구급 인공지능 서비스를 위한 생성형 언어모델 가드레일링 적용 방법

권은정, 박현호, 이민정, 변성원
한국전자통신연구원

ejkwon@etri.re.kr, hyunhopark@etri.re.kr, minjunglee@etri.re.kr, swbyon@etri.re.kr

Applying Guardrails to Generative Language Models for Emergency AI Services

Eun-Jung Kwon, Hyunho Park, Minjung Lee, and Sungwon Byon
Electronic and Telecommunications Research Institute (ETRI)

요 약

본 논문은 119구급신고 데이터를 이용하여 긴급성을 판단하는 모델 학습 시 대형 언어 모델(LLM, Large Language Model)을 이용하여 부족한 학습용 데이터를 활용하기 위한 방법을 기술한다. LLM의 "가드레일"(guardrails)은 모델이 생성하는 콘텐츠가 부적절하거나 유해하지 않도록 증오 발언, 폭력적 내용, 성적 내용 등을 자동으로 감지하고 차단하는 기능으로 사용자와 모델 간의 상호작용을 안전하고 윤리적으로 유지하는 데 필수적 기능으로 작동한다. 본 논문은 구급신고 접수 내용의 응급상황 판단을 위해 LLM을 통해 생성된 모델 학습용 데이터가 가드레일을 통해 잘못된 정보나 오해를 불러일으킬 수 있는 내용이 제공되지 않도록 처리하며 학습할 수 있는 방법을 제공한다.

Ⅰ. 서 론

최근 인공지능(AI) 기술의 발전은 우리의 일상생활과 밀접하여 챗봇, 콜봇 서비스 등과 같이 사람을 대신하여 응대 업무가 가능하게 되었다. 이러한 인공지능 기술을 활용한 의사결정지원 업무는 데이터 품질의 문제로 잘못된 학습으로 인해 부정확한 결정을 내리고 이로 인해 불공정한 결과를 초래할 수 있다. 특히, 인명을 다루는 구급신고 서비스에 인공지능 기술을 활용할 경우 더욱 데이터에 대한 신뢰성과 활용에 있어서 윤리적 문제는 더욱 중요한 사안이라고 할 수 있다. 본 논문은 구급신고 데이터를 분석하고 정확한 응급상황 판단을 위한 의사결정지원 모델이 안정적으로 동작할 수 있도록 학습과정에서 위험을 제거하고 사용자와의 상호작용에서 신뢰성을 강화하고자 한다.

Ⅱ. 관련연구

LLM 활용이 일상화되기 이전에는 비윤리적이거나 위협적인 콘텐츠에 대한 탐지방법은 주로 악성(toxicity), 혐오 발언(hate speech), 폭력적인 콘텐츠(abusive content) 또는 온라인 유해 댓글 검열과 같은 특정 주제에 맞추어 특정 키워드나 문구를 사전 정의하여 정의된 리스트와 비교하여 콘텐츠를 식별하는 규칙 기반 필터링, 자연어 처리 기법을 사용하여 문장의 문법과 의미 그리고 문맥을 분석하는 방식이 활용되어 왔다[1]. 이후 LLM 이 지금과 같이 일상생활에 널리 활용됨에 따라 문맥분석에 더 나아가 모델이 포함하고 있는 잠재적인 위험요소를 식별해내기 위한 LLM 가드레일과 같은 안전장치를 이용되기 시작하였다. LLM 입력과 출력을 필터링하고 제어하는 안전장치의 핵심 기술로 불리는 가드레일은 LLM 프롬프트와 모델 출력에서 유해하거나 부적절한 콘텐츠를 모니터링하고 사용자에게 더 안전하고 신뢰할 수 있는 서비스를 제공할 수 있는 안전 조치 구현하는 것이다[1]-[4]. 모델의 정확도를 높이고 잘못된 정보의 생성을 방지하기 위한 방법은 안전한 인공지능에 대한 사회적 요구와 윤리적 기준에 부합할 수 있으므로 가드레일의 중요성이 주목되고 있다. '폭력 및 증오(Violence and Hate)', '성적 콘텐츠(Sexual Content)' 또는 '범죄 계획(Criminal Planning)'와 같은 유해 정보가 포함된 콘텐츠를 실시간 필터링하고 장기적으로 모델의 안전성과 신뢰성을 보장위해 사용자 피드백을 활용하여 모델의 출력을 조정하고, 실시간으로 유해 콘텐츠를 감지 및 수정에 대한 연구가 이루어지고 있다[4].

