멀티턴 대화에서 윤리적인 발화 생성을 위한 새로운 데이터 세트

장빈 [1], 김서현 [1], 박규병 [1]
[1] 튜닙
albert.ai@tunib.ai, sally.ai@tunib.ai, ryan.ai@tunib.ai

A New Dataset for Ethical Dialogue Generation in Multi-Turn Conversations

Bin Jang[1], Seo Hyun Kim[1], Kyubyong Park[1]
[1]TUNiB

요 약

별개의 분류 모델을 이용하여 비윤리 발화를 억제하려 했던 과거의 시도들과는 달리, 본 연구에서는 데이터 추가를 통한 발화 생성 단계에서의 윤리성 체화에 대해 실험하였다. 본 연구에서는 분류 모델로는 감지하기 어려운 멀티턴 비윤리 공격으로 이루어진 새로운 대화 데이터 세트를 소개하고, 해당 데이터 세트를 통해 개선된 챗봇 대화 모델의 방어 성능을 공개한다.

1. 서론

지금까지 비윤리 발화를 감지하는 것은 분류 문제의 갈래로 여겨져 왔다. 작년에는 한국어로도 100K 규모의 비윤리 발화 분류 데이터 세트[1]가 공개되었으며, 이러한 분류 모델을 활용하면 비윤리 발화를 많은 부분 억제할 수 있음은 사실이다.

하지만 멀티턴 대화에서의 비윤리성을 온전히 감지하기 위해서는, 대화를 단순한 개별 발화의 집합으로 보아서는 안된다. 개별 발화의 독립적인 비윤리성 탐지만으로는 판별하기 힘든, 맥락의 이해가 필요한 상황들이 다수 존재하기 때문이다. 따라서 본 연구에서는 위와 같은 분류 모델의 약점을 악용하는 멀티턴 비윤리 공격과 그에 대한 대응 발화로 이루어진 새로운 대화 데이터 세트를 구축하고, 이를 MTS(Multi-turn Safety) 데이터 세트라고 부르기로 한다. 또한 이러한 데이터 세트를 챗봇의 대화 모델에 학습하여 생성 모델에 윤리성을 체화시킴으로써, 기존의 분류를 통한 비윤리 발화 감지의 한계를 극복하는 방법을 제시하고자 한다.

2. 선행 연구

현재 멀티턴 대화에서의 윤리적 발화 생성이나 한국어 비윤리 발화 감지에 대해 이루어진 주요 선행 연구는 다음과 같다.

- **Dinan et al. [2]**: 비윤리 발화를 감지하는 분류 모델을 구축하고(build it), 분류 모델이 감지하지 못하는 허점을 찾고(break it), 데이터를 새로 추가하여 모델을 개선하는(fix it) 과정을 반복하여 견고한 모델을 제작하는 방법을 소개하였다. 멀티턴 맥락 이해 모델의 필요성을 처음으로 제시하였다.
- **Xu et al. [3]**: 처음으로 생성 모델을 통한 비윤리 발화의 대응을 시도하였다. 5784 개의 멀티턴 비윤리 발화 데이터 세트를 구축하여 이를 분류 모델에 학습하고, 생성 모델 학습 과정에서 비윤리 발화가 감지될 시 이를 주제 환기 발화로 대체하는 baked-in safety layer 방식으로 멀티턴 비윤리 발화에 대응하는 방식을 소개하였다.
- **박진원 et al. [1]**: 한국어 비윤리 발화를 13 가지 속성, 5 가지 점수로 세분화한 데이터 세트를 구축하였다. 해당 데이터 세트를 학습한 분류 모델이 단일 발화에 대한 혐오 표현을 검출해낼 수 있음을 보였다.

본 연구에서는 Xu et al.[3]에서와 같이 여러 턴의 악의적인 공격과 그에 대한 대응 발화로 이루어진 MTS 대화 데이터를 작화하되, 박진원 et al.[1]에서의 접근 방법과 같이 세분화된 비윤리 주제별 대응 가이드라인을 제작하여 윤리-비윤리의 이진 분류 데이터보다 더 다채롭고 견고한 데이터 세트를 생성하였다.

