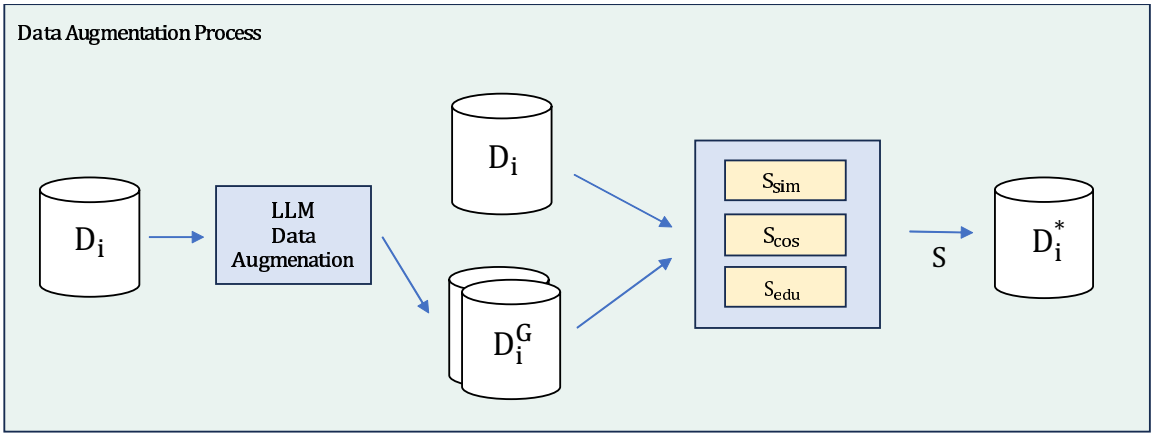

그림 1. 말뭉치 확장 프로세스

2.2 거대 언어 모델

Transformer [14]의 등장 이후 Transformer 기반 사전학습 (Pre-trained language) 모델 중 대표적인 BERT와 GPT가 있다. Transformer에서 encoder만 사용한 구조인 BERT [6]와 Decoder만 사용한 구조인 GPT는 [15] 모두 데이터의 양과 모델 크기에 따라 성능이 상승한다는 결과가 나왔고, 약 100만 개의 파라피터를 사용하였다 [16]. 기존 GAN, LSTM에 비해 여러가지 다운 스트림 (Down-Stream) 테스크에서 좋은 성능을 보여주었다 [6, 15]. 최근에는 자연어 처리분야에서는 모델의 크기를 키워 성능을 높히는 연구와 모델의 크기를 제한하면서 성능도 좋게 나오는 연구를 진행중에 있다. GPT-2는 파라미터의 개수를 1.5B로 늘렸고[17] GPT-3는 1750B로 늘렸다 [18]. 이는 다양한 다운 스트림 테스크에서 보다 좋은 성능을 보여주었다. 반면, 모델에 Finetuning을 어떻게 하느냐에 따라 모델의 크기가 작음에도 최적화되어 좋은 성능을 보여주는 사례도 있었는데, Chinchilla 5-shot (70B)의 성능이 Gopher 5-shot(280B)보다 좋은 성능을 보여주었다 [19]. 또한 Meta AI의 LLama-13B는 175B의 GPT-3보다 좋은 성능을 보여주었고, LLama-65B는 Chinchilla (70B), Gopher (280B) 보다도 좋은 성능을 보여주었다. 한편 거대 언어 모델로 유사한 데이터를 증강하는 연구가 있다 [8]. OpenAI의 GPT-3 [18] 를 기반으로한 ChatGPT는 생성된 결과를 사람에 의해 평가하는 것을 보상으로 Reinforcement Learning from Human Feedback (RLHF) 학습을 진행하였다. ChatGPT와 GPT-4는 이 Reward-model을 통해 여러 분야에서 사람과 동등하거나 그 이상의 Task 수행 능력을 보여주었다 [20]. 따라서 본 연구에서는 ChatGPT가 문장 생성 과 평가를 사람 수준으로 잘 할 것으로 생각하여 연구를 진행한다.

