사이버범죄 수사 지원을 위한 온프레미스 LLM 기반 비정형 텍스트 핵심 정보 추출

윤병휘[1], 박현호[2,3], 김지온[1,*]

[1]한림대학교, [2]한국전자통신연구원, [3]과학기술연합대학원대학교

pathofborealis04@gmail.com, hyunhopark@etri.re.kr, jion972@hallym.ac.kr

On-Premise LLM-Based Key Information Extraction from Unstructured Text for Cybercrime Investigation Support

Byung Hui Yoon[1], Hyunho Park[2,3], Ji-On Kim[1,*]

[1]Hallym University, [2]Electronics and Telecommunications Research Institute, [3]University of Science and Technology

요 약

사이버범죄 수사를 위해서는 수사 과정에서 확보한 비정형 텍스트(예: 카카오톡 대화)로부터 피해자 이름, 피해 금액 등 핵심 정보를 정확하고 신속하게 추출할 필요가 있다. LLM(Large Language Model)은 비정형 텍스트에서 핵심 정보를 정확하고 신속하게 추출할 수 있으나, 피해자 이름과 피해 금액 같은 개인정보는 유출 위험이 존재하므로 클라우드 기반 LLM(예: ChatGPT)을 활용하는 데에는 한계가 있다. 이에 본 연구는 외부 클라우드 연결 없이 활용 가능한 온프레미스(On-Premise) LLM인 EXAONE-3.5-7.8B-Instruct를 기반으로 한 핵심 정보 추출 방법을 제안한다. 또한, 본 연구에서 핵심 정보 추출 정확도 향상시킬 수 있는 프롬프트 기법을 적용하고, LLM 결과 후처리를 위한 정규표현식을 설계하였다. 본 연구에서 제안한 방법을 사이버범죄 관련 카카오톡 대화 모의 텍스트에 적용한 결과, 약 84%의 F1-score를 달성하였다. 본 연구의 핵심 정보 추출 기법은 개인정보 유출 우려 없이 사이버범죄 수사의 신속한 진행에 기여할 수 있다.

Ⅰ. 서 론

사이버범죄 수사에서 범죄의 상황을 파악하는 핵심 정보의 추출은 필수적이다. 경찰청에서 공개한 자료에 따르면 국내 사이버범죄는 2022년 217,807건에서 2023년 약 241,842건으로 증가하였고, 2023년 사이버범죄 241,842건에서 사이버사기가 전체 230,355건 중 155,715건(64.39%)을 차지하였고[1], 사이버사기의 피해액은 2022년 1조 8천억 원에 달했다[2]. 사이버사기와 같은 사이버범죄에서는 범죄자와 피해자의 관계와 범행 자금 흐름을 빠르게 식별할 수 있는 피해자의 이름과 피해 금액과 같은 핵심 정보를 신속하게 파악하는 것이 중요하다. 그러나, 수사 과정에서 확보하는 대부분의 단서는 범죄자와 피해자 사이의 카카오톡이나 문자 대화와 같은 단서 데이터들은 대부분 자연어 형태의 비정형 텍스트이기에 정형 데이터처럼 규칙 기반의 분석을 통한 핵심 정보 추출은 어렵다. 또한, 피해자가 100명 이상인 경우 수사관이 수작업으로 비정형 텍스트에서 핵심 정보를 일일이 추출하기에 상당한 시간이 소요되어 수사 진행에 부담이 될 것이다.

LLM(Large Language Model)을 활용하여 정확하고 신속하게 비정형 텍스트의 핵심 정보를 추출하는 방안에 관한 연구가 진행되었다. 연구 [3]에서 ChatGPT-3.5를 이용하여 병리보고서에서의 종양의 크기, 특징, 림프절의 침범여부 등의 정보를 추출하는 방안을 제안하였다. 그러나, ChatGPT-3.5와 같은 클라우드 기반 LLM을 통한 정보 추출은 범죄 수사와 같이 개인정보 유출에 민감한 분야에서는 적용에 한계가 있다.

본 연구는 온프레미스(On-Premise) LLM을 활용한 수사 관련 비정형 텍스트로부터 핵심 정보를 추출하는 방안을 제안한다. 온프레미스 환경은 모델과 데이터가 기관 내부 인프라에서만 운용되므로 외부 네트워크 연결이 차단되어 있어, 개인정보 및 수사 기밀 정보가 외부로 유출될 가능성을 최소화할 수 있다는 점에서 클라우드 기반 LLM보다 보안성이 우수하다. 온프레미스 LLM으로는 한국어에 대한 이해와 추론 성능이 우수한 EXAONE-3.5-7.8B-Instruct 모델을 활용하였다[4]. 핵심 정보 추출 정확도 향상을 위해 LLM의 성능을 개선시킬 수 있는 프롬프트 작성 원칙[5]에 따라 프롬프트를 설계하였다. 또한, EXAONE-3.5-7.8B-Instruct의 출력에서 핵심정보를 정제할 수 있는 후처리 방안도 설계하여 개발하였다. 본 연구에서 제안하는 온프레미스 LLM 기반 비정형 텍스트 핵심 정보 추출 기법은 사이버범죄 관련 모의 텍스트 데이터에서 약 84%의 F1-score를 기록하며 준수한 성능을 보였다. 이러한 결과는 피해자가 많은 범죄에서 핵심 정보 추출 시간을 단축 시켜 수사 진행을 지원할 수 있음을 시사한다.

