복수 대규모 언어 모델에 기반한 제어 가능형 데이터 증강 프레임워크

강현석°, 남궁혁, 정지수, 정상근*
충남대학교 바이오AI융합학과, 충남대학교 컴퓨터융합학부, 충남대학교 컴퓨터융합학부, 충남대학교 컴퓨터융합학부
{dnfldjaak11}@gmail.com, {hyuk199}@gmail.com, {jisu.jung5}@gmail.com, {hugmanskj}@gmail.com

Controllable data augmentation framework based on multiple large-scale language models

Hyeonseok Kang°, Hyuk Namgoong, Jeesu Jung, Sangkeun Jung†
Chungnam National University, Chungnam National University, Chungnam National University , Chungnam National University

요 약

데이터 증강은 인공지능 모델의 학습에서 필요한 데이터의 양이 적거나 편향되어 있는 경우, 이를 보완하여 모델의 성능을 높이는 데 도움이 된다. 이미지와는 달리 자연어의 데이터 증강은 문맥이나 문법적 구조와 같은 특징을 고려해야 하기 때문에, 데이터 증강에 많은 인적자원이 소비된다. 본 연구에서는 복수의 대규모 언어 모델을 사용하여 입력 문장과 제어 조건으로 프롬프트를 구성하는 데 최소한의 인적 자원을 활용한 의미적으로 유사한 문장을 생성하는 방법을 제안한다. 또한, 대규모 언어 모델을 단독으로 사용하는 것만이 아닌 병렬 및 순차적 구조로 구성하여 데이터 증강의 효과를 높이는 방법을 제안한다. 대규모 언어 모델로 생성된 데이터의 유효성을 검증하기 위해 동일한 개수의 원본 훈련 데이터와 증강된 데이터를 한국어 모델인 KcBERT로 다중 클래스 분류를 수행하였을 때의 성능을 비교하였다. 다중 대규모 언어 모델을 사용하여 데이터 증강을 수행하였을 때, 모델의 구조와 관계없이 증강된 데이터는 원본 데이터만을 사용하였을 때보다 높거나 그에 준하는 정확도를 보였다. 병렬 구조의 다중 대규모 언어 모델을 사용하여 400개의 원본 데이터를 증강하였을 때에는, 원본 데이터의 최고 성능인 0.997과 0.017의 성능 차이를 보이며 거의 유사한 학습 효과를 낼 수 있음을 보였다.

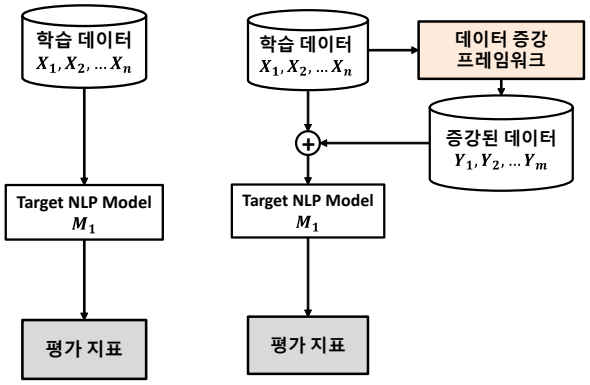
주제어: 대규모 언어 모델, 데이터 증강, 자연어 생성, 문장 생성, 유사의미 문장 생성

1. 서론

데이터 증강은 인공지능 모델이 현실 세계의 다양한 상황에 적응하고, 예측 오류를 줄이는 데 도움이 된다. 데이터 증강은 데이터의 양이 부족하거나, 편향이 심한 경우에도 유용하다. 자연어 데이터 증강 기법 중 하나로 유사의미 문장을 생성하는 기법이 존재한다[1]. 문장의 유사성을 정의하여 제어하는 방법이 연구되었으며 문장 제어를 위한 데이터 구축 및 생성 모델들이 개발되었다.

본 연구는 유사의미 문장을 생성하는 모델을 개발하기보다는, 최근 각광받는 대규모 언어 모델을 활용하여 유사의미 문장 생성을 할 수 있는 방법을 제시한다. 이러한 방법을 통해, 대규모 언어 모델을 활용해 훈련없이(model-less) 문장 생성과 프롬프트 입력을 통한 대규모 언어 모델의 제어가 가능해진다. 생성된 유사의미 문장들을 최종 언어 모델의 다중 클래스 분류 훈련에 사용하여 대규모 언어 모델을 활용한 데이터 증강 기법이 적합한지 평가한다.

