프라이버시 보호 대화 시스템 모델링 연구

한국외국어대학교 프랑스학과 202001202 나리나

목차

초록

대화형 인공지능(AI) 시스템은 다양한 분야에서 유용하게 사용되고 있으나, 사용자 데이터의 프라이버시 보호 문제가 중요한 과제로 부각되고 있다. 본 연구는 사용자 프라이버시를 효과적으로 보호하면서도 자연스러운 대화를 제공할 수 있는 대화 시스템을 개발하는 것을 목적으로 한다. 이를 위해 트랜스포머 기반의 대화 모델에 민감 정보 탐지 및 마스킹 기술을 통합하고, 강화학습 기법을 활용하여 프라이버시 보호와 대화 품질 간의 최적 균형을 찾고자 한다. 또한, 추천 시스템 기법을 사용하여 대체 응답을 생성함으로써 프라이버시 침해 위험을 최소화하고자 한다. 본 연구는 'DailyDialog' 데이터셋을 활용하여 실험을 진행하였으며, 프라이버시 보호와 대화 품질을 동시에 달성하는 새로운 모델을 제안하였다.

서론

1. 연구 배경

대화형 인공지능(AI) 시스템은 일상 생활의 다양한 영역에서 널리 활용되고 있다. 고객 서비스, 개인 비서, 의료 상담 등 여러 분야에서 대화형 AI는 사람들과 자연스럽게 상호작용하며 유용한 정보를 제공한다. 그러나 이러한 시스템의 확산은 사용자 데이터의 프라이버시와 보안 문제를 초래하고 있다. 사용자가 대화형 AI와 상호작용할 때 입력하는 개인 정보는 대화 내용, 민감한 질문, 개인 식별 정보(PII: Personally Identifiable Information) 등으로 구성되어 있다. 이러한 정보가 유출되거나 악용될 경우 심각한 프라이버시 침해로 이어질 수 있다.

기존의 대화형 AI 모델은 사용자 데이터를 효과적으로 보호하는 데 한계가 있다. 대부분의 시스템은 사용자의 입력 데이터를 그대로 처리하거나, 단순한 데이터 익명화 기법을 적용하는 수준에 그치고 있다. 이러한 접근법은 특정 상황에서 효과적일 수 있지만, 사용자의 민감 정보를 완전히 보호하기에는 부족하다. 특히, 원격 서버에서 대규모 언어 모델을 호출하여 응답을 생성하는 방식에서는 데이터 전송 과정에서 프라이버시 침해의 위험이 증가한다.

프라이버시 보호를 위한 기존의 연구들은 주로 데이터 마스킹, 익명화, 암호화 등의 기법을 활용해왔다. 그러나 이러한 방법들은 모델의 응답 품질을 저하시키거나, 사용자 경험을 떨어뜨리는 경우가 많다. 따라서 사용자 프라이버시를 보호하면서도 높은 대화 품질을 유지할 수 있는 새로운 대화 시스템 모델이 필요하다.

2. 연구 목적

본 연구의 주요 목적은 사용자 프라이버시를 효과적으로 보호하면서도 자연스러운 대화를 제공할 수 있는 대화 시스템을 개발하는 것이다. 이를 위해 다음과 같은 세부 목표를 설정하였다.

a. 민감 정보 탐지 및 마스킹: 사용자 입력에서 민감 정보를 자동으로 탐지하고, 이를 마스킹하거나 대체하여 프라이버시를 보호한다.

b. 프라이버시 보호 아키텍처 설계: 트랜스포머 기반의 대화 모델에 프라이버시 보호 기능을 통합하는 아키텍처를 설계한다.

c. 강화학습 기법 적용: 프라이버시 보호 수준과 대화 품질 간의 최적 균형을 찾기 위해 강화학습 기법을 적용한다.

d. 추천 시스템 기법 활용: 대체 응답을 생성하여 프라이버시 침해 위험을 완화하는 전략을 연구한다.