Ⅱ. 온프레미스 LLM 기반 비정형 택스트의 핵심 정보 추출

온프레미스 LLM 기반 비정형 텍스트 핵심 정보 추출 방법을 설명한다. 그림 1은 온프레미스 LLM 기반 비정형 텍스트로부터 사이버 범죄의 핵심 정보 추출하는 절차를 보여준다. 사이버범죄 핵심 정보를 추출하는 절차는 크게 단계 1 “핵심 정보 추출”과 단계 2 “핵심 정보 정제”로 구분된다. 단계 1에서는 핵심 정보 추출을 위해 설계된 프롬프트를 온프레미스 LLM에 입력하여 핵심 정보를 추론하고, 그 결과를 포함한 JSON 파일을 생성한다. 이후 단계 2에서는 정규 표현식 기반 후처리를 적용하여 불필요한 정보를 제거함으로써 핵심 정보 추출 성능을 개선할 수 있다.

본 연구에서는 비정형 텍스트로부터 핵심 정보를 추출하기 위해 온프레미스 LLM 모델로 EXAONE-3.5-7.8B-Instruct 모델[4]을 활용하였다. EXAONE-3.5-7.8B-Instruct 모델이 한국어에 특화된 대규모 언어 모델

* : 교신저자

로서, 복잡한 표현과 다양한 형태로 나타나는 한국어 텍스트를 정확하게 해석하고 의미 단위를 효과적으로 구분할 수 있다. 그래서, 비정형 텍스트에서 핵심 정보 추출 과정에서 높은 신뢰성을 제공할 수 있다.

또한, 본 연구에서는 사이버사기 중 작업 대출 사기의 시나리오를 바탕으로 사이버범죄 전문가와의 자문을 통해 만들어진 모의 카카오톡 대화 100건의 텍스트 데이터를 사용하였다. 모의 카카오톡 대화 텍스트의 시나리오는 범죄자가 SNS로 신용 불량자와 같이 대출이 어려운 대출 희망자를 모집하여 대출 희망자들의 서류를 위조하여 제 2 금융권으로 대출을 받은 뒤, 대출 금액을 갈취한 이력을 다루고 있다. 모의 카카오톡 대화 텍스트에서 “대출일시”, “대출자”, “대출금액” 등을 중심으로 주요 핵심 정보를 추출한다면, 작업 대출 사기의 현황을 신속하게 파악할 수 있다.

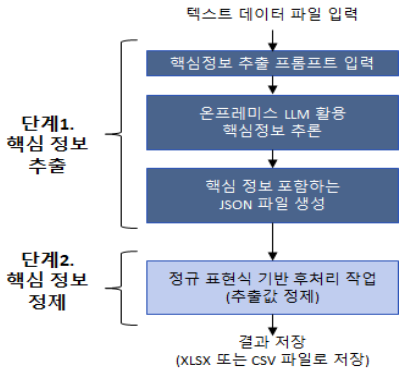

그림 1. 온프레미스 LLM 기반 사이버범죄 핵심 정보 추출 절차

단계 1에서 비정형 텍스트를 주면 온프레미스 LLM에 핵심 정보 추출을 위한 프롬프트가 입력되고, 온프레미스 LLM 모델은 프롬프트를 따라 핵심 정보를 포함하는 JSON 파일을 생성한다. 프롬프트는 연구 [5]에서 제시한 LLM 답변 품질 향상을 위한 지시문 작성 원칙을 참고하여 설계하였다. 프롬프트에서 EXAONE-3.5-7.8B-Instruct 모델의 역할을 기업용 비정형 한국어 텍스트 전용 추출 엔진으로 설정하고, 단계적 추론을 수행하도록 지시하였다. 또한 프롬프트에는 작업 대출 사기 시나리오와 관련된 “대출일시”, “대출자”, “대출금액” 등을 포함한 10개 필드로 구성된 핵심 정보 추출 스키마를 제시하고, 이를 단일 JSON 객체로 변환하도록 하였다. 이때 모델이 특정 필드의 핵심 정보를 식별하지 못하거나 추출된 값이 모호하다고 판단될 경우, 해당 값을 null로 표기하도록 명시하였다. 프롬프트를 따라 EXAONE-3.5-7.8B-Instruct 모델은 핵심 정보를 추론하여 JSON 파일 형식으로 출력한다.

