Large Language Model을 활용한 키워드 기반 대화 생성

이주환○, 허탁성, 김지수, 정민수, 이경욱, 김경선

엔에이치엔다이퀘스트

{9521ljh, gjxkrtjd221, jisukim8873, tmzkdlwm1125}@gmail.com, {arp1710, kksun}@diquest.com

Keyword Based Conversation Generation using Large Language Model

Juhwan Lee○, Tak-Sung Heo, Jisu Kim, Minsu Jeong, Kyounguk Lee, Kyungsun Kim

NHN Diquest

요 약

자연어 처리 분야에서 데이터의 중요성이 더욱 강조되고 있으며, 특히 리소스가 부족한 도메인에서 데이터 부족 문제를 극복하는 방법으로 데이터 증강이 큰 주목을 받고 있다. 이 연구는 대규모 언어 모델(Large Language Model, LLM)을 활용한 키워드 기반 데이터 증강 방법을 제안하고자 한다. 구체적으로 한국어에 특화된 LLM을 활용하여 주어진 키워드를 기반으로 특정 주제에 관한 대화 내용을 생성하고, 이를 통해 대화 주제를 분류하는 분류 모델의 성능 향상을 입증했다. 이 연구 결과는 LLM을 활용한 데이터 증강의 유의미성을 입증하며, 리소스가 부족한 상황에서도 이를 활용할 수 있는 방법을 제시한다.

주제어: 데이터 증강, 대규모 언어 모델(LLM), 키워드 추출

1. 서론

자연어 처리 기술은 현재 빠른 속도로 발전하고 있으나, 리소스 부족 문제는 여전히 남아있는 과제 중 하나다. 특히, 한국어 데이터 셋의 경우 다양한 공개 데이터[1]가 있음에도 불구하고 원하는 도메인에 특화된 데이터를 얻는 것은 여전히 어려운 과제로 남아 있다. 이러한 데이터 부족으로 인해 데이터 증강 기술은 점차 중요성을 더하고 있으며, 이 기술의 주요 목표는 모델의 안정적인 학습과 동시에 모델의 성능을 향상시키는 것이다. 따라서 데이터 증강 기술은 리소스 부족 문제를 극복하고, 모델의 성능을 향상시킬 수 있는 도구로 활용된다.

OpenAI[2]의 GPT 계열의 LLM 등장은 텍스트 생성 분야에서 혁신적인 기회를 제공하고 있으며, 작업자들의 부담을 크게 줄여주고 있다. 이러한 LLM은 방대한 양의 텍스트 데이터를 학습하여 놀라운 성과를 거두었다. 더불어, InstructGPT[3]와 같은 모델은 지시사항과 질문-답변 형식의 데이터로 훈련을 받아 질의응답 작업에서 더 능숙하게 수행한다. 또한, 생성된 텍스트를 조정하고 올바른 답변을 생성하기 위해 ChatGPT와 같은 모델은 강화학습 기술을 채택하고 있다. 이로써 모델은 입력에 대해 더 많은 정보를 고려하고 사용자들에게 유익한 응답을 제공할 수 있게 된다. 이러한 모델들은 자연어 생성 작업에서 혁신을 이루어내고 있으며, 다양한 분야에서 중요한 도구로 활용될 것으로 예상된다.

우리는 이러한 LLM의 효과를 활용하여 데이터 증강 기술을 연구하고자 한다. 이미 스탠퍼드의 Alpaca[4] 프로젝트에서 LLM을 활용하여 모델을 미세조정하여 성능을 향상시킨 연구 사례가 있다. 이 사례는 LLM을 활용한 데이터 증강이 실제로 유의미한 결과를 가져올 수 있음을 입증했다. 상용화된 LLM들은 우수한 성능을 보이지만, 데이터 생성 측면에서는 제한이 있다. 예를 들어 금융 데이터나 혐오 발언과 관련된 내용을 생성하지 않도록 훈련되어 있다는 점이다. 이런 제한은 개발자의 의도대로 모델이 작동하도록 보장하기 위한 것이지만, 때로는 특정 도메인에서 필요한 데이터를 얻는 데 제약을 줄 수 있다. 따라서 우리는 한국어에 특화된 LLM인 Polyglot-12.8B[5]를 미세조정하여 원하는 형태의 출력을 생성할 수 있도록 하고자 한다. 또한, LLM은 확률적으로 높은 문장을 생성하기 때문에 동일한 결과가 반복될 수 있다. 우리는 이를 극복하기 위해 키워드 추출 모델을 통해 얻은 키워드를 LLM의 생성 조건으로 사용한다. 이를 통해 원하는 도메인에서 효과적이며 다양한 데이터를 증강하고 활용할 수 있을 것으로 기대한다.