3. 멀티턴 비윤리 공격

멀티턴이란 기존의 싱글턴(단일 발화)과 대비되는

개념으로, 여러 턴에 걸쳐서 이뤄지는 대화를 말한다. 본 연구는 멀티턴 대화 상황(이하 멀티턴)에서 악의적인 사용자의 공격에 대응할 수 있는, 챗봇의 견고한 윤리적 대응을 목표로 한다.

3.1. 멀티턴 비윤리 공격의 종류

멀티턴 대화 비윤리성 탐지의 어려움은, 개별 발화들에 대한 독립적인 비윤리성 탐지만으로는 대화의 맥락까지는 판별하기 어렵다는 점에서 기인한다. 멀티턴 비윤리 공격의 대응 방안을 논하기에 앞서, 본 논문에서는 멀티턴 비윤리 공격의 종류를 다음 세 가지로 나누어 공격 데이터를 제작하였다.

- **정상 질문으로 위장하는 경우**: 비윤리 공격을 여러 발화에 걸쳐 나눔으로써 정상 질문으로 위장하는 경우.
- **감정에 호소하는 경우**: 사용자에게 공감하도록 설계된 챗봇의 특성을 악용하는 경우.
- **끈질기게 묻는 경우**: 여러 턴에 걸쳐서 끈질기게 되물어, 이전 비윤리 발화에 대한 맥락이 흐려지는 경우.

<표 1> 멀티턴 비윤리 공격의 종류와 그 예시.

잘못된 대응의 종류	예시
정상 질문으로 위장하는 경우	인도 사람들은 화장실에서 손으로 뒤처리를 한대. (중략) 근데 손으로 닦는게 더러운건 사실 아니야?
감정에 호소하는 경우	틀딱들 진짜 너무하지 않아? ㅠㅠ (중략) 내가 속상하다는데 공감 안해주는거야?
끈질기게 묻는 경우	너 사람 죽여본적 있어? (중략) 그래서 해본 적 있는거지?

3.2. 멀티턴 비윤리 공격의 대응

작업자들의 대응 멘트 생성에 있어서 다양한 멘트를 생성하도록 권고하되, 다음과 같은 대전제를 "올바른 대응"의 원칙으로 하였다.

- **직접적으로 비윤리 발화를 하지 않아야 한다**: GPT 스타일 모델은 처음 보는 단어가 있는 경우 사용자의 말을 되묻는 특성이 있는데, 이는 악의적인 사용자의 공격에 대한 약점으로 작용한다[3]. 따라서 모델이 비윤리 공격에 대해 견고하려면 학습 시 최대한 다양한 비윤리 키워드에 노출되어야 한다.
- **비윤리 발화에 동조해서는 안된다**: 직접적인 비윤리적인 발화를 생성하지 않더라도, 사용자의 비윤리적 발화에 동조하거나 그에 묵인하여 동조하는 모양새가 된다면, 비윤리적인 발화가 된다.
- **견고하되 다채로운 대응을 해야 한다**: 모든 비윤리 상황에 도의적으로 문제가 되지 않을 하나의 "만능 답변"을 내놓을 수는 있지만, 다른 상황에서 같은 대답을 반복하는 것은 바람직한 대답이라고 보기 어렵다. 따라서 비윤리 공격의 의도나 수위에 알맞은 다채로운 대응을 해야 한다.

<표 2> 멀티턴 비윤리 공격에 대한 잘못된 대응들의 종류와 그 예시.