본 연구에서는 복수 대규모 언어 모델들을 활용하는 방법으로 2가지 방안을 제시한다. 1)입력 문장과 생성되는 문장의 의미 유사도를 높이기 위해 대규모 언어 모델을 병렬로 구성하여 높은 의미 유사도의 생성 문장을 선별하는 방법이다. 2) 의미가 유사하면서도 다양한 어휘와 문장 구조를 가진 문장을 생성할 수 있도록 대규모 언어 모델을 순차적으로 구성하여 이

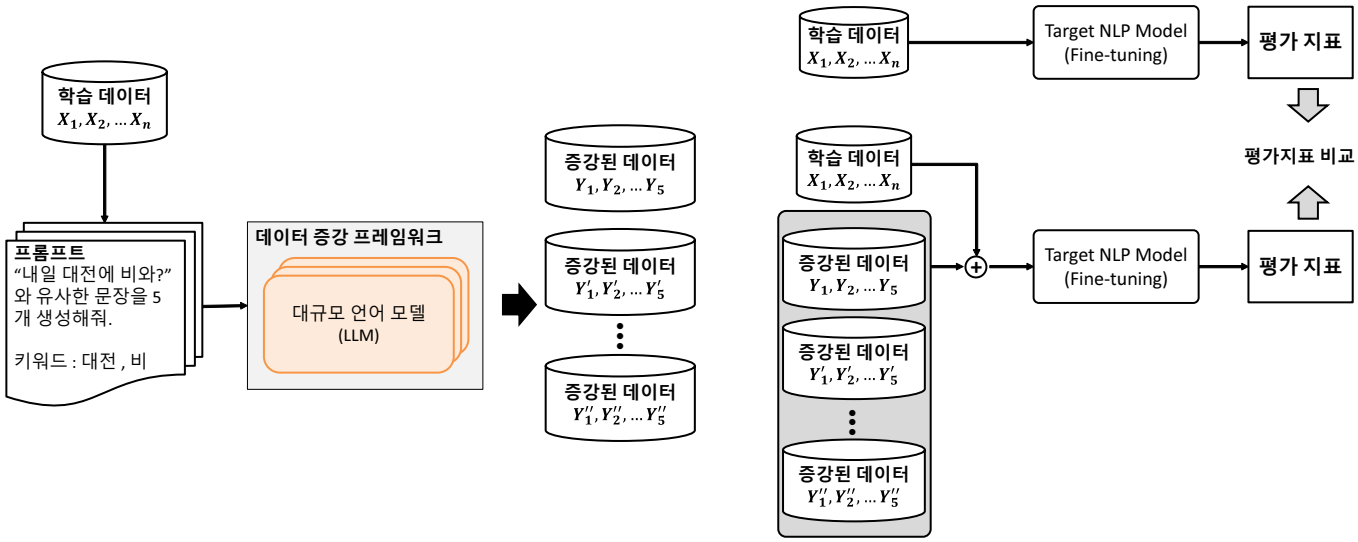


(a) 기본 훈련데이터 사용 (b) 증강데이터 사용

그림 1. 데이터 기반 언어 처리 훈련 방식

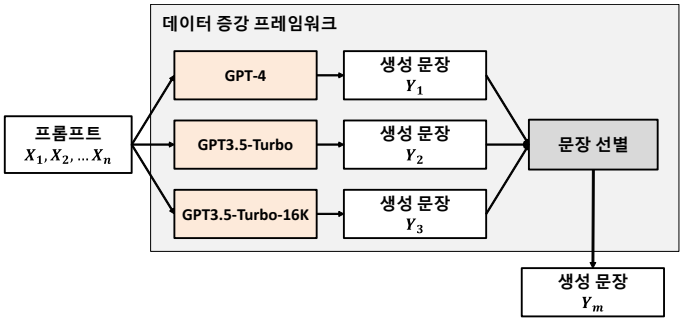
전 모델로부터 생성된 문장을 다음 모델이 수정하는 과정으로 생성 문장에 다양성을 주는 방법이다. 여러 언어 모델의 결과물을 보완하여, 보다 신뢰성 있는 결과물을 생성하는 새로운 방법론을 제시한다.