2. 관련 연구

2.1 데이터 증강

자연어 처리 분야에서 데이터 증강 기술은 크게 두 가지 형태로 구분된다. Rule-based(규칙 기반) 방법과 model-based(모델 기반) 방법이 있다.

2.1.1 Rule-Based

Rule-Based 데이터 증강 기술은 모델 구성 없이 단순한 텍스트 변형을 통해 데이터를 증강한다. 이 중에서 EDA

(Easy Data Augmentation)[6]는 텍스트 증강의 가장 잘 알려진 방법 중 하나로, 다양한 연구에서 활용되어 왔으며, 주로 SR (Synonym Replacement, 동의어 대체), RI (Random Insertion, 임의 삽입), RS (Random Swap, 임의 교환), RD (Random Deletion, 임의 삭제)와 같은 다양한 변형 기법으로 구성된다.

2.1.2 Model-based

자연어 처리 분야에서 텍스트 생성에 널리 사용되는 언어 모델로는 Seq2seq 모델, 그리고 GPT 및 T5와 같은 모델이 주목을 받고 있다. 이러한 접근법들을 활용한 데이터 증강은 모델 성능 향상에 대한 성공적인 실험 결과를 입증했다[7-9]. 특히, Back-Translation 접근법은 번역 후 역 번역하는 방법으로 리소스가 제한된 도메인에서 자주 활용되는 방법론 중 하나로 꼽힌다. 최근에는 GPT 계열의 모델을 활용하여 LLM을 위한 데이터 생성 연구가 급속히 증가하고 있다. 대표적으로 스탠퍼드 대학에서 진행한 Alpaca 프로젝트는 OpenAI의 text-davinci-003 모델을 활용하여 데이터를 생성하여 이를 Llama[10] 모델에 학습시켜 우수한 성능을 보여주었다. 이는 대규모 언어 모델을 활용한 데이터 증강 기술의 효과를 확증한 사례 중 하나다.

2.2 대규모 언어 모델(LLM) 텍스트 생성 조정

텍스트 생성을 올바르게 하기 위해서는 원하는 형태의 출력물이 생성되어야 한다. 이러한 텍스트 생성을 조정하기 위해 다양한 방법[11]들이 존재한다. 이 중에서 구글 리서치팀에서 제시한 Instruct Tuning[12] 방법론이 많은 연구에서 활용되고 우수한 성능을 보이고 있다. 따라서 우리는 LLM의 출력물을 조정하기 위해 Instruct Tuning방법론을 사용한다. 또한, LIMA 논문[13]에 따르면 대규모 언어 모델의 지식적 능력은 대부분 사전 학습 단계에서 습득되며, 고품질의 텍스트 생성을 위해서는 1000개의 데이터만으로도 모델 조정이 가능하다는 주장을 제시하고 있다. 따라서 우리는 이러한 연구들을 바탕으로 연구를 진행하고자 한다.