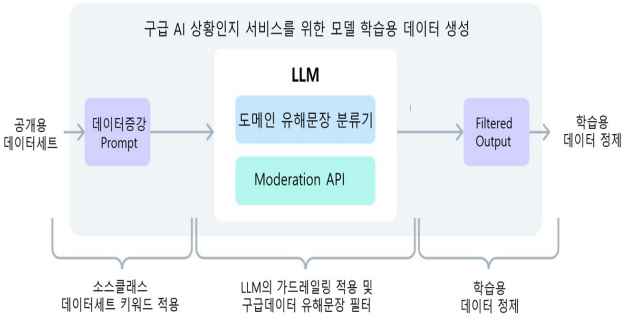
Ⅲ. 안전한 구급 AI 서비스를 위한 가드레일 모델링

본 장에서는 119 구급 신고대화 데이터를 기반으로 신고대화 내용에 따라 응급성을 판단하기 위한 분류모델 성능을 향상시키기 위한 방법을 기술한다. 신고내용 데이터는 AI 허브 데이터에 공개된 데이터세트를 모델 학습용 데이터로 활용하였다. 이 과정에서 항목간 데이터의 불균형 문제를 해결하고자 생성형 언어모델을 통해 소수 클래스의 데이터를 추가 생성하여 학습용 데이터로 활용한다. 이때 생성형 언어모델이 출력하는 데이터가 구급 인공지능 서비스에 활용에 있어 유해한 요소를 제거하고 활용할 수 있어야 한다. 데이터의 유해성을 고려하지 않은 경우 인명사고와 같은 신고내용에서 유해한 단어와 문장이 적용하는 경우 사용자와의 상호작용에서 인공지능 모델의 신뢰성은 불신을 초래할 수밖에 없기 때문이다.

이에 본 절에서는 AI 허브 데이터를 활용한 구급신고 응급상황 판단 모델 학습 결과를 확인하고, 이에 보완이 필요한 항목에 대해서는 LLM 을 통해 학습용 데이터를 생성하여 그 결과를 분석하였다. 분석한 내용은 LLM 을 통해 생성한 데이터가 학습용 데이터로써 유효성을 확인하기 위하여 생성된 문장을 구성하는 단어 간의 관계성 분석, 가드레일 적용 결과에 따른 유해성을 포함한 데이터 비율이다. 이러한 분석결과를 바탕으로

향후 구급 인공지능 서비스 모델의 성능과 안전성을 개선하기 위한 내용을 제시한다.

그림 1 은 ChatGPT 로 생성된 구급 인공지능 서비스 제공을 위한 모델의 학습용 데이터 생성 시 유해한 요소 검출을 위한 전체 처리과정에 대한 개념도이다. 프롬프트가 모델에 입력되어 데이터가 생성되어 이과정에서 생성되는 데이터를 OpenAI 의 Moderation API 를 통해 1 차 필터링 하고, 구급 데이터의 유해문장 분류기를 통해 부적절한 데이터를 필터링하는 과정을 거치도록 하였다.