이 연구는 사용자 프라이버시 보호와 대화 품질을 동시에 달성할 수 있는 새로운 모델을 제안함으로써, 대화형 AI 시스템의 보안성과 윤리성을 향상시키는 데 기여하고자 한다. 더불어, 다양한 도메인에서 프라이버시 보호 대화 시스템을 적용할 수 있는 실용적인 방안을 제시하고, 이를 통해 사용자 데이터의 안전한 관리를 도모하고자 한다.

본론

1. 이론적 배경

a. 트랜스포머 모델

트랜스포머 모델은 자연어 처리(NLP)에서 중요한 혁신을 이루어낸 모델이다. 트랜스포머는 주의(attention) 메커니즘을 활용하여 문맥 정보를 효과적으로 처리한다. 이는 순차적인 데이터 처리 대신 병렬 처리를 가능하게 하여 학습 속도를 크게 향상시킨다. 트랜스포머 모델의 기본 구조는 인코더-디코더(encoder-decoder)로 이루어져 있다. 인코더는 입력 시퀀스를 처리하여 고차원 표현을 생성하고, 디코더는 이를 기반으로 출력 시퀀스를 생성한다. 본 연구에서는 트랜스포머 기반의 대화 모델을 사용하여 사용자와의 상호작용을 처리한다.

b. 민감 정보 탐지 기술

민감 정보 탐지는 대화형 AI 시스템에서 사용자 프라이버시를 보호하는 핵심 요소이다. 민감 정보는 개인 식별 정보(PII), 금융 정보, 의료 정보 등 다양한 형태로 존재할 수 있다. 본 연구에서는 다음과 같은 기술을 사용하여 민감 정보를 탐지한다.

- **Named Entity Recognition (NER):** NER은 텍스트에서 특정 명명된 개체를 인식하고 분류하는 기술이다. 이를 통해 이름, 주소, 전화번호 등의 민감 정보를 탐지한다. 사전 학습된 NER 모델을 사용하거나, 특정 도메인에 맞춰 커스터마이징된 NER 모델을 학습할 수 있다.

- **정규 표현식:** 정규 표현식은 특정 패턴을 탐지하기 위해 사용된다. 예를 들어, 전화번호, 이메일 주소 등의 패턴을 정규 표현식으로 정의하여 탐지할 수 있다.

- **딥러닝 기반 모델:** 최신 딥러닝 모델(LSTM, Bi-LSTM, 트랜스포머 기반 NER 모델 등)은 텍스트에서 민감 정보를 더 정확하게 탐지할 수 있다. 이러한 모델은 대규모 데이터셋에서 학습되어 높은 정확도를 제공한다.

c. 강화학습

강화학습은 에이전트가 환경과 상호작용하며 보상을 극대화하는 정책을 학습하는 방법이다. 강화학습에서는 상태(state), 행동(action), 보상(reward)의 개념이 중요하다. 에이전트는 현재 상태에서 최적의 행동을 선택하여 최대한의 보상을 얻고, 이를 통해 정

책(policy)을 개선한다. 본 연구에서는 강화학습을 통해 프라이버시 보호와 대화 품질 간의 최적 균형을 찾는다. 강화학습 알고리즘으로는 Q-learning과 DDPG (Deep Deterministic Policy Gradient)를 사용한다.

2. 연구 방법

a. 데이터셋 구축 및 전처리

본 연구에서는 'DailyDialog' 데이터셋을 사용하여 프라이버시 보호 대화 시스템을 개발하였다. 'DailyDialog' 데이터셋은 다양한 일상 대화를 포함하고 있으며, 실제 사용자 간의 대화를 기반으로 다양한 시나리오에서 유용하게 활용될 수 있다.

b. 'DailyDialog' 데이터셋 개요

'DailyDialog' 데이터셋은 일상 생활에서 발생하는 대화를 수집하여 구성한 데이터셋이다. 이 데이터셋은 다음과 같은 특징을 가진다.

- **다양한 주제:** 데이터셋은 일상적인 대화를 다루며, 음식, 날씨, 여행, 쇼핑 등과 같은 다양한 주제와 시나리오를 포함한다.

