KIPS Trans. Softw. and Data Eng.
Vol.12, No.10 pp.455~460
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.10.455

TAGS: Text Augmentation with Generation and Selection

Kim Kyung Min† · Dong Hwan Kim† · Seongung Jo† · Heung-Seon Oh†† · Myeong-Ha Hwang†††

ABSTRACT

Text augmentation is a methodology that creates new augmented texts by transforming or generating original texts for the purpose of improving the performance of NLP models. However existing text augmentation techniques have limitations such as lack of expressive diversity semantic distortion and limited number of augmented texts. Recently text augmentation using large language models and few-shot learning can overcome these limitations but there is also a risk of noise generation due to incorrect generation. In this paper, we propose a text augmentation method called TAGS that generates multiple candidate texts and selects the appropriate text as the augmented text. TAGS generates various expressions using few-shot learning while effectively selecting suitable data even with a small amount of original text by using contrastive learning and similarity comparison. We applied this method to task-oriented chatbot data and achieved more than sixty times quantitative improvement. We also analyzed the generated texts to confirm that they produced semantically and expressively diverse texts compared to the original texts. Moreover, we trained and evaluated a classification model using the augmented texts and showed that it improved the performance by more than 0.1915, confirming that it helps to improve the actual model performance.

Keywords : Natural Language Process, Artificial Intelligence, Large Language Model, fEw-Shot Learning, Text Augmentation

생성-선정을 통한 텍스트 증강 프레임워크

김 경 민† · 김 동 환† · 조 성 웅† · 오 흥 선†† · 황 명 하†††

요 약

텍스트 증강은 자연어처리 모델의 성능 향상을 목적으로 원본 텍스트의 변환, 생성을 통하여 새로운 증강 텍스트를 생성하는 방법론이다. 기존 연구된 기법들은 표현적 다양성 부족, 의미 왜곡 , 한정적인 양의 증강 텍스트와 같은 한계점이 존재한다. 거대언어모델과 few-shot learning을 활용한 텍스트 증강은 이러한 한계점의 극복이 가능하지만, 잘못된 생성으로 인한 노이즈 발생의 위험성이 존재한다. 본 논문에서는 여러 후보 텍스트를 생성하고 적합한 텍스트를 증강 텍스트로 선정하는 TAGS를 제안한다. TAGS는 기존 텍스트 few shot learning을 통해 다양한 표현을 생성하면서 대조 학습과 유사도 비교를 통해 원본 텍스트가 적더라도 적합한 데이터를 효과적으로 선정한다. 이를 텍스트 증강이 필수적인 업무용 챗봇 데이터에 적용하여 60배 이상의 양적 향상을 달성하였다. 또한 증강 텍스트의 질적 향상을 확인하기 위해 실제 생성된 텍스트를 분석하여 원본 텍스트에 비해 의미론적, 표현적으로 다양한 텍스트를 생성함을 확인하였으며, 증강 텍스트로 실제 분류 모델을 학습하고 실험하여 실질적으로 자연어처리 모델 성능 향상에 도움이 되는 것을 확인하였다.

키워드 : 자연어 처리, 인공지능, 거대 언어 모델, 퓨샷 학습, 텍스트 증강

1. 서 론

자연어처리(Natural Language Process) 모델을 다양한 산업 분야에 적용하려면 해당 분야의 특수한 언어 특성을 반영한 학습 텍스트 데이터가 필수적이다. 데이터 구축은 복잡하고 어려운 작업이며, 편향된 데이터가 발생할 수도 있다. 텍스트 증강(Text Augmentation)은 원본 텍스트를 변형하거나 확장하여 새로운 증강 텍스트를 생성하는 기법으로, 모델의 일반화 능력과 강건성을 향상시킬 수 있다.

기존 텍스트 증강 기법은 원본 텍스트의 일부를 동의어 혹은 유사한 임베딩을 가진 단어로 대체하거나[1,2], 역번역을 통해 문장의 표현을 다양화하는 방법[3]들이 연구되었다. 그러나 이러한 방법론들은 원본 텍스트에 과도하게 의존하여 표현적 다양성 부족, 핵심 단어 변경으로 인한 의미 왜곡, 한정적인 양의 증강 텍스트와 같은 문제가 있다.