프롬프트 입력을 통한 대규모 언어 모델의 제어는 대규모 언어 모델이 생긴 이후 진행되었다. 우리 연구는 대규모 언어 모델에 중요한 키워드를 지정하고 프롬프트와 함께 입력한다. 이는 유사의미 문장을 생성할때 의미가 변형이 되는 것과 라

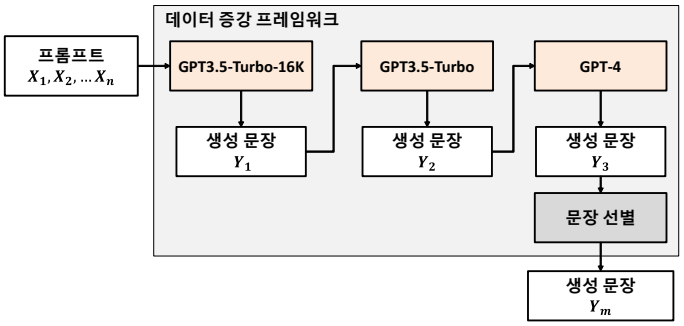


(a) 복수 대규모 언어 모델에 의한 데이터 증강 (b) 제안한 증강데이터 프레임워크의 성능 평가 방법

그림 2. 데이터 증강 프레임워크의 전체 구조. 유사의미 문장 프레임워크는 두개의 단계로 진행이 된다. 그림 2a는 첫 단계로 일부 훈련 데이터를 기반으로 대규모 언어 모델을 활용해 유사의미 문장을 생성하는 데이터 증강 단계이다. 그림 2b는 다음 단계로 증강된 데이터와 데이터 증강에 사용된 기존 훈련 데이터를 함께 언어 모델의 미세조정에 사용한다.



(a) 병렬 구조의 다중 대규모 언어 모델



(b) 순차 구조의 다중 대규모 언어 모델

그림 3. 다중 대규모 언어 모델 구조

벨에 적합하지 않은 문장을 생성하는 것을 방지한다. 생성된 문장 데이터로 훈련을 진행하며 키워드의 유무 차이로 인한 훈련 모델의 성능을 비교 평가해 프롬프트 입력 제어의 유효성을 증명하고자 한다.

본 논문은 다음과 같은 구성을 가진다. 2장에서는 관련 연구에 대해 소개하고, 3장에서는 대규모 언어 모델을 사용한 유사의미 문장 생성 방법과 다중 언어 모델의 구조에 대해 소개한다. 4장에서는 실험에 대한 구체적인 방법과 결과에 대해 소개하고, 마지막으로 5장에서는 본 연구에 대한 결론을 정리한다.

2. 관련 연구

2.1 프롬프트 기반 대규모 언어 모델 제어

대규모 언어 모델에 대한 대답을 사용자가 원하는 형태, 문체 문맥 등으로 출력하기 위한 입력 정보 제어(Prompt Engineering[2]) 방법이 존재한다. Prompt Engineering은 다양한 주제의 언어 모델을 효율적으로 제어할 수 있도록 입력 문장인 프롬프트를 개발하고 최적화하는 것을 목적으로 하는 비교적 새로운 분야이다. Prompt Engineering은 대규모 언어 모델과 상호 작용하고 개발하는 데 유용한 스킬과 기법을 포함하며, 대규모 언어 모델을 이해하는 데 중요한 기술이다.

2.2 유사의미 문장 생성

유사의미 문장 생성은 자연어 문장의 의미는 같지만 구문 또는 어휘의 형태가 다른 새로운 문장을 생성한다. 연구[3]에서는 유사의미 문장 생성 작업인 의역문 생성 문제를 다루었다. 다양한 언어 모델들의 의역문 생성에 대한 평가가 되어있으며 언어 모델로 유사의미 문장 생성에 대해 연구가 진행되었다.