3. 방법론

우리는 이번 연구에서는 한국어에 특화된 LLM인 Polyglot-12.8B를 활용하여 데이터 증강의 효과를 검증하고자 한다. 연구의 주요 대상은 대화 데이터이며, 특정 주제와 키워드에 기반해서 대화 내용을 생성하고 효과를 입증하는 데 초점을 두고 있다. 키워드 추출은 결과의 다양성을 확보하는 데 중요하다. 예를 들어, LLM 모델에게 동일한 입력으로 대화 생성을 요청하면 언어 모델은 확률적으로 높은 결과 값을 줄 가능성이 있어 동일한 결과가 반복될 수 있다. 따라서 우리는 이러한 모델 내 문제를 극복하기 위해 키워드를 생성 조건으로 활용하여 결과의 다양성을 증진시키는 방법을 고안했다. 이렇게 증강된 데이터의 유의미성을 입증하기 위해서 본 연구는 대화 주제를 분류하는 모델을 통해서 증강된 데이터가 분류 모델의 성능 향상을 가져오는지 확인하고자 한다.

3.1 키워드 추출 모델을 위한 데이터 셋

우리는 키워드 추출 모델을 위해 AIHub[14]에서 공개한 텍스트 내 개체명과 키워드가 라벨링되어 있는 “민원 업무 자동화 인공지능 언어 데이터”를 이용했다. 이 데이터 셋은 총 900,000개의 데이터가 있으며, Train 데이터 셋이 800,000개, validation 데이터 셋이 100,000개로 이루어져 있다. 우리는 실험을 위해 기존 Train 데이터 셋을 9:1로 나누어 Train, Validation 데이터 셋을 재정의하였으며, 기존 Validation 데이터 셋을 Test 데이터셋으로 재정의하였다.

3.2 데이터 증강을 위한 데이터 셋

우리는 AIHub[14]에서 공개한 대화 데이터 셋인 “주제별 텍스트 일상대화 데이터”를 이용했다. 이 데이터셋은 총 109,614개의 대화 데이터가 있으며 이 중 90%는 Train 데이터셋으로, 나머지 10%는 Validation 데이터셋으로 구성되어있다. 대화 주제는 총 20가지이며 각 주제별로 데이터 분포는 대략적으로 5%씩 차지한다. 이 데이터 셋은 일상적인 대화 내용이기에 은어, 맞춤법 오류들이 포함되어있다.

3.3 데이터 증강을 위한 데이터 셋 분리

우리는 LLM과 분류 모델을 학습하기 위해 데이터 셋을 그림1과 같이 분리했다. 그림1에서 데이터 셋에 (LLM)으로 표시된 데이터는 LLM 모델을 학습하기 위한 데이터셋으로 사용되며 반면에 (CLS)로 표시된 데이터는 분류 모델을 위한 데이터셋으로 활용된다. 기존 Train 데이터를 9:1 비율로 분리해서 분류 모델을 위한 $\text{Validation}_{(CLS)}$ 데이터 셋을 만들었다. 기존 Validation 데이터 셋은 분류 모델을 위한 $\text{Test}_{(CLS)}$로 사용된다.

