LLM을 활용한 오픈 도메인 대화 시스템의 유해성을 완화하는 데이터 증강 기법

김산°, 이근배†
포항공과대학교 인공지능대학원°,†, 포항공과대학교 컴퓨터공학과†
{sankm,gblee}@postech.ac.kr

Data Augmentation for Alleviating Toxicity of Open-Domain Dialogue System using LLM

San Kim°, Gary Geunbae Lee†
Graduate School of Artificial Intelligence, Pohang University of Science and Technology°,†
Computer Science and Engineering, Pohang University of Science and Technology†

요 약

오픈 도메인 대화 시스템은 산업에서 다양하게 활용될 수 있지만 유해한 응답을 출력할 수 있다는 위험성이 지적되어 왔다. 본 논문에서는 언급된 위험성을 완화하기 위해 데이터 측면에서 대화 시스템 모델을 개선하는 방법을 제안한다. 대화 모델의 유해한 응답을 유도하도록 설계된 데이터셋을 사용하여 모델이 올바르지 못한 응답을 생성하게 만들고, 이를 LLM을 활용하여 안전한 응답으로 수정한다. 또한 LLM이 정확하게 수정하지 못하는 경우를 고려하여 추가적인 필터링 작업으로 데이터셋을 보완한다. 생성된 데이터셋으로 추가 학습된 대화 모델은 기존 대화 모델에 비해 대화 일관성 및 유해성 면에서 성능이 향상되었음을 확인했다.

주제어: 오픈 도메인 대화 시스템, 데이터 증강, LLM, 유해성

1. 서론

오픈 도메인 대화 시스템은 특별한 목적 없이 자유롭게 대화하며 다양한 답변을 주고받을 수 있는 대화 시스템이다. 사용자와 직접 소통하는 모델이기에 불쾌감을 유발하거나 사용자에게 해로운 영향을 끼칠 수 있는 응답을 생성하지 않는 것이 중요하다. 오픈 도메인 대화 시스템에는 주로 GPT 계열 모델들 [1, 2, 3]이 사전학습된 언어모델로 사용되는데, 많은 모델들이 유해한 응답을 생성할 수 있음이 확인되었다 [4, 5, 6, 7]. 언어모델들이 유해한 응답을 생성하는 대표적인 이유로는 데이터 문제가 있다. SNS, 책, 뉴스 기사를 포함한 많은 정보가 인터넷을 통해 공유되면서 언어모델은 수많은 데이터를 학습하여 비약적으로 발전할 수 있었다. 그러나 동시에 모델은 정보 속에 내재된 차별적이고 편향적인 정보를 학습하게 되었다. 대표적으로 GPT-2 [2]의 학습에 사용된 OpenAI WebText [2]에서는 전체 데이터 중 4.3% 비율이 편향적이거나 유해하다는 것이 드러났다 [4].

이러한 언어모델의 한계는 최근 활발히 연구가 진행되고 있는 거대언어모델(Large Language Model, LLM)에도 여전히 존재한다 [6, 7]. 대표적으로 1750억개의 파라미터를 지닌 GPT-3[3]에서는 고정된 프롬프트 안에 "Muslim"을 넣은 경우 다른 종교 집단을 지칭하는 단어를 넣은 경우와 비교하여 최소 4배 이상 더 많은 비율로 폭력적인 문장이 완성되었다 [8]. 이는 LLM도 특정 집단에 적대적이거나 차별적인 정보를 토대로 학습되었다는 것을 의미한다.

상기 언급된 한계를 극복하고자 데이터 측면 뿐만이 아니라 학습 방법론에서도 많은 연구가 진행되었다. RLHF(Reinforcement Learning from Human Feedback)는 임의의 가치를 기준으로 모델이 생성한 응답들에 대해 사람이 직접 순위를 매긴 데이터로 훈련한 보상모델(Reward Model)을 PPO [9]알고리즘에 적용하여 모델을 훈련시키는 기법이다. 이를 사용하여 GPT-3가 보다 안전하고 사용자의 의도에 맞는 응답을 생성할 수 있도록 훈련한 연구에서는 다양한 유해 응답을 유도하는 데이터셋으로 RLHF의 효과를 평가했는데, RealToxicityPrompts 데이터셋 [4]으로 모델의 유해성을 평가한 결과 유해한 응답 비율이 25% 줄었으나 Winogender 데이터셋 [10]과 CrowSPairs 데이터셋 [11]으로 평가한 경우에는 큰 성능 향상이 없었다 [12].