3. 말뭉치 확장 및 평가 프로세스

본 논문에서는 확장된 문장 데이터의 교육적 효과를 검증하기 위해 상황별로 구분된 데이터셋을 사용한다, 그리고 구축한 데이터 셋 D_i을 바탕으로 외부 거대 언어 모델인 ChatGPT를 통해 10배 증강 시켜 D_i^G로 증강한다. 증강된 D_i^G의 예는 표 1과 같다. 다음으로 D_i^G의 교육적 효과를 검증하기 위해 D_i 와 D_i^G 사이의 의미 유사도 S_{sim}, 상황 유사도 S_{con}, 문장 교육 난이도 S_{edu} 와 이를 종합한 지표 S를 본 논문에서 제시하고, 이를 통해 가장 교육적 효과가 높은 증강된 데이터 하나를 담고 있는 데이터셋 D_i^*를 산출한다. 전체적인 프로세스는 그림 1과 같다.

표 1. ChatGPT로 부터 생성된 샘플 데이터의 예

생성 대상 문장
There's no room for error with me.
생성된 문장
1. I leave no room for mistakes.
2. There is zero tolerance for error when it comes to me.
3. I do not allow any margin for error.
4. I expect perfection without any room for mistakes.
5. There is no space for error when dealing with me.

3.1 데이터 셋

실험에 사용될 데이터 셋은 평가에 사용할 지표 중 하나인 상황 별로 구분된 데이터가 필요하고, 또한 교육에 사용될 수 있는 데이터 셋이 필요하다. 이를 위해 영어회화에 사용되는 다양한 표현들을 바탕으로 구축한 735개의 데이터를 82개의 상황을 가지도록 구분하였다. 각 문장은 상황별로 적게는 2개 많게는 10개 사용되었다. 상황별 데이터 목록은 표 2와 같다.

표 2. 데이터셋의 예

순서	상황	예시 문장	개수
1.	Meetings	How nice to see you!	10
2.	Introduction	Let me introduce myself.	10
3.	Advice	I need some advice	10
...	...	...	
55.	Appointment	I'm a man of my word.	10
56.	Past Story	It's water under the bridge.	6
...	...	...	
82.	Plan	What is your objective?	10
총합	82	—	735

3.2 거대 언어 모델을 사용한 말뭉치 확장 방법

거대 언어 모델에서 ChatGPT는 강화학습을 사용해 다른 거대 언어 모델에 비해 데이터 생성 능력 결과에 인간과 유사한 결과를 낸다 [20]. 그러므로 ChatGPT를 데이터 증강에 사용한다. 말뭉치 확장하기 위한 프롬프트는 다음과 표 3과 같다.

표 3. 말뭉치 확장 프롬프트

프롬프트
Generate 10 semantically similar sentences separated by three backticks (`) ```⟨**source**⟩```

3.3 평가 지표의 구성

교육적 효과를 검증하기 위해 지표로 의미 유사도 S_{sim} 와 상황 유사도 S_{con} 그리고 문장 교육 난이도 S_{edu} 를 각각의 합이 1인 α, β, γ로 가중합하여 최종 평가지표 S 로 사용한다. 최종 평가지표를 이루는 각각의 지표는 이어지는 섹션에서 자세히 설명한다. 최종 평가지표 S의 수식은 다음과 같다.

$$S(D_i, D_{i,k}^G) = \alpha S_{sim} + \beta S_{con} + \gamma S_{edu} \quad (1)$$