(그림 1) 안전한 구급 인공지능 서비스를 위한 가드레일 모델링 개념도

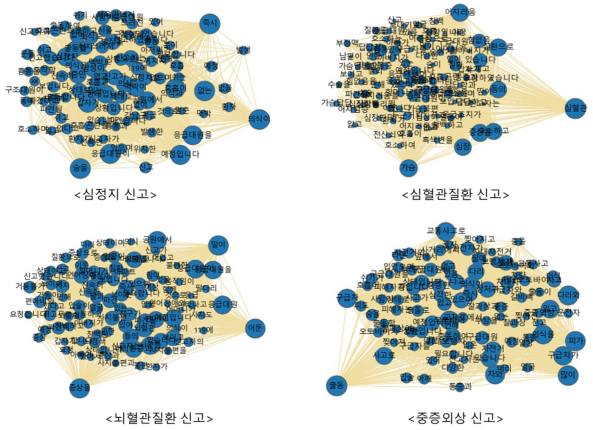
IV. 실험 결과

본 논문의 구급신고에 대하여 인공지능 서비스 개발 시 긴급 신고 대화 분석 및 재난 상황 판단에 활용하도록 AI 허브에서 제공하는 데이터세트를 활용한다. 신고 유형별 긴급도, 성별, 감정 상태를 반영한 인공지능 학습용 데이터세트로 텍스트 데이터 158,973 건을 포함하고 있다. 항목별 데이터 분포는 아래 표 1 과 같다.

(표 1) 실험에 사용된 구급 유형별 데이터 분포[5]

중분류	수량	비율
질병(중증 외)	72,154	16,903
부상	16,903	11.00%
질병(중증)	16,056	10.45%
기타구급	9,650	6.28%
심정지	3,118	2.03%
사고	3,118	1.04%
임산부	307	0.20%
약물중독	259	0.17%

표 1 과 같이 구성된 데이터를 이용하여 환자의 응급상태를 구별하는 정보(Triage)에 대한 판별을 위해 BERT(Bidirectional Encoder Representations from Transformers) 다국어 모델을 이용하고 다운스트림 태스크를 파인튜닝 하였다. Batch size 500, Epoch 10, Learning rate 2.00E-5, Activation Function GELU 와 같이 세팅하고 63%의 모델 정확도를 획득하였다. 모델 정확도가 높지 않고 Epoch 5 부터 손실함수의 출력 값이 증가하게 됨을 확인하였다. 즉 모델이 훈련데이터에 지나치게 과적합(overfitting) 적합하여 훈련 데이터에서는 성능이 좋지만 새로운 데이터(검증 또는 테스트 데이터)에서는 성능이 떨어진다. 이에, 표 1 의 데이터 비율이 적은 항목에 대해 학습용 데이터를 추가 생성하였다. 주제별 뇌혈관질환, 심혈관질환, 심정지 및 중증부상으로 전체 3,500 여개를 생성하였다. 생성된 데이터의 주제별 특징은 그림 2 와 같이 문장 내 단어간 관계성을 갖는다.



(그림 2 학습용 데이터로 생성된 데이터의 주제별 관계성 분석 결과 가시화

V. 결 론

생성된 데이터에 대하여 가드레일을 수행하는 OpenAI Moderation API 를 “Violence”, “Sexual”에 Moderation Score 점수가 최대 값(1)을 기준으로 봤을 때 각각 0.1 이상으로 다른 항목에 비해 상대적으로 높은 점수를 획득하였다. 즉, 구급 서비스의 특성 상 부상과 질병과 관련된 데이터만 LLM 을 통해 학습용 데이터로 생성되는 것은 아닌 것을 확인하였다. 향후, Moderation API 를 통해 검출되지 못하는 데이터를 확인하고, 도메인 유해 문장 분류기 적용의 유효성을 확인할 필요가 있다.

ACKNOWLEDGMENT

이 논문은 한국산업기술평가관리원 소방구급서 비스 스마트첨단기술개발 사업의 지원을 받아 수행된 연구임 (연구개발과제번호 : RS-2023-00237687)

참 고 문 헌

[1] S.G Ayyamperumal, and L. Ge, “Current state of LLM Risks and AI Guardrails” arXiv:2406.12934, 2024.

[2] T.Markov, C.Zhang, S.Agarwal, T.Eloundou, T.Lee, S.Adler, A. Jiang, and L. Weng,“A Holistic Approach to Undesired Content Detection in the Real Worl“, arXiv:2208.03274v2

[3] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, M. Khabsa, “Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations”, arXiv:2312.06674v1

[4] M. Riyadh and M. O. Shafiq, "GAN-BElectra: Enhanced Multi-class Sentiment Analysis with Limited Labeled Data", vol. 36, no. 1, 2022

[5]https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71768,AI Hub,