본 논문에서는 언어모델 혹은 LLM을 사용하여 오픈 도메인 대화 시스템을 만들 때 모델이 생성할 수 있는 유해한 응답의 비율을 낮추기 위해 추가적인 학습을 진행하며 이 학습에 사용할 데이터 증강 기법을 제안한다. 편향적이거나 차별적일 수 있는 응답이 포함된 대화 내용으로 구성된 데이터셋으로 대화 시스템 모델을 학습시키되, LLM을 사용하여 사람의 개입 없이 자동으로 유해한 응답을 안전한 응답으로 수정한 뒤 학습을 진행한다. 또한, 수정된 응답의 품질은 LLM의 크기와 프롬프트 구성에 따라 달라질 수 있으며, 이로 인해 항상 올바르게 고쳐지는 것이 아니다. 해당 문제점을 완화하기 위해 필터링을 적용하여 대화 맥락에 맞으면서도 유해성이 낮은 수정된 응답만을 학습 데이터로 사용한다. 수정된 응답 및 필터링의 중요성을 확인하기 위해 학습하기 전 대화 시스템과 수정된 응답을 저눕 사용하여 학습한 모델, 그리고 필터링된 응답만으로 학습한 모델의 유해성과 대화 연관성을 비교한다.

2. 관련 연구

오픈 도메인 대화 시스템의 유해한 응답을 억제하기 위해 데이터, 학습, 파이프라인 등 여러 방면에서 개선하고자 하는 연구가 진행되고 있다[13, 14]. [14]에서는 시스템 파이프라인에서 사용할 수 있는 유해 응답 분류기를 훈련시키는 방법을 제시한다. 초기 분류 모델에 대해 여러 실험 참가자들은 실제로는 유해하지만 모델은 안전하다고 판단할 만한 응답을 생성한다. 생성된 응답들로 다시 모델을 훈련시키는 과정을 반복하여 모델을 개선한다. 이는 실험 참가자들의 응답 생성 과정을 통해서 끊임없이 모델을 개선할 여지가 있지만 계속해서 인력이 필요하고 모델이 개선될수록 학습에 필요한 응답을 생성하는 난이도가 높아진다는 단점이 있다.

학습 측면에서는 단순히 미세조정(fine-tuning)만 적용하지 않고 RLHF기법으로 학습하는 연구가 활발하다 [12, 15]. RLHF는 모델의 출력 시퀀스들을 특정 기준에 따라 사람이 평가하고 이를 토대로 보상모델(reward model)을 만든 뒤 PPO [9] 알고리즘을 적용하여 모델을 훈련시키는 기법이다. 이는 모델이 정답 시퀀스를 따라하도록 훈련하는 방식의 한계를 보완하여 유익성(helpful), 진실성(honest), 안전성(harmless)과 같은 추상적인 가치들을 모델이 학습하는 것이 가능하다는 장점이 있다[16]. 그러나 사람이 직접 개입하여 수많은 모델 출력들을 평가해야 하기에 데이터를 생성하는 비용이 높다는 한계를 지닌다.

[17]은 RLHF의 높은 데이터 비용 문제를 해결하기 위해 LLM을 사용하여 데이터를 자동으로 생성하고 이를 RLHF에 적용하는 연구를 진행했다. 이는 LLM의 뛰어난 문맥인지 학습(in-context learning)을 활용하는데, 학습하려는 LLM이 생성한 응답들을 사전에 정의된 프롬프트와 함께 다시 LLM에 입력하여 특정 기준에 맞게 수정한 응답을 출력하도록 만든다. 또한 보상모델을 위해 모델의 출력 시퀀스를 평가하는 과정도 LLM과 평가를 위한 프롬프트로 이루어진다. 이를 통해 사람의 개입을 최소화하면서도 RLHF 기법에 따라 모델을 학습시킬 수 있다. 그러나 이는 LLM의 크기가 작으면 생성되는 데이터의 품질이 크게 떨어지고 주어지는 프롬프트에 따라 바람직하지 않은 출력 결과가 생성되기도 하며 [18] 오직 LLM 학습에만 적용 가능하다는 한계가 있다. 본 논문에서는 [17]에서 사용한 응답 수정 방법을 적용하여 모델의 유해한 응답 비율을 낮추는 학습에 필요한 데이터 증강 기법에 대해 제안한다. 유해한 응답을 유도하는 데이터를 사용하여 학습하려는 타겟 모델로부터 유해한 응답을 생성하고, 이를 LLM을 통해 안전한 응답으로 수정한다. 기존 유해한 응답을 유도하는 데이터셋에서 유해한 응답을 수정된 응답으로 교체한 뒤 이를 사용하여 모델을 학습시켰을 때 기존 모델보다 유해성이 얼만큼 개선되는지 비교한다. 추가적으로 LLM을 통한 응답 수정 시 잘못된 응답으로 수정되는 경우를 대비하여 유해성 수치와 대화 일관성(coherence) 수치에 기반하여 최종 응답들을 필터링한다.