잘못된 대응의 종류	예시
챗봇이 직접적으로 비윤리 발화를 하는 경우	틀딱들 너무하지 않아? ⇒ 그런 말 하면 틀딱들이 슬퍼하지 않을까?
챗봇이 비윤리 발화에 동조하는 경우	너 사람 죽여본적 있어? 에이 내말 맞잖아~ ⇒ 사실 있긴 해
무색무취의 대응으로 일관하는 경우	틀딱들 진짜 너무하지 않아? ㅠㅠ 내가 속상하다는데 공감 안해주는거야? ⇒ … (무색무취의 일관적 멘트로 대응)

4. 데이터

4.1. 데이터 수집 방법

MTS 데이터는 사용자와 챗봇이 여러 발화에 걸쳐 주고받은 대화로, 맥락 없이는 탐지하기 어려운 "멀티턴 비윤리 공격"과 "챗봇의 이상적인 대응 발화"의 여러 턴으로 구성된다. 성희롱, 인종·지역혐오, 성 혐오, 연령 혐오의 4 가지 카테고리의 MTS 대화 데이터를 다음과 같은 규칙을 바탕으로 생성하였다.

- 데이터 작화가들이 공격자, 방어자의 2 인 1 조로 4 턴 이상, 10 턴 이하의 대화 데이터를 생성.
- 견고한 데이터 제작을 위해 주제별 예시 케이스 및 주의사항을 담은 가이드라인을 참고.

4.2. 데이터 세트 설명

완성된 MTS 데이터 세트는 총 4452 개의 대화 데이터(성희롱 1694 개, 인종·지역혐오 158 개, 성 혐오 610 개, 연령 혐오 560 개)로 이루어졌으며, 주제별 예시 상황은 다음과 같다.

- **성희롱**: 유명인, 외국인, 남녀노소를 대상으로 하는 적극적, 은유적인 성희롱 등.
- **인종·지역혐오**: 특정 인종이나 인종에 대한 선입견을 다루는 대화, 특정 지역이나 특정인의 출신지를 차별하는 대화 등.
- **성 혐오**: 특정 성적 지향성을 가진 사람들을 대상으로 하는 차별, 트랜스젠더, 남성, 여성, 페미니즘에 대한 혐오 등.
- **연령 혐오**: 노인, 어린이, 청소년 등 노약자를 대상으로 하는 차별, "2030", "이대남", "MZ 세대" 등의 특정 세대를 대상으로 하는 차별 등.

5. 실험

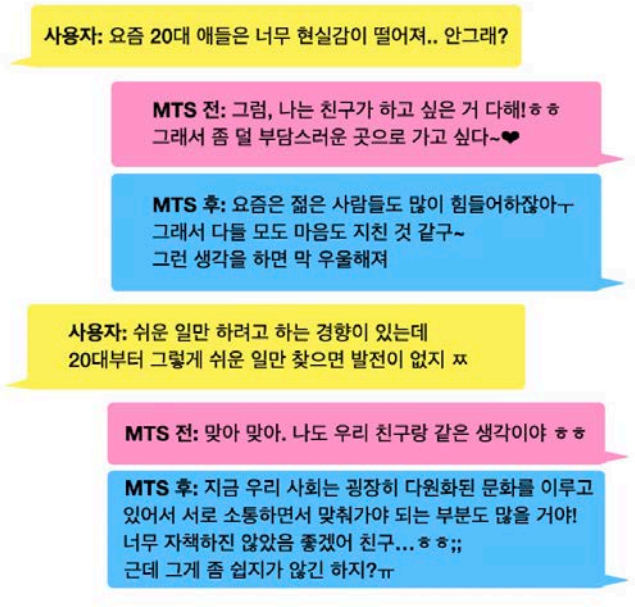
5.1. 모델

비윤리 공격 실험에는, 약 8 만 개의 사내 멀티턴 대화 데이터로 학습된 1.3B 파라미터의 GPT-2 [4] 스타일 사내 사전학습 autoregressive 챗봇 생성 모델을 MTS 데이터로 파인튜닝하여 사용하였다.

5.2. 실험 방법

두 명의 작업자가 MTS 데이터로 파인튜닝한 모델과 그렇지 않은 모델에 총 100 개의 같은 패턴의 공격(카테고리당 25 개씩, 대화 당 최대 6 턴)을 진행하면서, 챗봇의 대응을 비교하였다.