Few-Shot Learning(FSL)은 매우 적은 데이터로 과업을 수행하는 기법으로, 거대언어모델(LLM, Large Language Model)과 함께 크게 발전하였다. 특히, 텍스트 스타일 변환[4], 텍스트 요약[5]과 같이 기존 문장을 변형하는 과업에서 높은 성능을 달성하였다. 그러나 이러한 성과에도 불구하고, FSL을 사용한 텍스트 증강 연구는 거의 이루어지지 않았다. Table 1과 같이 LLM과 FSL을 활용한 텍스트 증강은 “어떤 자료가 있어야 할까”처럼 다양한 표현이나, “비과세소득”과 같이 의미론적으로 새로운 내용을 추가할 수 있다. 또한, few-shot 샘플에

※ This work was funded by the Korea Electric Power Corporation (KEPCO) (R22XO02-30) and “Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (MOE) (2021RIS-004).
† 비 회 원 : 한국기술교육대학교 컴퓨터공학과(석사)
†† 정 회 원 : 한국기술교육대학교 컴퓨터공학부 교수
††† 비 회 원 : 한국전력공사 전력연구원 디지털솔루션연구소 연구원
Manuscript Received : June 16, 2023
First Revision : August 7, 2023
Accepted : August 22, 2023
* Corresponding Author : Heung-Seon Oh(ohhs@koreatech.ac.kr)

Table 1. Examples of augmented text

(1) withholding_tax	
Original text	원천세 기준이 뭐야? (What is the criterion for withholding tax?)
Synonym	원천세 규격이 뭐야 (What is the basis for withholding tax?)
Embedding	원천세 폭탄이 뭐야 (What is the crisis for withholding tax?)
Back Translation	원천징수 기준은 무엇인가요?(What is the withholding tax threshold?)
FSL	원천세를 계산하기 위해 어떤 자료가 있어야 할까?(What materials do you have to work with withholding tax?)
	연말정산 간소화 서비스 제공(Provide year-end tax settlement simplification service)
(2) reasearch_expanse_withholding_tax	
Original text	위촉연구원 급여 원천세 증빙 작성(Documentation for Withholding Tax on Commissioned Researcher's Salary)
Synonym	위촉연구원 보수 원천세 증빙 집필(Documentation for Withholding Tax on Commissioned Researcher's Earnings)
Embedding	위촉연구원 급여 원천세 증빙해라(Apply for Withholding Tax on Commissioned Researcher's Salary)
Back Translation	위촉연구원 급여에 대한 원천징수 증명서 작성(Withholding Tax Certificate for Commissioned Researcher's Salary)
FSL	위촉연구원 원천세 증빙서 작성 시 비과세소득 입력 방법(Entering Tax Exempt Income for Commissioned Researchers'Withholding Tax Certificate)
	원천세 계산방법(Method of Withholding Tax Calculation)

따라 생성결과가 달라지므로, 이를 다양화하면 기존 방법론보다 훨씬 많은 텍스트 증강이 가능하다. 그러나 Table 1-1의 “연말정산 간소화 서비스 제공”과 같이 대상 도메인에 포함되지 않는 Out-of-Domain(OOD) 클래스 문장이 생성되거나, Table 1-2의 “원천세 계산방법”처럼 도메인 안의 다른 클래스에 해당하는 in-domain negative 클래스 문장이 생성될 가능성이 있다. 이러한 문장들은 노이즈로 작용하여 모델의 성능에 치명적인 영향을 줄 수 있다.

FSL기반 텍스트 스타일 변환을 수행한 prompt and rerank [4]에서는 여러 개의 후보 텍스트를 생성한 후, 목표 스타일과의 일치도 등을 기준으로 적합한 후보들을 선택하는 선정 과정을 거쳐 높은 성능을 달성하였다. 이와 유사하게, FSL을 활용한 텍스트 증강에서도 여러 후보 텍스트를 생성하고, 선정 과정을 거침으로써 노이즈를 완화할 수 있다. 적합한 후보 텍스트의 선정 기준은 OOD 클래스와 negative 클래스에 해당되는 텍스트들을 제외하고 목표 클래스에 해당하는 텍스트만 선정할 수 있어야 한다. 또한, 텍스트 증강의 목적성을 고려하였을 때 선정 과정은 적은 원본 텍스트 데이터를 가지고 있을 때에도 동작되어야 한다.