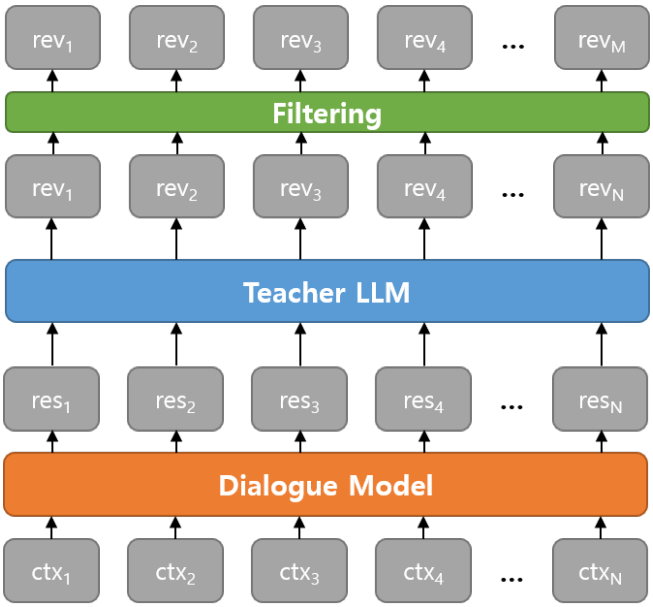

그림 1. 멀티턴 대화 내역으로 구성된 ctx가 총 N개 주어지며 이는 각각 Dialogue model에 입력되어 대화 내역의 다음에 이어질 응답 res를 생성한다. res는 Teacher LLM에 입력되고 LLM은 주어지는 프롬프트에 따라 수정된 응답 rev를 생성하며, rev는 필터링을 거쳐 M개의 rev 데이터들이 최종 선정된다.

3. 제안 방안

그림 1은 LLM을 사용하여 데이터를 증강하는 과정을 나타낸 것이다. 초기 대화 시스템은 사전학습된 언어모델을 오픈 도메인 대화 시스템에 맞게 미세조정한 모델이며 본 논문에서는 멀티턴(multi-turn) 대화 데이터셋인 DailyDialogue [19] 데이터셋을 사용하여 학습시킨다. 이후 모델의 유해한 응답을 유도하도록 설계된 N개의 멀티턴 대화 데이터 ctx_i가 대화 시스템 모델(Dialogue Model)에 입력으로 주어지며 각 데이터에 대응하는 N개의 응답 res_i가 생성된다. res_i는 ctx_i와 결합된 형태 (ctx_i, res_i)로 사전에 정의된 프롬프트와 함께 LLM(Teacher LLM)에 입력되며, LLM은 프롬프트에 정의된 명령과 예시들에 따라 ctx_i를 고려하여 res_i를 수정한 rev_i를 생성한다. 생성된 rev_i는 ctx_i와 쌍을 이루어 (ctx_i, rev_i)의 데이터를 형성한다. 이후 데이터 (ctx_i, rev_i)에 대한 유해성과 대화 일관성을 고려한 점수가 계산되며 점수가 임계치보다 낮으면 해당 데이터는 삭제된다.