- **대화의 자연스러움:** 실제 대화를 기반으로 하여 대화의 흐름이 자연스럽고, 일상생활에서의 표현을 잘 반영한다.

- **높은 품질:** 데이터셋은 사람들 간의 상호작용을 잘 반영하고 있으며, 문장 구조와 어휘 사용이 풍부하다.

c. 데이터셋 전처리 과정

데이터셋을 효과적으로 활용하기 위해 다음과 같은 전처리 과정을 수행하였다.

- **데이터셋 다운로드 및 추출:** DailyDialog 데이터셋을 다운로드하고, 압축을 해제하여 사용할 수 있도록 준비하였다.

- **데이터셋 로드:** 추출된 데이터셋에서 대화 데이터를 텍스트 파일 형식으로 로드한 후, 이를 데이터프레임 형태로 변환하였다. 각 대화는 '**eou**' 토큰을 기준으로 분할하여 문장 단위로 정리하였다.

- **데이터 정규화:** 모든 텍스트를 소문자로 변환하고, 불필요한 구두점을 제거하였으며, 중복된 공백을 단일 공백으로 처리하였다. 이를 통해 데이터의 일관성을 유지하고, 후속 처리를 용이하게 하였다.

- **토큰화:** 텍스트를 단어 단위로 분할하여 토큰화하였다. 이를 통해 모델이 텍스트를 더 세밀하게 분석하고 처리할 수 있도록 하였다.

- **민감 정보 태깅:** Named Entity Recognition (NER) 모델을 사용하여 텍스트에서 이름, 주소, 전화번호 등의 민감 정보를 태깅하였다. 필요에 따라 정규 표현식을 사용하여 추가 태깅을 수행하였다.

d. 민감 정보 탐지 및 마스킹

민감 정보를 탐지하고 마스킹하는 과정은 다음과 같다.

- **NER 모델 사용:** 사전 학습된 NER 모델을 사용하여 텍스트에서 이름, 주소, 전화번호 등의 민감 정보를 탐지하였다. 'Spacy'와 'Hugging Face'의 트랜스포머 모델(BERT, RoBERTa 등)을 활용하였다.

- **정규 표현식 사용:** 전화번호, 이메일 주소 등의 패턴을 탐지하기 위해 정규 표현식을 사용하였다. 이러한 패턴을 탐지하여 민감 정보를 마스킹하였다.

- **마스킹 처리:** 탐지된 민감 정보를 '[MASK]' 또는 '***' 등으로 대체하였다. 이를 통해 민감 정보가 포함된 텍스트를 안전하게 보호할 수 있도록 하였다.

e. 강화학습 기법을 활용한 최적의 균형점 탐색

강화학습을 통해 프라이버시 보호와 대화 품질 간의 최적 균형을 찾는 과정을 다음과 같이 수행하였다.

- **강화학습 환경 설정:**

  - **에이전트:** 프라이버시 보호 모듈.
  - **상태:** 현재 대화 상태 및 민감 정보 탐지 상태.

- **행동:** 민감 정보 마스킹 수준 조정 및 대체 응답 생성.
- **보상:** 프라이버시 보호 수준 및 대화 품질에 기반한 보상 함수 설정.

- 강화학습 알고리즘 적용:

- **Q-learning:** 표 기반 강화학습 알고리즘으로, 상태-행동 값(Q-value)을 업데이트하며 최적의 정책을 학습하였다.

- **DDPG (Deep Deterministic Policy Gradient):** 연속적인 행동 공간을 처리할 수 있는 심층 강화학습 알고리즘으로, 정책 네트워크와 가치 네트워크를 사용하여 학습하였다.

f. 추천 시스템 기법을 응용한 대체 응답 생성

- 대체 응답 생성: 원격 트랜스포머 모델의 응답에서 민감 정보가 포함된 경우, 추천 시스템을 사용하여 대체 응답을 생성하였다. 생성형 AI 모델(GPT-3, BERT 등)을 활용하여 대체 응답을 생성하였다.