본 논문에서는 생성 과정과 선정 과정을 통한 텍스트 증강 프레임워크인 TAGS(Text Augmentation with Generation and Selection)을 제안한다. TAGS는 클래스가 존재하는 데이터에 대해 작동하며, 두 과정 모두 클래스의 정보를 활용한다. 생성 과정에선 디코더 구조의 LLM과 FSL을 통해 여러 개의 후보 텍스트를 생성한다. 선정 과정에서는 대조 학습(Contrastive Learning)으로 in-domain class 정보를 미리 학습시킨 인코더 구조의 LLM을 사용하여 원본 텍스트와 후보 텍스트의 임베딩을 생성하고, 이를 바탕으로 OOD 클래스와 in-domain 클래스를 모두 고려한 적합한 텍스트를 선정한다. 또한, 생성된 증강 텍스트를 few-shot 샘플로 재사용하여 증강 가능한 데이터의 수가 한정적이라는 기존 방법들의 단점을 극복하였다. TAGS를 실무용 챗봇 텍스트 데이터에 실험하여 기존대비 60배 이상에 달하는 양적 증가를 달성하였다. 또한, 증강된 데이터로 의도 분류 모델을 학습하여 기존 대비 0.1915의 정확도 향상을 달성하였고 이를 통해 증강 데이터의 질적 우수성을 간접적으로 확인하였다.

본 연구의 기여점은 다음과 같다.

1. FSL을 통해 다양하게 생성하면서 별도의 선정 과정을 통해 노이즈까지 완화할 수 있는 텍스트 증강 프레임워크 TAGS를 제안하였다.

2. 실제 업무용 챗봇 데이터에 TAGS를 적용하여 60배 이상의 양적 증가를 달성하였으며, 증강 데이터로 학습한 모델에서의 성능 향상을 통해 질적 향상을 간접적으로 확인하였다

2. 관련 연구

2.1 텍스트 증강

텍스트 증강은 자연어처리 모델의 성능을 향상시키기 위해 학습 텍스트 데이터를 인공적으로 변형, 생성하여 다양성과 양을 증가시키는 방법론이다. 텍스트 증강은 분류 태스크에서 특히 유용하게 활용되는데, 이때는 증강 데이터의 다양성 뿐만 아니라 원본 텍스트 데이터의 클래스 정보 또한 고려해야 한다.

이전 연구들 에서는 주로 원본 텍스트의 일부 단어 혹은 전체 문장을 변형하는 방법론들이 제안되었다. 단어 단위의 변형을 통한 텍스트 증강 기법으로는 몇 개의 단어를 동의어[2] 혹은 비슷한 임베딩[1]을 가진 단어로 대체하는 등의 방법론들이 제안되었다. 한편, 문장 단위의 변형으로는 원본 텍스트를 다른 언어로 번역 후 기존 언어로 다시 번역하는 역번역을 진행하거나[3] 문장의 요약을 통해 의미는 유지하면서 표현을 다향화 하는 방법[6]들이 연구되었다. 그러나 이러한 방법들은 원본 텍스트에 크게 의존하고, 때로는 클래스 정보가 훼손될 수도 있다. 또한, 생성 가능한 증강 텍스트 데이터의 수 역시 매우 한정적이다.

AugGPT의[7] 경우 chatGPT를 활용하여 생성 기반의 텍스트 증강을 진행한다. 그러나 chatGPT 자체 성능에 크게 영향을 받으며, 생성 후 별도의 선정 과정이 없어 noise 발생에 취약하다는 단점이 있다.