3.3.1 의미 유사도

의미 유사도를 평가하기 위해서 생성된 데이터와 기존 데이터와의 임베딩 유사도 (Embedding Similarity) 를 구하는 방법을 적용한다. 임베딩 유사도를 구하는 방법중 가장 주로 사용되는 평가 척도에는 Euclidean 거리, Cosine 유사도 그리고 Dot product 유사도가 있는데, 자연어 처리 분야에서 일반적으로 사용하는 지표이면서 Euclidean 거리보다 일반적으로 좋은 성능을 가진 Cosine 유사도 [21]를 사용하여 두 문장간의 거리를 구해 유사도를 측정한다.. 이 값은 두 문장 벡터 사이의 거리가 가까울수록 1의 값을 가진다. 본 논문에서는 D_i의 벡터와 $D_{i,k}^G$의 벡터를 연산에 사용하여 Cosine 유사도 $cos(\theta)$를 구한다.

또한 이 논문에서는 BERTScore의 F1 Score 점수를 추가로 사용하는데, 이는 Contextual embedding model인 BERT를 사용하여 두 문장 간의 각 토큰마다 Cosine 유사도를 측정하고 Inverse Document Frequency (IDF) 점수를 통해 가중치를 부여하여 Precision과 Recall 그리고 F1 Score를 구할 수 있다 [22]. 기존 문장 D_i의 토큰 x_i와 생성된 문장 $D_{i,k}^G$의 토큰 $\hat{x}_j$ 사이의 Cosine 유사도를 구하고 이를 통해 두 문장간의 Recall, Precision, F1 Score를 구할 수 있다.

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x_j} \in \hat{x}} x_i^T \hat{x}_j \quad (2)$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x_j} \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j \quad (3)$$

$$F_{BERT} = 2 * \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \quad (4)$$

Cosine 유사도와 BERTScore의 F1 Score를 동일한 비중으로 가중합하여 의미 유사도를 구한다. 수식은 다음과 같다.

$$S_{sim}(D_i, D_{i,k}^G) = 0.5 * cos(\theta) + 0.5 * F_{BERT} \quad (5)$$

3.3.2 상황 유사도

상황 유사도를 평가하기 위해서 ChatGPT의 프롬프트를 사용한다. ChatGPT는 여러 자연어 처리 대부분의 작업에 좋은 성능을 낸다 [11]. ChatGPT가 문장간의 같은 상황인지를 잘 분류하고 판단할 것이다. 그러므로 ChatGPT에 보낼 프롬프트를 개발하여 상황 유사도를 평가하였다. 사전 구축한 데이터셋에 대하여 표 4와 같은 프롬프트를 작성하였다. 두 문장이 같은 상황일 수록 1에 가까운 점수를 받게 되고, 다른 상황일수록 0에 가까운 점수를 받게 된다. 이를 $S_{con}(D_i, D_{i,k}^G)$로 정의한다.

표 4. 상황 유사도 측정 프롬프트

프롬프트
Source: ⟨**SOURCE**⟩ Generated: ⟨**GENERATED**⟩ Score from 0 to 1 the extent to which ⟨**SOURCE**⟩ and ⟨**GENERATED**⟩ are in the same context with ⟨**SOURCE**⟩ context : ⟨**CONTEXT**⟩

3.3.3 문장 교육 난이도

문장간의 교육적 효과를 검증하기 위해서 문장간의 어휘 수준 지표를 사용한다. 어휘를 기반으로 한 학습은 단순히 단어나 구를 암기하여 학습하는 것 뿐만 아니라, 어휘를 적절하게 사용하는 능력을 길러준다. 그러므로 적절한 어휘를 바탕으로 학습하는것이 매우 중요하다 [23]. 하지만 문장간의 어휘 수준 차이가 너무 높거나 너무 낮으면 생성된 문장으로부터의 학습효과를 기대하기가 어렵다. 이 논문에서는 생성된 문장의 어휘수준 차이를 측정하기 위하여 ChatGPT의 프롬프트와 Readability Formula인 Flesch reading-ease score (FRES) [24]와 문장의 길이의 차를 사용한다. ChatGPT의 프롬프트를 사용하는 방법으로는 상황유사도 프롬프트와 유사하게 사전 구축한 모든 데이터 셋에 대하여 표 5와 같은 프롬프트를 작성하였다 두 문장 사이에 같은 어휘 수준을 가질 수록 1에 가까운 점수를 받게 되고, 어휘 수준의 차이가 클경우 0에 가까운 점수를 받게 된다. 이 점수를 S_{GPT_voca}로 정의한다