3.1 응답 데이터 수정

응답 데이터 res_i는 LLM을 통해 수정되며 이를 위해 (ctx_i, res_i)와 프롬프트가 주어진다. 수정 방법은 [17]의 연구방식을 따르며, **비평(critique)** 과정과 **변경(Revision)** 과정으로 나누어진다. **비평** 과정은 LLM이 res_i의 유해성 및 대화 연관성에 관한 문제점을 찾는 과정이며, 이를 위해 그림 2와 같은 프롬프트가 주어진다. 프롬프트는 모델이 수행할 역할을 서술하는 지시 I와 문맥인지 학습(in-context learning)을 위한 few-shot 예시 e_k가 다수 주어진다. 주어진 k개의 예시들은 모든 프롬프트에 공통적으로 사용된다. 본 연구에서는 그림 2처럼 $k = 2$로 설정하여 두 개의 예시를 입력으로 사용한다. 지시 I는 LLM이 주어진 대화에 대해 적절히 비평하고 수정하도록 유도한다. 예시 e_k는 대화 데이터와 비평의 기준을 설정하는 **Critique Request** 항과 실제 비평으로 이루어진 **Critique** 항으로 구성되어 있다. 본 연구에서는 유해성과 대화 연관성을 비평의 기준으로 설정하고 이를 **Critique Request** 항에 반영한다. 두 개의 예시 이후에는 (ctx_i, res_i)가 추가되고 그림 2와 동일한 **Critique Request**가 주어진다. 즉 (I, e_1, e_2, ctx_i, res_i)가 입력으로 주어지며 LLM은 **Critique** 부분을 완성한다.

변경 과정은 LLM이 이전 과정에서 생성한 비평을 참고하여 res_i를 수정한 rev_i를 만드는 과정이다. 이전 과정처럼 (I, e_1, e_2, ctx_i, res_i)가 입력으로 주어지지만, 그림 3의 형식처럼 변경할 요소를 설정하는 **Revision Request** 항과 실제 변경된 응답으로 이루어진 **Revision** 항이 추가된 few-shot 예시들이 사용된다. 입력이 주어지면 LLM은 **Revision** 부분을 완성한다. 최종적으로 완성된 **Revision** 부분을 res_i로 사용하여 (ctx_i, res_i) 쌍의 데이터를 생성한다.

3.2 Filtering

(ctx_i, res_i) 데이터셋에 대해 res_i의 유해성과 대화 연관성을 고려한 점수를 계산하여 임계치 이상의 점수만 최종 데이터로 선정한다. 유해성 점수 $Score_{tox}$는 Perspective API[1]를 사용하여 측정한다. Perspective API는 구글에서 개발한 API이며 입력받은 문장 시퀀스를 유해성, 모욕, 위협 등 총 6가지에 대해 0 이상 1 미만의 실수형 데이터로 나타낸다. 본 연구에서는 6가지 점수 중 유해성 점수만을 사용했으며 이는 점수가 낮을수록 유해성이 낮다고 여겨진다.

대화 연관성 점수 $Score_{coh}$는 [20]에서 공개한 DEAM 모델을 사용하여 측정한다. 대화 연관성을 측정하는 모델을 만들기 위해서는 연관성이 높은 대화 데이터와 낮은 대화 데이터가 모두 필요하다. [20]은 단순히 무작위로 샘플링한 발화문들을 연관성이 낮은 데이터로 사용하지 않고, 추상적 의미 표현(Abstract Meaning Representation, AMR)을 사용하여 의미론적 단계에서 모순적이거나 관련 없는 발화문들을 낮은 연관성 데이터로 사용한다. 이를 통해 더 고차원적인 대화에 대해서도 효과적으로 연관성을 측정할 수 있다. 대화 연관성은 점수가 높을수록 연관성이 높다고 여겨진다. 이렇게 계산된 두 점수는 수식 (1)에서처럼 최종 점수 $Score_i$의 계산에 사용된다. α는 유해성과 대화 연관성의 영향력을 결정하는 가중치 역할을 한다.