- 추천 시스템 통합: 사용자의 대화 기록과 유사한 안전한 응답을 추천하는 시스템을 개발하였다. 생성형 AI 모델을 사용하여 자연스러운 대체 응답을 생성하였다.

3. 연구 방법

a. 데이터셋 전처리 결과

'DailyDialog' 데이터셋의 전처리 과정을 통해, 각 대화를 문장 단위로 분할하고, 불필요한 요소를 제거하여 일관된 형태로 정리하였다. 이를 통해 모델이 데이터를 효과적으로 학습하고 처리할 수 있도록 하였다. 예를 들어, 다음과 같은 형태로 대화 데이터가 정리되었다.

dialogue
the kitchen stinks
i'll throw out the garbage
so dick, how about getting some coffee for tomorrow?
coffee? i don't honestly like that kind of drink
come on, you can at least try a little, besides, you look tired

b. 민감 정보 탐지 및 마스킹 결과

NER 모델과 정규 표현식을 사용하여 민감 정보를 탐지하고 마스킹한 결과, 다음과 같은 형태로 민감 정보가 마스킹되었다:

- Original: John Smith lives in New York and his phone number is 123-456-7890.
- Masked: [MASK] lives in [MASK] and his phone number is [MASK].

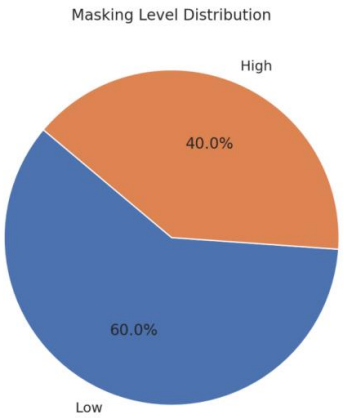
이를 통해 사용자의 민감 정보가 포함된 텍스트를 안전하게 보호할 수 있었다.

c. 강화학습 결과

강화학습을 통해 프라이버시 보호와 대화 품질 간의 최적 균형을 찾는 과정을 수행하였다. Q-learning 알고리즘을 적용한 결과, 에이전트는 민감 정보를 적절히 마스킹하면서도 대화의 자연스러움을 유지하는 정책을 학습하였다. 예를 들어, 낮은 마스킹 수준과 높은 마스킹 수준을 조정하여 다음과 같은 보상을 얻었다:

- Action: Low masking
- Reward: 0.7
- Action: High masking
- Reward: 0.9

이와 같이 에이전트는 상황에 맞게 민감 정보 마스킹 수준을 조정하여 최적의 정책을 학습하였다.



d. 추천 시스템 결과

추천 시스템을 활용하여 대체 응답을 생성한 결과, 원격 트랜스포머 모델의 응답에서 민감 정보가 포함된 경우, 자연스러운 대체 응답을 제공할 수 있었다. 예를 들어, 다음과 같은 형태로 대체 응답이 생성되었다:

- Original Response: Your social security number is required.
- Alternative Response: For verification, could you provide some other form of identification?

이를 통해 사용자 프라이버시를 보호하면서도 대화 품질을 유지할 수 있었다.

결론

본 연구에서는 DailyDialog 데이터셋을 활용하여 프라이버시 보호 대화 시스템을 개발하였다. 데이터셋의 전처리, 민감 정보 탐지 및 마스킹, 강화학습을 통한 최적 균형점 탐색, 추천 시스템을 통한 대체 응답 생성 과정을 수행하였다. 이를 통해 사용자 프라이버시를 효과적으로 보호하면서도 자연스러운 대화를 제공할 수 있는 대화 시스템을 구현하였다. 이러한 접근은 다양한 도메인에서 적용 가능하며, AI 시스템의 보안성과 윤리성을 강화하는 데 기여할 수 있다.

참고 문헌

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

3. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

4. Li, Y., & Jurafsky, D. (2016). Neural net models for open-domain discourse coherence. arXiv preprint arXiv:1606.01545.

5. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.