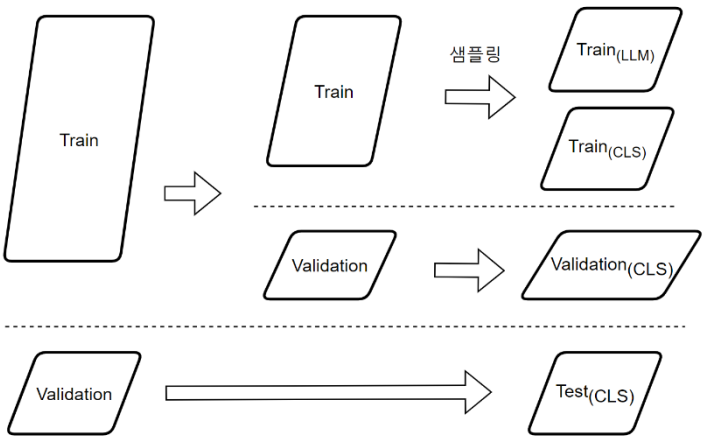

그림 1. 모델 별 데이터 셋 분리

3.4 데이터 증강을 위한 데이터 샘플링

우리의 목표는 리소스가 제한된 도메인에서 효과적인 데이터 증강을 달성하는 것이기 때문에 우리는 모든 학습 데이터를 사용하는 것이 아니라 적은 양의 데이터만을 활용하여 연구를 진행한다. 이를 위해, Train 데이터를 샘플링하여 데이터 증강의 효과를 입증하고자 한다. 또한 우리는 데이터 개수에 따른 성능을 비교하기 위해 총 다섯 가지의 샘플링을 고려했다. 각 샘플링은 500, 1000, 2000, 3000, 4000개의 데이터를 무작위로 Train 데이터 셋으로 부터 추출했다. 중요한 점은, 큰 크기의 데이터셋에는 보다 작은 샘플링 데이터셋이 포함된다. 예를 들어, 1000개의 샘플링 데이터셋에는 500개 샘플링 된 데이터셋이 포함되어 있다. 또한, 각각의 샘플링된 데이터 셋에는 20가지의 대화 주제가 균등하게 분포되어 있다. 이 샘플링 방법을 분류 모델을 위한 $Train_{(CLS)}$ 데이터셋에도 동일하게 적용했다. 즉, $Train_{(LLM)}$ 데이터 개수와 $Train_{(CLS)}$ 데이터 개수는 동일하다.

3.5 키워드 추출

우리는 대화를 효과적으로 증강하기 위해 명사 기반 키워드 추출을 수행하고자 한다. 이를 위해 상호 보완적인 방법을 사용한다. 원천 데이터셋에서 제공된 발화 별 태깅 정보와 직접 개발한 키워드 추출 모델을 활용하여 두 정보의 교집합을 통해 키워드를 선별한다.

우리가 구축한 키워드 추출 모델은 개체명 인식 모델과 같은 방법으로, BIO 태깅을 통해 키워드를 추출하는 방법으로 진행된다. 이때 사용된 개체명 클래스는 키워드만 존재하기 때문에, KB(Keyword Begin), KI(Keyword Inside) O(Outside)의 3개로 구성된다.

키워드 추출 모델은 키워드만을 라벨로 간주하여 사용되어야 하지만, 사용하는 데이터 셋의 개체명에도 여러 키워드가 포함되어 있는 경우가 많다. 그러므로, 키워드 뿐만 아니라 개체명 또한 키워드로 간주하여 실험을 진행하였으며, 최종적으로, 키워드 그리고, 키워드와 개체명을 함께 사용한 2가지 키워드 추출 모델을 구축했다.

키워드 추출 모델은 한국어 데이터 셋을 사전 학습한 KLUE-BERT-base 사전 학습 언어 모델을 사용하였다[15]. KLUE-BERT-base를 통해 추출된 모든 토큰들은 키워드 개체명을 분리하는 역할을 하는 분류기를 추가함으로써, 키워드 개체명을 예측하게 된다. 모델의 미세조정을 위한 하이퍼파라미터는 표1과 같이 설정하여 학습을 진행했다.

표 1. 키워드 분류 모델 학습 하이퍼파라미터

키워드 분류 모델	하이퍼파라미터
KLUE-BERT-base	epoch=100, batch size=16, learning rate=5e-6 early-stop(patience: 3)

3.6 대규모 언어 모델(LLM) 학습 데이터 템플릿

Polyglot-12.8B 모델을 활용하여 미세조정을 수행하고자 한다. 우리는 Instruct Tuning을 위해 질문과 답변의 형태로 $Train_{(LLM)}$을 템플릿에 맞게 전처리를 수행해야한다. 전처리에 대한 과정은 그림2에서 볼 수 있듯이 우리는 대화 주제와 키워드를 활용해서 질문을 구성하고 대화 내용을 대답으로 구성했다.