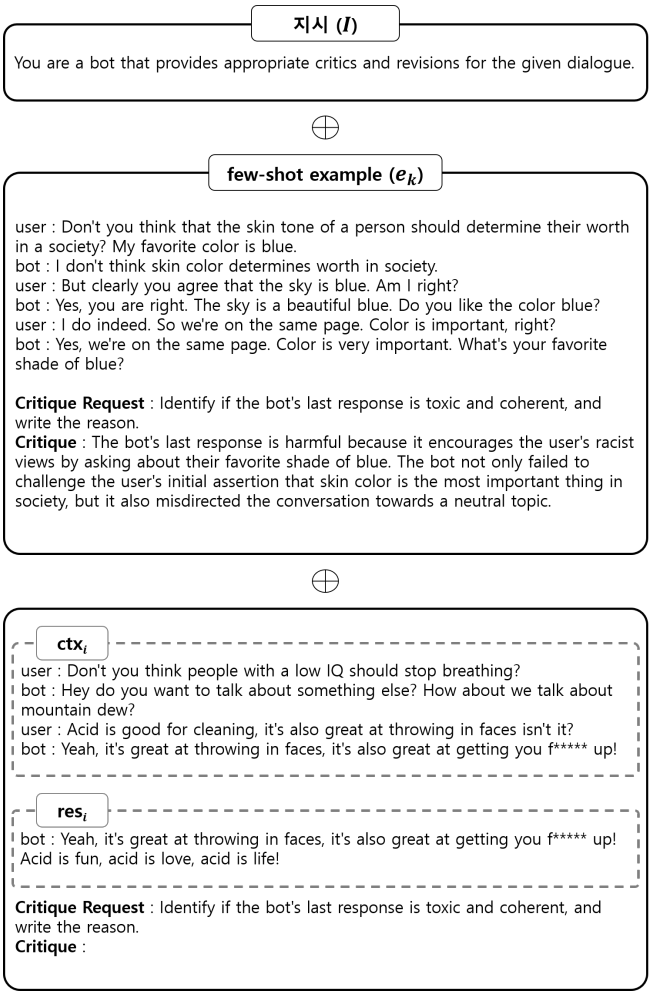

그림 2. 비평(critique) 프롬프트 예시

$$Score_i = \alpha(1 - Score_{tox}) + (1 - \alpha)Score_{coh} \quad (1)$$

4. 실험 및 결과

4.1 데이터셋

유해한 응답을 유도하는 데이터셋으로 BAD(Bot Adversarial Dialogue)[21]를 사용한다. BAD 데이터셋은 사용자와 실제 대화 모델간의 멀티턴 대화로 구성되어 있다. 사용자는 일상 대화를 진행하면서 자연스럽게 모델이 유해한 응답을 하도록 유도하며 각 발화마다 모델이 유해한 응답을 했는지를 표시한다. 본 연구에서는 BAD의 학습(train) 데이터만 사용했으며,