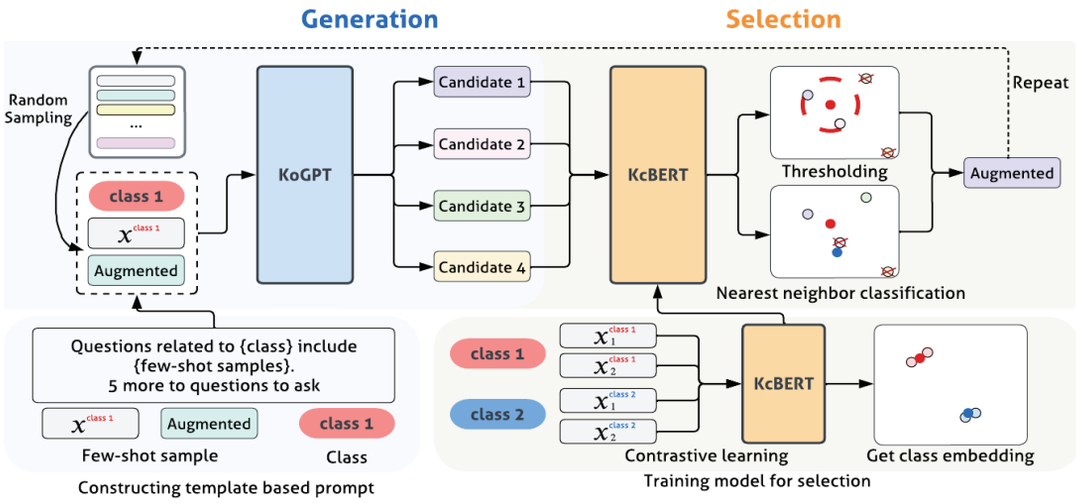


Fig. 1. Overall Architecture of TAGS

2.2 Few-shot learning

Few-shot learning(FSL)은 딥러닝 모델이 학습 데이터의 수에 의존적이라는 한계점을 극복하기 위한 방법들이다. 거대 언어모델(LLM, Large Language Model)을 활용한 FSL은 대량의 말뭉치로 사전학습한 지식들을 하위 태스크에 활용하여 다양한 태스크에서 높은 성능을 달성하였다[8].

특히, 생성기반 LLM에 프롬프트를 구성하여 FSL을 진행하는 방식은 감정분석[9], 개체명 분석[10] 등의 분류 태스크 뿐 아니라 텍스트 스타일 변환[4], 텍스트 요약[5]과 같이 기존 문장을 변형하거나 생성하는 태스크에서도 높은 성능을 달성하였다. 그러나 이러한 기법들은 잘못된 텍스트가 생성될 가능성이 있으며, 프롬프트만으로 이를 제어하는 것에는 한계가 있다. FSL을 통한 텍스트 스타일 변환 프레임워크인 Prompt and Rerank [4]에서는 여러 개의 후보 텍스트들을 생성하고, 목표 스타일과의 일치성들을 바탕으로 적합한 텍스트들을 선정하는 방식으로 비교적 작은 사이즈의 LLM으로 높은 성능을 달성했다.

그러나 이러한 성과에도 불구하고 FSL을 활용한 텍스트 증강 연구는 거의 진행되지 않았다. 텍스트 스타일 변환, 텍스트 요약 등을 통한 텍스트 증강은 가능하지만, 원본 텍스트 데이터의 클래스 정보가 고려되지 않기 때문에 별도의 처리없이 이를 증강 텍스트 데이터로 활용하는 것은 부적절하다.

3. 제안방법

TAGS의 전체 프레임워크는 Fig. 1과 같다. 도메인 내에 존재하는 임의의 클래스를 l라 하고, 해당 클래스의 원본 텍스트를 x^l이라 한다. TAGS는 x^l가 입력으로 들어가 후보 텍스트를 만든 후, 후보 텍스트에서 증강 텍스트를 선정하는 과정으로 진행된다.

3.1 생성과정

생성 과정에서는 FSL을 통해 여러 개의 후보 텍스트를 생성한다. 우선, l의 정보와 무작위로 선정된 k개의 x^l를 few-shot 샘플로 선택하고, 미리 작성한 템플릿에 삽입하여 프롬프트를 구성한다. l의 정보는 클래스의 이름을 사용하였다. 이후, 구성된 프롬프트를 디코더 구조의 LLM에 입력하여 여러 개의 후보 텍스트를 생성한다.

이 때, 사용된 few-shot 샘플과 완전히 동일한 텍스트가 생성될 가능성이 높다. 프롬프트만으로 이러한 중복 생성을 제어하는 것은 한계가 있기 때문에 CTRL[11]에서 사용된 repetition penalty를 적용하였다. Repetition penalty는 이미 존재하는 토큰들에 대해서는 토큰 샘플링 시에 페널티를 주는 방식으로, 원본 텍스트와 동일한 생성을 방지하고 더욱 다양한 표현을 생성할 수 있도록 유도한다.