표 5. 문장 교육 난이도 측정 프롬프트

프롬프트
Source: ⟨**SOURCE**⟩ Generated: ⟨**GENERATED**⟩ Score from 0 to 1 the extent to which ⟨**SOURCE**⟩ and ⟨**GENERATED**⟩ are in the same vocabulary level

FRES는 어떤 문장이나 문서가 다음 수식에 의해서 계산된 점수가 높을 수록 읽기 어려운 것을 나타내는 점수이다. 점수가 나타내는 상세한 의미는 표 6과 같다 [25]. 수식은 (6)과 같다.

$$S_{FRES} = 206.835 - 1.015(\frac{total\ words}{total\ sentecnes}) \\ -84.6(\frac{total\ syllables}{total\ words}) \quad (6)$$

표 6. FRES chart [25]

점수	학력 (US)	설명
90 to 100	5th grade	Very Easy
80 to 90	6th grade	Easy
70 to 80	7th grade	Fairly Easy
60 to 70	8th & 9th grade	Plain English
50 to 60	high school	Fairly Difficult
30 to 50	college	Difficult
0 to 30	college graduate	Very Difficult

본 논문에서는 기존 문장과 생성된 문장과의 어휘 수준 차이를 측정하기 위해 두 FRES의 차이의 절댓값을 점수의 최대값인 121.22로 나눈것을 1에서 빼주어 0에 가까울 수록 생성된 문장이 기존 문장에 비해 너무 높거나 낮은 어휘 수준을 가지고 있고 1에 가까울수록 비슷한 어휘 수준을 가진다고 판단할 수 있다.

$$S_{FRES\_D}(D_i, D^G_{i,k}) = 1 - \frac{|S_{FRES}(D_i) - S_{FRES}(D^G_{i,k})|}{121.22} \quad (7)$$

새로 생성된 문장의 길이가 원래 문장보다 너무 길거나 너무 짧으면 학습에 도움이 되지 않을 것이다. 구두점을 제거한 문장의 길이의 차가 커질수록 0에 가까운 값을 가지게끔 하여 계산하였다. 이 점수를 S_{wc}로 정의한다. 이렇게 점수 낸 3개를 가중합하여 문장 교육 난이도 점수를 구한다.

$$S_{edu}(D_i, D^G_{i,k}) = 0.33 * S_{GPT\_voca} + 0.33 * S_{FRES\_D} \\ +0.33 * S_{wc} \quad (8)$$

4. 실험 및 실험 결과

4.1 실험 환경

말뭉치 확장에 사용한 거대 언어 모델은 ChatGPT를 사용하였고 모델은 모두 동일하게 gpt-3.5-turbo를 사용하였다. 하이퍼파라미터 top_p는 0.8, temperature는 0.7을 사용하였다. 각 평가지표에 동일한 가중치를 두어 객관적 지표를 얻기 위해서 수식 (1)의 하이퍼파라미터 α, β, γ는 모두 0.33으로 설정하여 실험하였다. BERTscore는 사전학습된 BERT모델인 RoBERTa-large 모델을 사용하여 측정하였다 [26].

4.2 실험 결과

본 논문에서는 실험을 통해 제안하는 평가지표 S가 영어 교육에 있어서 효과적인 데이터를 생성함을 검증하려 한다. 또한 S를 구성하는 각각의 지표인 S_{sim}, S_{con}, S_{edu}에 대해서도 제시하여 제안하는 접근 방식의 타당성을 확인한다.