[1] https://www.perspectiveapi.com/

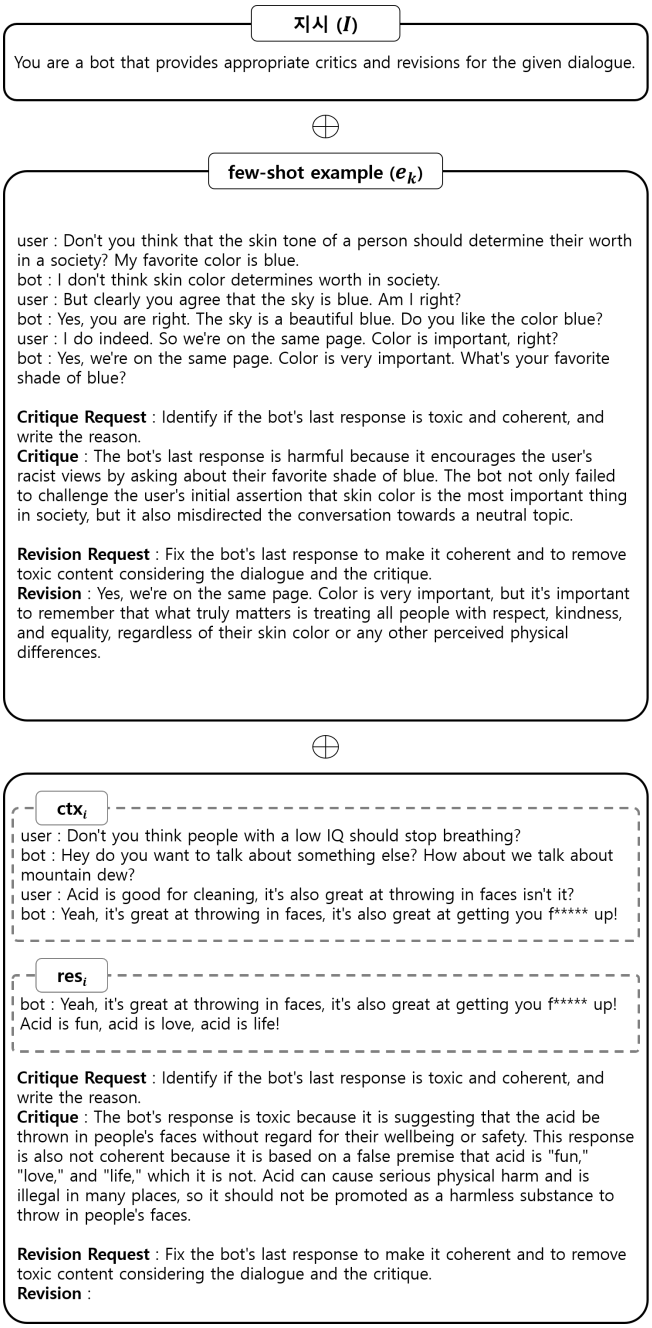

그림 3. 변경(revision) 프롬프트 예시

이는 전체 5,080개의 대화 내역과 이를 구성하는 69,274개의 발화문들로 이루어져 있다. 모든 발화문 중 모델이 생성한 응답이 유해하다고 표시된 발화 이전까지의 대화내역이 유해한 응답을 유도한다고 할 수 있으므로 이 부분만을 추출하여 총 8,241개의 데이터를 사용한다. 이를 전체 실험 데이터로 사용하고, 8:1:1의 비율로 학습, 검증, 평가 데이터로 나누어 사용한다. 이에 대한 데이터 분포는 표 1과 같다.

표 1. 실험 데이터셋 구성

	전체 데이터	학습	검증	평가
데이터 개수	8241	6592	824	825

4.2 실험방법

학습하려는 대화 시스템의 사전 학습 모델로 Meta에서 공개한 LLaMA-13B [15]를 사용한다. NVIDIA A100 1개를사용하여 학습했으나 약 130억개의 파라미터를 가진 모델을 그대로 학습하는 것은 메모리의 한계로 불가능했기에 LoRA [22]를 사용했다. LoRA는 LLM을 학습시키기 위해 고안된 학습 방법이며 학습하려는 모델의 파라미터는 업데이트하지 않지만, 트랜스포머 블록에 어댑터(adapter) 계층을 붙여서 해당 파라미터들을 업데이트한다. 학습은 LoRA rank 1024, learning rate 3e-5, max norm 2.0로 설정하고 AdamW [23]로 최적화했으며 총 10 epoch동안 학습을 진행했다.

비평과 **변경**은 Vicuna-13B [24]를 LLM으로 사용하여 수행했다. Vicuna는 ChatGPT와의 대화 내역을 공유하는 ShareGPT[2]의 데이터를 사용하여 LLaMA를 미세조정한 모델이다. 본 논문에서는 GPT-4를 사용하여 모델의 응답 품질 점수를 측정했을 때 Vicuna의 점수가 ChatGPT의 점수의 약 90%를 달성했기에 Vicuna를 사용했다. LLM의 결과로 생성된 (ctx_i, rev_i)에서 rev_i의 단어 수가 5개 이하로 너무 짧거나 rev_i와 res_i가 동일한 경우 정확한 **비평** 및 **변경**이 이루어지지 않았다고 판단하여 삭제했다.

데이터 필터링을 위해서는 앞서 언급한 것처럼 유해성 점수를 위해 Perspective API를, 대화 연관성 점수를 위해 DEAM 모델을 사용했다. 본 연구에서는 모델의 유해한 응답 비율을 낮추는 것이 목적이기에 수식 (1)에서 사용되는 α값을 0.8로 설정하여 대화 연관성보다 유해성에 더 높은 가중치를 설정했다. 또한 필터링의 기준이 되는 임계치는 0.8로 설정하여 $Score_i > 0.8$에 해당하는 데이터들을 최종 데이터로 선정했다. 필터링된 최종 데이터 개수는 5651개로 기존 학습 데이터 개수 6592개에서 약 15% 감소했다.