3.2 선정과정

선정 과정에서는 생성된 후보 텍스트를 기준에 따라 증강 텍스트로 선정한다. 적합한 증강 텍스트는 목표 클래스의 정보를 충분히 포함하여 해당 클래스로 분류 가능한 텍스트로 정의 가능하다.

원본 텍스트 x^l은 클래스 l에 속하는 텍스트로, 후보 텍스트가 x^l과 유사하다면 후보 텍스트가 목표 클래스에 해당한다고 할 수 있다. 인코더 구조의 LLM을 통해 임베딩을 생성하고, 후보 텍스트와 원본 텍스트의 임베딩 간 유사성을 측정할 수 있다.

그러나 범용 LLM에는 도메인의 클래스 정보가 포함되어 있지 않다. 따라서 같은 클래스이지만 문장의 표현이 크게 달라 상이한 임베딩을 가지거나, 다른 class이지만 문장이 비슷하여 유사한 임베딩을 가질 수 있다. 이러한 문제를 해결하기 위하여 임베딩의 유사도 기반 비교에 적합할 뿐 아니라[12], few-shot에서도 높은 일반화 능력을 가지는[13,14] Supervised Contrastive Learning (SCL loss)를 사용하여 미세조정을 진행하였다.

$$L_{SCL} = \sum_{i=1}^{I} -\log \frac{\sum_{j=1}^{J} e^{\mathrm{sim}\left(x_i^l, x_j^l\right)}}{\sum_{j=1}^{J} e^{\mathrm{sim}\left(x_i^l, x_j^l\right)} + \sum_{k=1}^{K} e^{\mathrm{sim}\left(x_i^l, x_k^{l^-}\right)}} \quad (1)$$

이때 l^-은 l이 아닌 모든 클래스이며, I, J는 클래스 l에 해당하는 배치 안의 텍스트의 수, K는 l^-에 해당하는 배치 안의 텍스트 수 이다. SCL은 동일한 클래스를 가진 임베딩은 가깝도록, 다른 클래스를 가진 임베딩은 서로 멀도록 학습시킨다. 이를 통해 임베딩 공간에 클래스 정보를 간접적으로 주입하여 클래스 정보가 포함되지 않는다는 문제를 완화할 수 있다. 이후, 학습에 사용된 x^l들의 평균 임베딩을 클래스 임베딩으로 사용하여 후보 임베딩과 비교를 진행한다.

임베딩 간의 비교는 임계값 측정(thresholding)과 최근접 분류(nearest neighbor classification)를 통해 진행된다. 임계값 측정은 목표 클래스의 클래스 임베딩과 후보 텍스트 임베딩의 유사도가 τ 이상일 경우에만 선정한다. 이는 out-of-domain class에 해당하는 텍스트를 필터링하는데 효과적이다. 최근접 분류는 후보 임베딩과 모든 클래스 임베딩의 유사도를 구하고, 목표 클래스 임베딩과 가장 유사할 경우에만 선정한다. 이는 in-domain negative class에 해당하는 텍스트를 필터링하는데 효과적이다. 최종적으로 두 조건을 모두 만족하는 후보 텍스트만 증강 텍스트로 선정한다.

3.3 반복과정

생성 과정과 선정 과정을 반복적으로 수행함으로 많은 양의 증강 텍스트 데이터를 생성할 수 있다. 더 나아가, 생성 과정에 원본 텍스트 뿐 아니라 증강 텍스트를 few-shot 샘플로 활용하여 새로운 증강 텍스트를 생성하는 것이 가능하다. 위 과정을 반복하여 생성 가능한 증강 텍스트 데이터의 수를 극대화시킬 수 있다.

4. 실 험

본 논문에서는 업무용 챗봇을 위한 데이터셋을 활용하여 실험하였다. 챗봇 데이터 셋은 업무 관련 클래스(예시: “정부과제 민감부담금”, “SW개발 용역 검수”)와 일상 관련 클래스(예시:다이어트, 음악)를 포함한 총 403개의 클래스로 구성되며 각 클래스는 업무에 따라 “원천세 징수”와 “원천세”처럼 매우 세부적으로 구분될 수 있다. 불특정 다수에게 사용되는 챗봇의 특성상 모델의 강건성이 중요하며, 사람을 통한 데이터 구축은 구축자 개인의 특성이 반영된 구축만 가능하므로, 별도의 텍스트 증강이 필수적이다.