이러한 생성된 문장들은 자연어 데이터 증강에 사용될 수 있으며 적은 데이터에서 다양한 방법으로 양을 늘려 신경망

표 1. 데이터 증강 실험을 위한 샘플링 데이터셋

데이터 구성	훈련 데이터	검증 데이터	평가 데이터
전체 데이터	6,993	3,009	2,998
N=100	100	20	
N=200	200	40	
N=400	400	80	

모델의 성능을 향상시키는데 사용된다. 자연어 데이터 증강에 관하여 연구[4]에서는 동의어 대체, 무작위 삽입, 무작위 교체, 무작위 삭제 등 간단하지만 강력한 기법이 사용하였으며, 언어 모델로 인한 문장 생성이 아닌 다양한 기법들로 유사의미 문장을 생성한 것이다.

3. 복수 대규모 언어 모델을 사용한 유사의미 문장 생성 방법

본 연구에서는 복수 대규모 언어모델을 사용한 유사의미 문장 생성을 통한 데이터 증강 방법을 제안한다. 제안한 방법을 통해 최소한의 인적 자원으로 훈련 데이터의 생성이 가능하다. 대규모 언어 모델을 문장 생성은 (1) 키워드 반영 프롬프트 기반의 출력 제어 (2) 복수 대규모 언어 모델의 조합 방법의 두 단계로 구성된다.

3.1 키워드 반영 프롬프트 기반의 출력 제어

대규모 언어 모델을 사용한 문장 생성의 첫 단계로, 생성할 문장의 특징과 구조와 같은 제어 조건을 입력 문장과 함께 프롬프트를 구성한다. 본 연구에서는 유사의미 문장 생성을 위해 입력 문장에 포함되어 있는 키워드와 키워드의 등장 순서를 제어 조건으로 지정하여 프롬프트를 구성하였다. 그리고 인적 자원의 최소화를 위해 생성된 문장과 입력 문장을 비교하여 문장 평가와 선별하는 과정인 의미 유사도를 대규모 언어 모델의 활용으로 점수화하였다. 대규모 언어 모델의 의미 유사도는 0부터 10점 사이의 점수로 표현하도록 프롬프트의 제어 조건을 구성하였다.

3.2 복수 대규모 언어 모델의 조합 방법

단일 모델의 사용과 달리 복수의 대규모 언어 모델 사용에서는 다양한 모델의 배치를 고려해 볼 수 있다. 본 연구에서는 GPT-3.5-turbo [5], GPT-3.5-turbo-16K, GPT-4를 사용하여 그림 3와 같이 2가지 형태의 모델로 구성하였다. 이 3개의 단일 대규모 언어 모델과 함께 2가지 다중 대규모 언어 모델을 사용하여 생성한 문장으로 언어 모델을 훈련하여 나온 성능에 대해 비교 평가를 진행하였다.

병렬 구조 그림 3a는 세 모델을 병렬로 구성하여 각 모델의 출력을 병합하는 방법을 사용하였다. 입력 문장과 제어조건이 입력된 프롬프트를 3개의 대규모 언어 모델에 입력하여 생성된 문장을 취합한다. 생성된 문장은 각각의 대규모 언어 모델에 의해 의미 유사도가 점수화되어 출력되며, 병합한 문장은 선별할 문장의 개수만큼 의미 유사도가 높은 문장부터 순서대로 추출한다. 단일 대규모 언어 모델에서 다수의 문장을 생성하게 되면 의미 유사도 점수의 편차가 커지는 현상이 발생한다. 이러한 문제를 복수의 대규모 언어 모델로부터 얻어진 문장들로부터 높은 유사도를 가지는 문장들만 선별하는 것으로 데이터 증강에 사용되는 생성 문장들의 의미 유사도 편차를 줄일 수 있다.

순차 구조 그림 3b는 한 모델의 출력이 다음 모델의 입력으로 사용되는 순차적인 방법이다. 이전 단계의 대규모 언어 모델에서 생성된 문장을 다음 모델에 프롬프트와 함께 입력하여, 앞서 생성한 문장들을 유사한 의미의 다른 문장으로 출력하는 구조이다. 이러한 방법은 프롬프트 입력으로 한번 생성한 문장보다 문법적 구조나 어휘 등을 다양하게 사용하여 문장을 생성한다. 이때, 다수의 대규모 언어 모델을 거치면서 생성된 문장이 입력 문장의 라벨을 벗어나게 되면 모델의 학습에 부정적 영향을 주게 되므로 키워드와 같은 제어 조건은 모든 문장에서 반영할 수 있게 구성하였다.