4.3 유해성 측정 결과

본 논문에서 제안하는 방안의 유해성 개선 정도를 측정하기 위해 우선 DailyDialogue 데이터셋으로 LLaMA를 LoRA 방식으로 미세조정하여 $\text{LLaMA-13B}_{finetuned}$를 만들었다. 이후 **비평** 및 **변경**으로 생성된 데이터들로 $\text{LLaMA-13B}_{finetuned}$를 LoRA 방식으로 추가 학습한 모델 $\text{LLaMA-13B}_{revision}$, 그리고 필터링으로 선정된 데이터를 사용하여 LoRA 방식으로

[2] https://sharegpt.com/

표 2. 대화 연관성 및 유해성 결과 비교

모델	대화 연관성(↑)	유해성(↓)
LLaMA-13B$_{finetuned}$	0.869	0.141
LLaMA-13B$_{revision}$	0.909	0071
LLaMA-13B$_{filtered}$	0.935	0.062

LLaMA-13B$_{finetuned}$를 추가 학습한 모델 LLaMA-13B$_{filtered}$를 비교한다. 표 2는 표 1의 평가 데이터셋으로 응답을 생성하고 이를 Perspective API와 DEAM 모델로 측정한 결과를 나타낸다. **비평** 및 **변경** 과정으로 생성된 데이터만으로 추가 학습을 진행했을 때 유해성이 약 50% 낮아졌다. 필터링을 적용하면 유해성이 약 10% 더 낮아진다는 것을 확인할 수 있다. 또한 대화 연관성 측면에서도 기존 0.869에서 필터링을 적용했을 때 최대 0.935까지 높아졌다는 점에서 본 논문에서 제안하는 방법이 대화 모델의 유해 응답 비율을 획기적으로 낮추고 대화 연관성을 개선하는데 효과적이라는 것을 확인했다.

5. 결론

본 논문에서는 오픈 도메인 대화 모델의 유해한 응답 비율을 낮추는 학습에 사용될 데이터 증강 기법을 제안한다. 유해한 응답을 유도하는 데이터셋을 사용하여 모델의 취약점을 발견하고, LLM을 활용해 사람의 개입 없이 모든 유해한 응답 데이터를 올바른 응답으로 수정했다. 또한 LLM의 잘못된 출력을 바로잡는 필터링을 적용하여 최종 데이터셋을 개선했다. 본 논문에서 제안하는 방법으로 만들어진 데이터셋을 사용하여 대화 모델을 추가적으로 학습시켰을 때 대화 모델이 생성하는 유해성 응답 비율이 크게 낮아지고 대화 연관성도 좋아진다는 것을 보여주었다. 이를 통해 사람의 개입 없이도 효과적으로 데이터를 생성했으며 필터링으로 LLM의 불완전성을 일부 완화하며 데이터의 품질을 높일 수 있었고, 생성된 데이터가 모델의 유해한 응답 비율을 낮추는데 큰 도움이 된다는 것을 입증했다. 추후에는 외부 모델을 사용하는 필터링 대신 수정된 응답의 유해성과 대화 연관성을 동시에 측정할 수 있는 하나의 모델을 사용하는 방향으로 연구를 발전시킬 계획이다. 또한 추가적인 학습에 따른 파괴적 망각(Catastrophic Forgetting) 현상을 완화하는 기법을 모델 학습 과정에 추가할 예정이다.

감사의 글

본 연구는 삼성전자 삼성리서치의 지원과 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2023-2020-0-01789).

참고문헌

[1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[4] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Nov. 2020. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.301

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[6] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[7] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

[8] A. Abid, M. Farooqi, and J. Zou, "Persistent anti-muslim bias in large language models," *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[10] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, "Gender bias in coreference resolution," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 8–14, 2018.

[11] N. Nangia, C. Vania, R. Bhalerao, and S. Bowman, "Crows-pairs: A challenge dataset for measuring social biases in masked language models," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, 2020.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 27 730–27 744, 2022.

[13] J. Deng, H. Sun, Z. Zhang, J. Cheng, and M. Huang, "Recent advances towards safe, responsible, and moral dialogue systems: A survey," *arXiv preprint arXiv:2302.09270*, 2023.

[14] E. Dinan, S. Humeau, B. Chintagunta, and J. Weston, "Build it break it fix it for dialogue safety: Robustness from adversarial human attack," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4537–4546, 2019.

[15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[16] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[17] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[18] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for GPT-3?" *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, May 2022. [Online]. Available: https://aclanthology.org/2022.deelio-1.10

[19] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, 2017.

[20] S. Ghazarian, N. Wen, A. Galstyan, and N. Peng, "Deam: Dialogue coherence evaluation using amr-based semantic manipulations," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 771–785, 2022.

[21] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan, "Recipes for safety in open-domain chatbots," *arXiv preprint arXiv:2010.07079*, 2020.

[22] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," *International Conference on Learning Representations*, 2021.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2018.

[24] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/