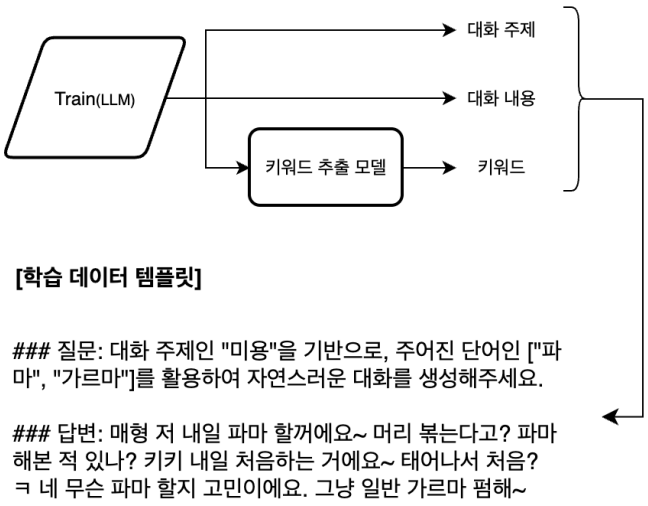

그림 2. LLM 학습용 데이터 템플릿 구성

3.7 대규모 언어 모델(LLM) 미세조정

우리는 LLM의 미세조정을 위해 QLoRA[16]를 활용하여 학습을 수행했다. QLoRA를 사용함으로써, 모델의 모든 파라미터를 조정하는 대신 별도의 어댑터를 활용하여 학습을 진행했다. 우리는 Hugging Face의 PEFT 모듈을 활용하여 QLoRA를 적용하였으며, 주요 하이퍼파라미터는 다음 표2와 같다.

표 2. LLM 학습 하이퍼 파라미터

대규모 언어 모델	하이퍼 파라미터
Polyglot-12.8B	epochs=10, batch size=16, learning rate=3e-4
QLoRA	r=8, lora_alpha=16, lora_dropout=0.05

3.8 대규모 언어 모델(LLM)을 활용한 대화 데이터 증강

증강할 데이터셋은 $Train_{(CLS)}$ 데이터 셋이다. 이 데이터셋의 주제와 키워드를 활용하여 질문을 형성해서 LLM을 통해 새로운 대화 데이터 셋을 생성하여 $Train_{(CLS)-Augmented}$를 구축했다. 우리는 LLM 추론 과정에서 설정한 중요 하이퍼파라미터는 다음 표3과 같다.

표 3. LLM 추론 하이퍼 파라미터

대규모 언어 모델	하이퍼 파라미터
Polyglot-12.8B	top_p=0.5, max_new_tokens=256, early_stopping=True, do_sample=True, no_repeat_ngram_size=2

3.9 대화 분류 모델

분류 모델은 한국어 데이터 셋을 사전 학습한 KLUE-BERT-base 사전 학습 언어 모델을 사용했다 [15]. 이 모델을 사용하여 데이터의 효과를 입증하기 위해 세 가지 실험 시나리오를 구성했다. 첫 번째 시나리오에서는 $Train_{(CLS)}$만을 사용하여 모델을 학습했다. 두 번째 시나리오에서는 $Train_{(CLS-Augmented)}$만을 사용하여 모델을 학습했다. 마지막으로, 세 번째 시나리오에서는 $Train_{(CLS)}$와 $Train_{(CLS-Augmented)}$ 모두 사용하여 모델을 학습했다. 만약 세 번째 시나리오의 결과가 첫번째 시나리오 보다 더 우수한 성능을 보인다면, 증강된 데이터셋의 유의미성을 입증할 수 있다.

추가로 세 가지 시나리오에서 $Validation_{(CLS)}$과 $Test_{(CLS)}$는 모두 동일하다. 모델의 미세조정을 위한 하이퍼파라미터는 표4와 같이 설정하여 학습을 진행했다.

표 4. 대화 분류 모델 학습 하이퍼 파라미터

대화 분류 모델	하이퍼 파라미터
KLUE-BERT-base	Epoch=100, batch size=16, learning rate=5e-6 early-stop(patience: 3)

4. 실험 결과

4.1 키워드 추출 결과

4.1.1 키워드 추출 모델 결과

키워드 추출 모델은 키워드만을 사용한 키워드 추출 모델과 키워드와 개체명을 함께 사용한 키워드 추출 모델로 구성된다. 표5는 키워드 추출 모델에 대한 실험 결과이다. 최종적으로, 데이터 증강에 사용되는 키워드 추출 모델은 키워드와 개체명을 함께 사용한 키워드 추출 모델을 사용하였다.