4. 실험 및 실험 결과

4.1 데이터셋 구성

본 연구에서 사용하는 Weather 데이터셋은 날씨에 대한 내용을 주제로 하는 문장들로 이루어진 자체 제작된(In-house) 데이터셋이다. 각각의 데이터는 문맥와 문장을 구성하는 주요 키워드, 그리고 날씨를 나타내는 14개의 라벨로 구성된 한국어 데이터셋으로, 전체 데이터의 크기는 13,000개이며, 훈련 데이터 6,993개(60%)와 검증 데이터 3,009개(20%), 평가데이터 2,998개(20%)로 구성되어있다. 실험에서는 증강된 데이터가 원본 데이터셋과 의미적 유사성을 가지고 있음을 정량적으로 비교하기 위해, 다중 클래스 분류에서 원본 데이터를 사용하였을 때와 유사한 성능을 가진다는 점을 보여주고자 한다. 이를 위해, 훈련 데이터의 크기에 따른 모델 훈련 성능을 비교하고자 훈련 및 검증 데이터의 일부만 사용하여 표 1와 같이 데이터를 구성하였다.

4.2 실험 방법

Weather 데이터셋에서 추출한 데이터를 단일 대규모 언어 모델, 병렬 및 순차 구조의 다중 대규모 언어 모델에 제어 조건과 함께 입력하여 유사의미 문장을 생성한다. 생성된 문장으로 훈련 데이터를 증강하고, 최대 길이(Max length) 64, batch size 64, learning rate 5e-5, AdamW 최적함수(Optimizer)를 사용하여 KcBERT [6]를 파인튜닝한다. 생성된 문장의 의미 유사도

표 2. 다중 클래스 분류 성능평가 결과. N_s는 데이터 증강에 사용된 샘플 데이터를 의미하며, N_t는 증강 데이터(샘플 데이터 + 생성 문장)을 의미한다. 샘플 데이터 400개를 사용하여 병렬 구조의 복수 대규모 언어 모델로 데이터를 증강하였을 때, 샘플 데이터셋을 사용하여 학습한 N_t=400의 정확도보다 0.001 높은 성능이 나왔다.

데이터 종류	모델 종류	학습 데이터 크기							
		N_t=100	N_t=200	N_t=400	N_t=600	N_t=800	N_t=1200	N_t=1600	N_t=2000
샘플링된 데이터셋		0.770	0.938	0.979	0.988	0.991	0.995	0.997	0.996
증강에 사용된 데이터셋									
N_s=100	GPT-3.5-turbo	-	0.773	0.813	0.866	0.864	0.865	0.887	0.856
	GPT-3.5-turbo-16K	-	0.701	0.793	0.857	0.852	0.835	0.856	0.811
	GPT-4	-	0.730	0.857	0.866	0.871	0.884	0.876	0.845
	Parallel Structure	-	0.727	0.879	0.886	0.894	0.887	**0.899**	0.874
	Sequential structure	-	0.675	0.820	0.825	0.838	0.851	0.838	0.850
N_s=200	GPT-3.5-turbo	-	-	0.884	0.917	0.947	0.925	0.917	0.912
	GPT-3.5-turbo-16K	-	-	0.845	0.937	0.932	0.919	0.917	0.891
	GPT-4	-	-	0.833	**0.968**	0.963	0.958	0.951	**0.968**
	Parallel Structure	-	-	0.864	0.955	0.954	0.963	0.951	0.923
	Sequential structure	-	-	0.876	0.929	0.924	0.944	0.949	0.938
N_s=400	GPT-3.5-turbo	-	-	-	-	0.943	0.943	0.940	0.932
	GPT-3.5-turbo-16K	-	-	-	-	0.970	0.971	0.974	0.968
	GPT-4	-	-	-	-	0.964	0.972	0.981	0.975
	Parallel Structure	-	-	-	-	**0.980**	**0.980**	0.966	0.970
	Sequential structure	-	-	-	-	0.973	0.935	0.960	0.958

와 학습데이터로 활용 가능성을 평가하기 위해 원본 데이터와 증강된 데이터에 대해 각각 문장 분류(Sentence Classification)을 수행하고 평가 지표인 정확도(Accuracy)를 비교하였다.