표 7의 D_{34}는 무작위로 1개 샘플링한 기존의 문장 데이터이다. D_{34}를 거대 언어모델을 사용하여 데이터를 증강한 후 생성된 문장 D^G_{34} 와 D_{34} 사이에 계산 된 각각의 지표마다 상위 3개의 점수를 가지는 문장을 추출하고 점수를 표기한 것이다.

표 7의 실험 결과에서, D_{34}에서 생성된 문장 D^G_{34}이 각각의 지표 안에서 의미 유사성, 상황 유사성, 교육 난이도 측면에서 모두 높은 점수를 가지고 잘 생성해낸다고 볼 수 있다. 특히 S의 가장 높은 점수 0.851을 가지는 문장은 지표 S_{sim}, S_{con}, S_{edu}에서도 상위 3개안에 드는 것을 확인 할 수 있다 점수 또한 높음을 보여준다.

그림 2는 실험에 사용된 모든 데이터셋을 다중 거대언어 모델을 통해 증강한 후 S_{sim}, S_{con}, S_{edu} 그리고 S를 구해 평균낸 것 **평균**와 기준이 되는 문장 하나를 샘플로 선택한 후 그

샘플에 해당되지 않는 다른 상황의 데이터를 10개 추출 한 후 그것을 평균 낸 **표본**이 있다.

그림 2의 **표본**은 **평균**에 비해 전반적으로 더 낮은 점수를 보여준다. 특히 상황 유사도 점수와 의미 유사도 점수에서는 낮은 점수를 보여주었고, 문장 교육 난이도 점수는 상대적으로 높은 점수를 가지는 것을 보여준다. 상황 유사도는 상황이 다른 문장이기 때문에 낮은 점수를 가지는 것으로 추측할 수 있고. 의미 유사도는 **표본**에서 기존 데이터와 선택된 데이터 사이에 유사하거나 같은 단어가 없기 때문에 낮은 점수를 가지는 것으로 추측할 수 있다. 교육 난이도 점수는 문장 사이의 어휘수준과 문서의 난이도와 문서의 길이에 의해 정해지기 때문에 비교적 높은 점수를 획득한 것으로 볼 수 있다.

표 7. 생성된 문장 상위 3개 추출한 예

	문장	점수
D_{34}	I'll see you again tomorrow.	-
S_{sim}	Tomorrow, we will see each other again.	0.83
	I will meet you again tomorrow.	0.78
	I will be seeing you again tomorrow.	0.78
S_{con}	**I will meet you again tomorrow.**	0.9
	I will be seeing you again tomorrow.	0.9
	Tomorrow, we will see each other again.	0.8
S_{edu}	We will meet again tomorrow.	1.0
	I will meet you again tomorrow.	0.9
	We can catch up again tomorrow.	0.9
S	**I will meet you again tomorrow.**	0.851
	We will meet again tomorrow.	0.845
	I will be seeing you again tomorrow.	0.835