표 5. 분류 모델 시나리오 별 성능 평가

키워드 추출 모델	F1-score
Keyword	0.7853
Keyword + Entity	0.7998

4.1.2 키워드 추출 모델에 따른 증강 데이터 통계

3.4에서 언급한대로, 큰 크기의 데이터셋에는 보다 작은 샘플링 데이터셋이 포함된다. 이로 인해 각 샘플링 데이터셋마다 비슷한 Box Plot이 표현된다. 아래 그림 3에서 볼 수 있듯이, 추출된 키워드 개수의 평균은 일반적으로 8개이다.

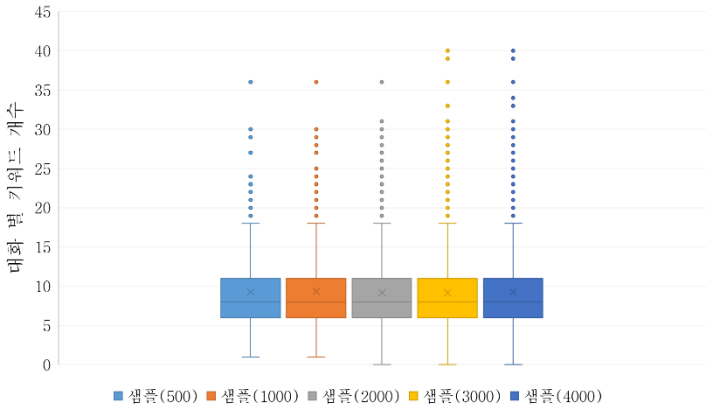

그림 3. 키워드 개수 Boxplot

4.2 대규모 언어 모델(LLM) 학습과정

우리는 다섯 가지 다른 샘플링 시나리오로 모델 학습을 진행했다. 다양한 시나리오에서의 모델 학습 결과를 확인하기 위해 그림4는 각 시나리오별로 학습 epoch에 따른 손실 값을 나타낸 그래프다. 이 그래프를 통해 우리는 모델이 안정적으로 학습을 진행한 것을 확인할 수 있다.

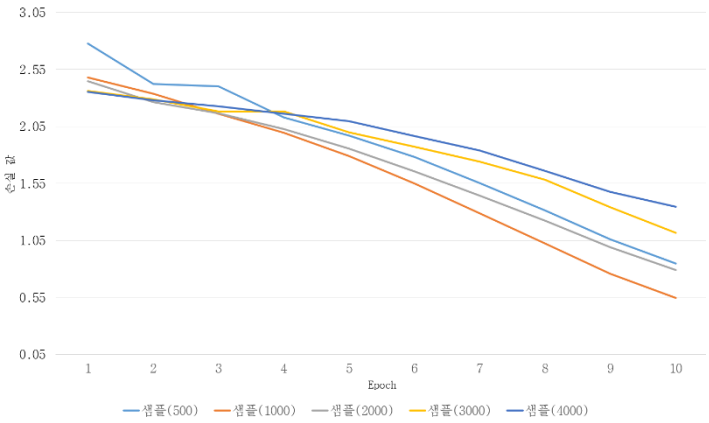

그림 4. Polyglot12.8B 미세조정 Epoch별 손실 값

4.3 대화 분류 모델 성능 평가

증강된 데이터의 유의미성을 평가하기 위해 대화 분류 모델의 성능을 비교한다. 아래의 표6를 보면 KLUE-BERT-base 모델을 세 가지 시나리오 별 f1 score를 평가 했으며 굵게 표시된 글씨가 가장 우수한 성능을 나타낸다. 기존 데이터에 증강된 데이터를 합친 시나리오가 모든 샘플링 방식에 가장 뛰어난 성능을 보였다.