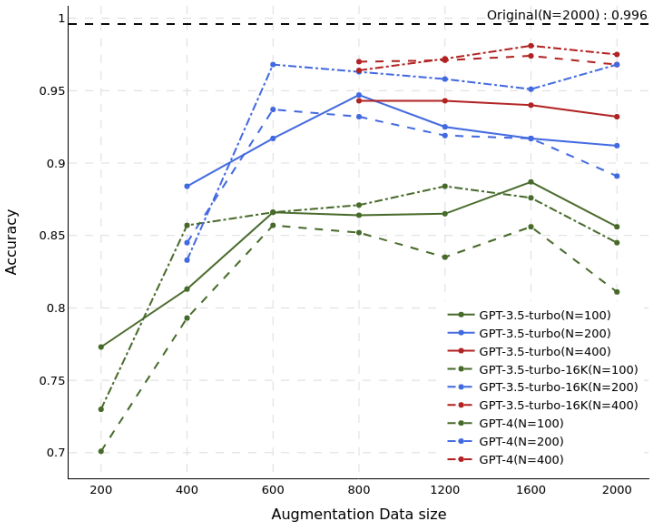
4.3 다중 클래스 분류 실험

대규모 언어 모델로 생성한 유사의미 문장으로 훈련 데이터를 증강하였을 때, 대규모 언어 모델에 입력하는 데이터의 수나 모델 구조에 관계없이 훈련 모델의 성능이 증가하는 것을 볼 수 있다. 실험에서는 원본 데이터를 사용한 다중 클래스 분류에서 정확도가 0.997로 최고 성능을 보였으나, N_s=400으로 했을 때, 병렬 구조의 다중 대규모 언어 모델로 증강한 데이터셋을 사용하였을 때, 0.980의 성능을 보이며 원본 데이터셋과 성능차이가 0.017로 거의 유사함을 보였다. 그러나 증강에 사용되는 원본 데이터의 수가 400보다 적은 경우 데이터 증강으로 높일 수 있는 최고 성능도 같이 낮아졌으며, 생성하는 문장의 수를 200보다 증가시켜도 평가 모델의 성능은 비슷하거나 낮아지는 모습을 보였다. 그 이유는 표3의 예시들을 통해서 유추할 수 있다.

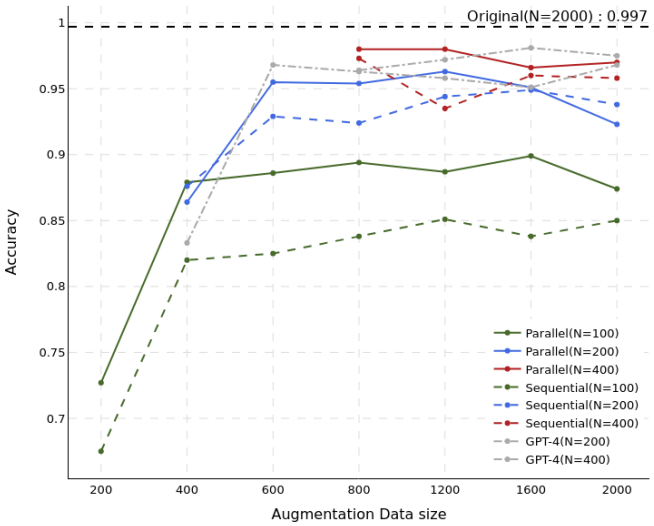
원본 데이터에서는 동일한 라벨을 가지는 문장들의 단어 구성과 문장 구조가 다양하지만 대규모 언어 모델에 의해 생성된 문장들은 키워드를 포함하는 조건에 맞추어 문장을 생성하기에 데이터가 단조롭고 모델이 다양한 문장을 훈련하기 어렵다. 문장의 다양성을 위해 키워드 포함 조건을 제외하게 되면 표 4와 같이 라벨에 적합하지 않은 문장을 생성하게 되어 잘못된 라벨 데이터 학습으로 모델의 성능을 저하시키게 된다. 프롬프트에 키워드 사용 유무에 따른 성능 차이는 표5에서 볼 수 있다.