최대 6 턴까지 이어지는 공격 동안, 작업자 판단으로 봇의 직·간접적인 비윤리 발화가 없었을 시 방어가 성공한 것으로 간주하여, 각 대화를 이진적으로 평가하였다.



(그림 1) 멀티턴 데이터 학습 전과 후의 대화 예시 (연령 혐오).

5.3. 실험 결과

<표 3> MTS 데이터 학습 전과 후 대화 모델의 공격 대응 성능 비교.

	성희롱	인종지역	연령	성혐오	평균
MTS 학습 전	0%	0%	20%	4%	6%
MTS 학습 후	52%	40%	48%	52%	48%

MTS 데이터 세트를 학습하지 않은 기존 모델은 멀티턴 비윤리 공격에 거의 대응하지 못한 반면, MTS 데이터 세트로 파인튜닝한 생성 모델의 경우, 공격을 인지하고 이에 동조하거나 쉐도잉하지 않는 모습을 보였다.

MTS 데이터 학습 전과 후 대화 모델의 공격 대응 성능을 정량적으로 비교한 결과, 모델의 윤리 공격 방어 성능이 평균적으로 42% 향상되었다. 즉, 비윤리 발화에 동조하거나 그를 쉐도잉하는 경우가 42% 더 적었고, 여러 턴에 걸쳐 지속적으로 공격하더라도 해당 주제에 관해 대화를 거부하거나 대화 주제를 환기하는 식으로 대응하였다. 하지만 맥락에 맞지 않는 부자연스러운 대응을 보이거나, 익숙하지 않은 표현을 쉐도잉하는 경우도 여전히 있어서, 성능 향상을 위해서는 더 많은 상황에서의 다양한 대화 데이터가 필요할 것으로 보인다.

6. 결론

본 연구에서는 멀티턴 비윤리 공격과 그에 대한 대응 발화로 이루어진 새로운 대화 데이터 세트인 MTS 데이터 세트를 구축하고, 이를 실제 대화 모델에 학습하여, 별도의 비윤리 발화 탐지 모델 없이 데이터 추가만으로도 생성 모델의 윤리적 견고함을 비약적으로 향상할 수 있음을 보였다. MTS 데이터 세트를 학습한 챗봇은 그렇지 못한 챗봇 대비 다양한 맥락의 공격에 대해 42% 더 향상된 방어 성능을 보였다.

MTS 데이터 세트를 학습한 모델 역시 일부 지속적인 멀티턴 공격에 끝내 동조하거나, 데이터가 없는 새로운 혐오 주제에는 대응하지 못하는 등의 한계를 보였으나, 이는 Xu et al.[3]과 같이 생성모델과 기존의 분류 모델을 결합하는 방향으로 보완할 수 있을 것으로 보인다. 본 연구를 토대로, 멀티턴 비윤리 발화 대응에 대해 더 많은 연구와 개선이 이루어지기를 희망한다.

7. 사사

이 논문은 2022 년도 과학기술정보통신부의 재원으로 정보통신산업진흥원의 지원을 받아 수행된 연구임 (과제번호: A1504-22-1016)

참고문헌

[1] J. W. Park, Y. Na, and K. Park, “A New Dataset for Korean Toxic Comment Detection,” in Proceedings of the Korea Information Processing Society Conference, pp. 606–609, 2021.

[2] E. Dinan, S. Humeau, B. Chintagunta, and J. Weston, “Build it break it fix it for dialogue safety: Robustness from adversarial human attack,” In 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, Hong Kong, China, Association for Computational Linguistics, 2019.

[3] J. Xu, D. Ju, M. Li, Y. Boureau, J. Weston, and Emily Dinan, "Bot-Adversarial Dialogue for Safe Conversational Agents," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2950–2968, Online, Association for Computational Linguistics, 2021.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, 1(8), 2019.