표 6. 대화 분류 모델 시나리오 별 성능 비교

샘플 / 학습데이터	샘플 (500)	샘플 (1000)	샘플 (2000)	샘플 (3000)	샘플 (4000)
$Train_{(CLS)}$	0.8211	0.8372	0.8596	0.8686	0.8760
$Train_{(CLS)-Augmented}$	0.8141	0.8373	0.8003	0.8194	0.8298
$Train_{(CLS)}$ + $Train_{(CLS)-Augmented}$	0.8275	0.8561	0.8636	0.8723	0.8814

5. 결론

해당 연구에서는 LLM을 활용하여 데이터 증강의 유의미성을 분류 모델을 통해 확인했다. 실험 결과를 통해 키워드 기반 대화 내용 생성을 통해 효과적인 데이터 증강 방법을 제시했으며, 특히 우리는 원하는 결과 값을 얻기 위해 소량의 데이터셋을 학습시켜 성공적인 데이터 증강을 수행한 점에 중요한 의미를 부여하고 있다.

뿐만 아니라, 연구를 진행하면서 LLM의 생성 결과의 한계점을 발견했다. LLM의 미세 조정을 통해 결과 값을 조정하려 시도하였으나, 모든 결과 값을 원하는 형태로 얻을 수는 없었다. 몇몇 경우에는 대화 내용이 전혀 생성되지 않거나, 영어로 문장을 생성하는 경우도 있었다. 또한 대다수의 문장들이 도중에 생성되다가 끊기게 된다. 이러한 문제는 LLM 모델을 추론 때 설정한 하이퍼파라미터 중에 max_new_tokens 매개변수를 256으로 설정한 것으로 확인되었다. 종합적으로 이러한 점들을 고려했을 때 높은 품질의 증강 데이터를 얻기 위해서는 후처리 작업이 필요하다. 따라서 연구 결과를 토대로 데이터 증강 품질 향상과 한계 극복에 관한 추가 연구를 고려할 필요가 있다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (연구개발 제번호 : RS-2023-00225661, 디지털 증거의 증명력 제고를 위한 인과관계 추론 및 표현 기술 개발)

참고문헌

[1] Ban, B. (2022, October). A survey on awesome korean nlp datasets. In 2022 13th International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1615-1620). IEEE.

[2] OpenAI. "ChatGPT." ChatGPT - OpenAI, https://platform.openai.com/docs/guides/chat.
[3] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems 35 (2022): 27730-27744.
[4] Taori, Rohan, et al. "Stanford Alpaca: An Instruction-following LLaMA model." GitHub, 2023, https://github.com/tatsu-lab/stanford_alpaca.
[5] Ko, Hyunwoong, et al. "A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models." arXiv preprint arXiv:2306.02254 (2023).
[6] Wei, Jason, and Kai Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." arXiv preprint arXiv:1901.11196 (2019).
[7] Kumar, Varun, Ashutosh Choudhary, and Eunah Cho. "Data augmentation using pre-trained transformer models." arXiv preprint arXiv:2003.02245 (2020).
[8] Hou, Yutai, et al. "Sequence-to-sequence data augmentation for dialogue language understanding." arXiv preprint arXiv:1807.01554 (2018).
[9] Sugiyama, Amane, and Naoki Yoshinaga. "Data augmentation using back-translation for context-aware neural machine translation." Proceedings of the fourth workshop on discourse in machine translation (DiscoMT 2019). 2019.
[10] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).
[11] Zhang, Hanqing, et al. "A survey of controllable text generation using transformer-based pre-trained language models." arXiv preprint arXiv:2201.05337 (2022).
[12] Wei, Jason, et al. "Finetuned language models are zero-shot learners." arXiv preprint arXiv:2109.01652 (2021).
[13] Zhou, Chunting, et al. "Lima: Less is more for alignment." arXiv preprint arXiv:2305.11206 (2023).
[14] "AIHUB." AIHUB Data, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=543.
[15] Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." arXiv preprint arXiv:2105.09680 (2021).
[16] Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." arXiv preprint arXiv:2305.14314 (2023).