TAGS의 성능을 확인하기 위하여 각 클래스 당 k개의 원본 텍스트를 통해 최대 300개의 증강 텍스트를 생성하고, 원본 텍스트와 BLEU 점수를 계산하여 다양성에 대한 직접적인 평가를 수행하였다. 또한, 증강된 텍스트로 의도 분류 모델에 학습하여 별도로 구축한 테스트셋에 실험하여 간접적인 평가를 진행하였다. 의도 분류 모델은 110M개의 파라미터를 가진 KoELECTRA-base[15]를 사용하여 batch size 256으로, 200 epoch으로 학습하였다. 테스트셋은 클래스당 50개 이상, 총 22,447개의 텍스트로 구성하였다. TAGS의 생성 과정에서는 KoGPT-base[16]를 사용하여, “{class}와 관련된 질문은 {few shot samples} 등이 있다. 추가로 만들 수 있는 질문 5개는}”를 프롬프트를 통해 한번의 생성에서 5개의 새 후보 텍스트를 생성하였다. 선정 과정에서는 KcBERT-base[17]를 활용해 original text로만 대조 학습을 진행한 모델을 사용하였다.

Table 2. Result Comparison of Augmentation Methods

Method	Accuracy	Macro F1
2-shot		
Baseline	0.3842	0.3614
Back Translation	0.3936	0.3778
Synonym Replacement	0.4075	0.4331
Embedding-based	0.4478	0.4247
TAGS-4	0.5072	0.4747
TAGS-250	**0.7102**	**0.7432**
5-shot		
Baseline	0.7070	0.6810
Back Translation	0.7728	0.7957
Synonym Replacement	0.7795	0.7905
Embedding-based	0.7223	0.7078
TAGS-10	0.8460	0.8529
TAGS-250	**0.8985**	**0.8947**

Table 2는 각 텍스트 증강 기법 별 성능을 비교한 표이다. 모든 기법은 원본 텍스트 하나 당 두 개의 증강 텍스트를 생성하여 2-shot에서는 클래스당 4개의 증강 텍스트, 5-shot에서는 클래스당 10개의 증강 텍스트를 생성하였고, TAGS의 경우 추가적으로 클래스 당 250개의 증강 텍스트를 생성하여 실험하였다. Back translation에는 영어와 일본어 번역을 사용하였고, embedding-based에는 KcBERT-base가 활용되었다. 2-shot과 5-shot에서 모두 TAGS로 생성한 증강 텍스트가 성능 향상에 도움이 되며, 기존 텍스트 증강 기법들과 비교하여도 0.06 이상 성능 향상에 도움이 되는 것을 알 수있다. 또한, 기존 텍스트 증강 기법은 생성 가능한 텍스트의 수가 한정적인 반면, TAGS의 경우 기존 대비 100배 이상의 텍스트 생성이 가능하며, 생성된 데이터가 모델 성능 향상에 많은 도움을 주는 것을 알 수 있다.

Table 3은 생성한 증강 텍스트 수에 따른 실험 결과로, 전체적으로 낮은 BLEU 점수를 가지면서 0.1632이상의 성능 향상을 보인다. 이를 통해 TAGS가 원본 텍스트와 다른 다양한 표현을 가지면서 클래스 정보는 유지하는 증강 텍스트를 생성함을 확인할 수 있다. 또한, 증강 텍스트의 수를 증가시키면 BLEU 점수는 감소하지만 분류 모델의 성능은 향상되는 경향을 보이는데, 반복 과정이 증강 텍스트의 양적, 질적 향상에 도움이 되는 것을 알 수 있다. 그러나 증강 텍스트가 250개 이상 넘어갈 경우 미세한 성능 감소가 발생하였다. 이는 반복이 진행되면서 실제 클래스에 해당하는 문장은 점점

Table 3. Results of Varying the Amounts of Augmented Text

Augmented text	Accuracy	BLEU
0(Baseline)	0.7070	-
50	0.8702	0.1274
100	0.8822	0.1030
150	0.897	0.0900
200	0.8965	0.0840
250	**0.8985**	**0.0790**
300	0.8962	0.0760