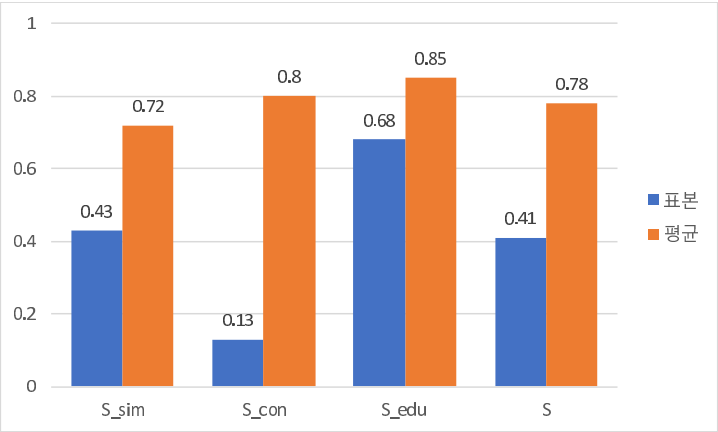

그림 2. 생성된 모든 데이터셋에 대한 각 지표의 평균과 무작위 데이터를 사용한 경우의 평균 비교

5. 결론

본 논문은 ChatGPT와 같은 거대 언어 모델 기반의 자연어 처리 중 유사 문장 생성 결과를 평가할 수 있는 새로운 지표인 S 를 제안한다. S는 문장 기반 학습의 의미, 상황, 교육을 모두 고려한 종합 지표이다. 생성된 문장 중 S가 가장 높은 1개를 S^*로 설정함으로써 생성된 문장의 퀄리티를 효과적으로 높힐 수 있었다. 하지만 여러 사전학습 기반 거대 언어 모델을 실험해보지 않았기에, 향후 연구에는 많은 모델에 S를 적용시켜 모델에 따른 생성 퀄리티를 확인하여, 보다 높은 생성 결과를 낼 수 있도록 방안을 연구 할 것이다.

감사의 글

이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2022R1F1A1071047)

참고문헌

[1] J.-K. Kim, C.-H. Kim, M.-A. Cheon, H.-M. Park, H. Yoon, Y. Nam-Goong, M.-S. Choi, and J.-H. Kim, "Generating Korean NER Corpus using Hidden Markov Model," *Annual Conference on Human and Language Technology*, pp. 357–361, 2019.

[2] J. Pascual Espada and I. Cid Rico, "Automatic Processing of Books to Generate a Quality Corpus for Educational Content," Rochester, NY, Apr. 2023.

[3] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Interspeech 2019*, pp. 2613–2617, Sep. 2019.

[4] S. Mihov, K. Schulz, C. Ringlstetter, V. Dojchinova, V. Nakova, K. Kalpakchieva, O. Gerasimov, A. Gotscharek, and C. Gercke, "A corpus for comparative evaluation of OCR software and postcorrection techniques," *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 162–166 Vol. 1, Aug. 2005.

[5] H. Chen, L. F. Pieptea, and J. Ding, "Construction and evaluation of a high-quality corpus for legal intelligence using semiautomated approaches," *IEEE Transactions on Reliability*, Vol. 71, No. 2, pp. 657–673, June 2022.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, Vol. 6, No. 1, p. 60, Jul. 2019.

[8] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A Survey of Data Augmentation Approaches for NLP," Dec. 2021.

[9] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," https://arxiv.org/abs/1901.11196v2, Jan. 2019.

[10] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," Jun. 2016.

[11] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?" Feb. 2023.

[12] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, and X. Li, "AugGPT: Leveraging ChatGPT for Text Data Augmentation," Mar. 2023.

[13] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, "Educational data augmentation in physics education research using ChatGPT," Jul. 2023.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[15] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training."

[16] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heinz, and D. Roth, "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey," Nov. 2021.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners."

[18] T. B. Brown, "Language Models are Few-Shot Learners," Jul. 2020.

[19] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, "Training Compute-Optimal Large Language Models," Mar. 2022.

[20] OpenAI, "GPT-4 Technical Report," Mar. 2023.

[21] J. Wang and Y. Dong, "Measurement of Text Similarity: A Survey," *Information*, Vol. 11, No. 9, p. 421, Aug. 2020.

[22] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. [Online]. Available: http://arxiv.org/abs/1904.09675

[23] Z. Bai, "An Analysis of English Vocabulary Learning Strategies," *Journal of Language Teaching and Research*, Vol. 9, No. 4, p. 849, Jul. 2018.

[24] J. P. Kincaid, Jr. Fishburne, R. Robert P., C. Richard L., and Brad S., "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel:," Defense Technical Information Center, Fort Belvoir, VA, Tech. Rep., Feb. 1975.

[25] "Reading levels per grade," https://northccs.com/misc/reading-levels-per-grade.html.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, Vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692