5. 결론

대규모 언어 모델을 사용하여 유사의미 문장 생성을 통해 학습 데이터를 증강시키는 방법이 유효함과, 단일 모델을 사용하기보다 여러 모델을 사용하였을 때, 더 유효한 문장을 생성할 수 있음을 확인하였다. 또한, 프롬프트에 입력 문장과 함께 키워드와 같은 문장생성 기준을 입력하는 것으로 더 학습에 적합한 데이터를 생성할 수 있음을 보았다. 본 연구에서는 프롬프트의 구성에 있어, 입력 문장만 사용하는 One-Shot 방식을 사용하였다. 예시 문장과 키워드 정보의 입력으로 문장 구조와 어휘를 제한하게 된다는 점을 정성 분석으로 알 수 있었으며, 키워드 외에도 동일한 라벨을 가지는 학습 데이터의 문장을 예시로 제공하는 Few-Shot 방식을 사용하여 원본 데이터에 가까운 데이터 증강을 수행할 수 있을 것으로 생각된다. 본 연구를 통해 대규모 언어 모델을 사용한 데이터 증강 방식은 짧은 문장에 대한 분류에서는 사람에 의해 구축된 학습데이터에 준하는 성능을



(a) 단일 대규모 언어 모델 증강에서 성능 그래프



(b) 다중 대규모 언어 모델 증강에서 성능 그래프

그림 4. 다중 클래스 분류 성능평가 그래프

표 3. 유사의미 문장 생성 예시(키워드가 있는 경우)

입력 프롬프트 "오늘 동대문에 눈 와?"와 유사한 의미의 문장을 키워드를 포함해서 5개 만들어줘. **키워드: 오늘, 동대문**
생성 문장 1. 오늘 동대문에 눈이 올까? 2. 오늘 동대문에 눈이 왔어요? 3. 오늘 동대문에 눈이 올까요? 4. 오늘 동대문에 눈이 오나요? 5. 오늘은 동대문에 눈이 왔나요?
동일 라벨의 원본 데이터 예시 1. 내일 도봉에 눈 오니? 2. 남산타워에는 눈 안 오지? 3. 오늘 눈이 올지 안올지 좀 알아봐 줘. 4. 오후에 눈 올 거 같니? 5. 서울 갈 건데 눈 소식 있어?

보았으며, GLUE Task의 Question-answering NLI(QNLI)[7]와 같은 문단 구조의 데이터셋에서도 유효한 데이터 증강을 수행할 수 있는 프레임워크를 제안할 계획이다.

표 4. 유사의미 문장 생성 예시 (키워드가 없는 경우)

입력 프롬프트 "서울에 눈 오고 있나요?"와 유사한 의미의 문장을 5개 만들어줘.
생성 문장 1. 서울은 눈이 오고 있나요? 2. 현재 서울에는 눈이 내리고 있나요? 3. 지금 서울에는 눈이 내리고 있나요? **4. 서울에 비가 오고 있나요?** 5. 혹시 서울에 눈이 내리고 있는지 알고 있나요?

표 5. 키워드 사용 유무에 따른 다중 클래스 분류 성능 비교

키워드 사용 여부	**모델 종류**	**학습 데이터 크기** N=800	N=1200
Control	Parallel	**0.980**	**0.980**
	Sequential	0.973	0.935
Non-Control	Parallel	0.953	0.963
	Sequential	0.958	0.925

감사의 글

이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업이며 (2022R1F1A1071047), 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155857, 인공지능융합혁신인재양성 (충남대학교))

참고문헌

[1] H. Seo, S. Jung, and J. Jung, "Semantic and syntax paraphrase text generation," *Annual Conference on Human and Language Technology*, pp. 162–166, 2020.

[2] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," 2023.

[3] T. Hosking, H. Tang, and M. Lapata, "Hierarchical sketch induction for paraphrase generation," 2022.

[4] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," 2019.

[5] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," 2023.

[6] J. Lee, "Kcbert: Korean comments bert," *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pp. 437–440, 2020.

[7] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," 2019.