Table 4. Examples of Augmented Texts from TAGS

(1) 음악
Original text
- 아이돌 노래도 좋은거 같아 - 이 가수가 국힙원탑이지 - 음악 콘서트 버전이 훨씬 좋아 - 풍성한 사운드가 있는 음악을 좋아해 - 아침에 재즈 듣는 게 너무 좋아
Augmented text
- 흑인음악을 들으면 마음이 편해지는 느낌이 들어 - 요즘 난 매일매일 새로운 곡들을 찾아 듣곤 해 - 가요와 팝을 모두 즐겨 들어요 - 아이돌 노래 중에는 빅뱅 노래가 최고야 - 팝송 중에서 내가 가장 좋아하는 노래는 마이클 잭슨의 빌리 진 (Billie Jean)이야 - 예술
(2) 원천세
Original text
- 원천징수된 건가요? - 원천세 처리 - 원천세는 어디서 확인 가능해? - 원천세 기준이 뭐야? - 원천세 업무
Augmented text
- 간이세액표에 의한 원천징수세율 인하로 인한 추가적인 세수 감소액은 어떻게 해야 하나? - 원천세를 계산하기 위해 어떤 자료가 있어야 할까? - 원천세 통합에 따른 과세체계 개편방안 - 원천세 경정청구 - 원천세는 언제까지 내야 하나? - 부가가치세법 제12조 제1항 제6호의 규정에 의한 의제매입세액공제제도

포화되어 이미 존재하는 문장과 유사하게 생성되는 경향을 보이나, 노이즈에 해당하는 문장은 그렇지 않기 때문에 발생하는 문제로 분석된다.

Table 4는 실제 TAGS를 통해 생성한 증강 텍스트들이다. 단순히 원본 텍스트들의 문맥을 활용할 뿐 아니라, LLM의 내부 지식을 활용하여 완전히 새로운 텍스트를 생성하는 것을 볼 수 있다. 특히, Table 4-1의 “빌리 진”, ”빅뱅”처럼 구체적인 예시를 생성하거나, Table 4-2의 “간이세액표”, “경정청구”처럼 전문적인 내용을 포함하는 것을 확인할 수 있다. 그러나 “예술”처럼 해당 클래스에 속하는 너무 포괄적인 개념이나, “부가가치세법”처럼 과하게 세부적인 내용을 생성, 선정하는 문제 또한 존재한다.

Table 5는 선정과정을 진행하지 않고 생성한 증강 데이터와의 성능 비교이다. 선정과정을 진행하지 않은 경우 TAGS에 비해 0.05이상의 성능 하락이 나타나며 선정과정이 실제로 증강 텍스트의 질적 향상에 도움이 되는 것을 확인할 수 있다.

5. 결 론

본 논문에서는 LLM을 활용한 텍스트 증강 프레임워크인 TAGS(Text Augmentation with Generation and Selection)를 제안한다. TAGS는 생성 과정과 선정 과정을 반복적으로 진행하여 양적, 질적으로 우수한 증강 텍스트를 생성한다. 생성 과정에서는 few-shot learning을 활용하여 여러 개의 후보 텍스트를 생성하고, 선정 대조 학습과 임베딩 간의 비교를 통하여 in-domain negative와 OOD 클래스가 모두 고려된 증강 데이터를 선정하였다.

Table 5. Results with Absence of Selection Process

Augmented text	TAGS	w/o selection
250	0.8985	0.8485

그러나 본 연구에서는 선정 과정에 사용되는 모델은 원본 텍스트 데이터로만 학습되어 있기 때문에 원본 텍스트 데이터의 질과 양에 의존적이라는 한계점이 있다. 또한, 반복이 진행됨에 따라 노이즈의 발생 또한 증가할 수 있다는 문제가 있다. 이를 극복하기 위하여 단순 증강 데이터를 단순히 few-shot sample로 활용할 뿐 아니라 분류 모델 또한 학습시키는 방법이 향후 연구되어야 한다.

References

[1] W. Y. Wang and D. Yang, “That’s so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal: Association for Computational Linguistics*, pp.2557-2563, Sep. 2015.

[2] E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, “PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China: Association for Computational Linguistics, pp.425-430, Jul. 2015.

[3] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Berlin, Germany: Association for Computational Linguistics, pp. 86-96, Aug. 2016.

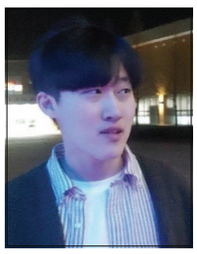
[4] M. Suzgun, L. Melas-Kyriazi, and D. Jurafsky, “Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2195-2222, Dec. 2022.

[5] A. Fabbri et al., “Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, pp.704-717, Jun. 2021.

[6] V. Gangal, S. Y. Feng, M. Alikhani, T. Mitamura, and E. Hovy, "Nareor: The narrative reordering problem," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp.10645-10653, 2022.
[7] H. Dai et al., "AugGPT: Leveraging ChatGPT for Text Data Augmentation," 2023.
[8] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," ACM Computing Surveys (CSUR), Vol.53, No.3, pp.1-34, 2020.
[9] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," *CoRR*, Vol. abs/1909.00161, 2019,
[10] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-based named entity recognition using BART," in *Findings of the Association for Computational Linguistics*: ACL-IJCNLP 2021, pp.1835-1845, 2021.
[11] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," *arXiv preprint arXiv:1909.05858*, 2019.
[12] T. Gao, X. Yao, and D. Cehn, "SimCSE: Simple contrastive learning of sentence embeddings," *2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2021. Association for Computational Linguistics (ACL), 2021.
[13] C. Liu, "Learning a few-shot embedding model with contrastive learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.35, No.10, pp.8635-8643, May 2021.
[14] J. Zhang et al., "Few-shot intent detection via contrastive pre-training and fine-tuning," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp.1906-1912, 2021.
[15] Park, Jangwon, KoELECTRA: Pretrained ELECTRA Model for Korean [Internet], https://github.com/monologg/KoELECTRA
[16] Ildoo Kim and Gunsoo Han and Jiyeon Ham and Woonhyuk Baek,KoGPT: KakaoBrain Korean(hangul) Generative Pre-trained Transformer [Internet], https://github.com/kakaobrain/kogpt
[17] Lee, Junbum, "KcBERT: Korean Comments BERT," *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pp.437-440, 2020.

김 경 민

https://orcid.org/0009-0009-7692-4334
e-mail : dukes123@koreatech.ac.kr
2022년 한국기술교육대학교 컴퓨터공학부 (학사)
2023년 ~ 현 재 한국기술교육대학교 컴퓨터공학과(석사)
관심분야 : Deep learning, Natural Language Process



김 동 환

https://orcid.org/0009-0008-0442-8211
e-mail : hwan6615@koreatech.ac.kr
2021년 한국기술교육대학교 컴퓨터공학부 (학사)
2022년 ~ 현 재 한국기술교육대학교 컴퓨터공학과(석사)
관심분야 : Deep learning, Natural Language Process

조 성 웅

https://orcid.org/0000-0003-3325-0412
e-mail : oowhat@koreatech.ac.kr
2020년 한국기술교육대학교 컴퓨터공학부(학사)
2023년 한국기술교육대학교 컴퓨터공학과(석사)
관심분야 : Deep learning, Natural Language Process

오 흥 선

https://orcid.org/0000-0002-9193-8998
e-mail : ohhs@koreatech.ac.kr
2009년 한국과학기술원 전산학(석사)
2014년 한국과학기술원 전산학(박사)
2013년 ~ 2018년 한국과학기술정보연구원 (KISTI)
2018년 ~ 현 재 한국기술교육대학교 컴퓨터공학부 교수
관심분야 : Deep learning, Natural Language Process

황 명 하

https://orcid.org/0000-0002-6887-8552
e-mail : mh.hwang@kepco.co.kr
2015년 충남대학교 정보통신공학과(학사)
2018년 과학기술연합대학원대학교 정보통신네트워크공학(석사)
2022년 한국과학기술원 바이오 및 뇌공학(박사수료)
2019년 ~ 현 재 한국전력공사 전력연구원 디지털솔루션연구소 연구원
관심분야 : Natural Language Processing(NLP), Deep Learning, Bioinformatics