단계 2에서는 그림 2와 같이 정규 표현식을 활용하여 추론 결과에서 핵심 정보를 정제하였다. 정보 추출 과정에서 모델이 생성한 값 중 일부는 준비된 정답과 일치했으나, 동시에 불필요한 값(garbage value)이 포함되거나 날짜 표기 방식의 불일치, 공백 미제거 등의 문제가 확인되었다. 이를 해결하기 위해 정규 표현식을 적용하여 핵심 정보의 정규화 과정을 수행하였다. 그림 2는 “대출일시”를 “10월 14일”과 같이 월과 일을 기준으로 표준화하는 정규 표현식의 예시를 보여준다.

```
_DATE_SLASH_RE = re.compile(r'(?<!\d)(\d{1,2})\s*[./]\s*(\d{1,2})(?!\d)')
_DATE_MONTHDAY_RE = re.compile(r'(?<!\d)(\d{1,2})\s*월\s*(\d{1,2})\s*일?')
_PAREN_RE = re.compile(r'\([^)]*\)')
_MULTI_SPACE_RE = re.compile(r'\s+')

def _clean_dateish(s: str) -> str:
    """대출일시: '잔금일', '까지' 제거, 10/14 → 10월14일, 공백 정리."""
```

그림 2. 정규표현식을 통한 “대출일시” 정보의 정규화

EXAONE-3.5-7.8B-Instruct 모델을 기반으로 한 작업 대출 사기 카카오톡 대화 텍스트의 핵심 정보 추출 성능은 그림 3에 제시하였다. 데이터 추출 프로그램의 weighted average F1-score는 84.33%로 산출되었다. 특히 “대출일시”, “대출자”, “대출금액”과 같은 주요 정보에서는 각각 87.43%, 85.71%, 93.00%의 F1-score를 기록하여 우수한 성능을 보였다. 이러한 결과는 온프레미스 LLM 기반 핵심 정보 추출이 수사관에게 범죄 현황을 신속히 파악하고 본격적인 수사 준비를 수행하는 데 실질적인 도움을 줄 수 있음을 시사한다.

```
── Per-field Precision / Recall / F1 ──────────────────────────
          대출일시  P: 94.12%  R: 81.63%  F1: 87.43%  (TP:80, FP:5, FN:18, N:100)
          사용기간  P: 96.94%  R: 97.94%  F1: 97.44%  (TP:95, FP:3, FN:2, N:100)
           대출자  P: 85.71%  R: 85.71%  F1: 85.71%  (TP:84, FP:14, FN:14, N:100)
      담보물건지 주소  P: 85.00%  R: 85.00%  F1: 85.00%  (TP:85, FP:15, FN:15, N:100)
          대출금액  P: 93.00%  R: 93.00%  F1: 93.00%  (TP:93, FP:7, FN:7, N:100)
       대출금액(숫자)  P: 87.88%  R: 87.00%  F1: 87.44%  (TP:87, FP:12, FN:13, N:100)
         대출수수료  P: 73.00%  R: 73.00%  F1: 73.00%  (TP:73, FP:27, FN:27, N:100)
          투자자  P: 70.00%  R: 72.92%  F1: 71.43%  (TP:70, FP:30, FN:26, N:100)
            금액  P: 82.00%  R: 83.67%  F1: 82.83%  (TP:82, FP:18, FN:16, N:100)
         금액(숫자)  P: 80.00%  R: 80.81%  F1: 80.40%  (TP:80, FP:20, FN:19, N:100)

Weighted F1:  84.33%  (total samples: 1000)
```

그림 3. 작업 대출 사기 대화 텍스트의 핵심 정보 추출 성능

Ⅲ. 결론

본 논문에서는 온프레미스 LLM을 기반으로 한 비정형 텍스트 핵심 정보 추출 방법을 제안하였다. 이를 위해 EXAONE-3.5-7.8B-Instruct 모델을 적용하고, 성능 향상을 위한 프롬프트 설계 원칙을 반영하였으며, 정규 표현식을 통해 추출된 정보를 정제하였다. 작업 대출 사기 대화 텍스트를 대상으로 실험한 결과, 약 84%의 F1-score를 기록하며 준수한 성능을 확인할 수 있었다. 제안한 방법은 사이버범죄 수사 초기 단계에서 범죄 상황을 신속히 파악하고 수사 준비를 지원하는 데 기여할 수 있다. 향후 연구에서는 2025년 8월에 공개된 gpt-oss를 포함한 다양한 온프레미스 LLM 모델을 비교·분석하여 최적의 모델을 탐색하고, 더 많은 사이버범죄 텍스트에 적용함으로써 핵심 정보 추출 기법을 고도화할 예정이다.

ACKNOWLEDGMENT

이 논문은 25년도 정부(경찰청)*의 재원으로 과학치안진흥센터 사이버범죄 수사단서 통합분석 및 추론시스템 개발 사업의 지원을 받아 수행된 연구임(No. RS-2025-02218280)

참 고 문 헌

[1] 김영명, “경찰청 사이버수사 인력 증가, 사이버범죄 건수·검거율은 감소,” 보안뉴스, 2024. [Online]. Available: https://www.boannews.com/media/view.asp?idx=133306

[2] 김유리, “작년 사이버사기 피해 1조8천억원…4년새 8배↑,” 한국세정신문, 2024. [Online]. Available: https://www.taxtimes.co.kr/news/article.html?no=26621

[3] J. Huang, et al., “A critical assessment of using ChatGPT for extracting structured data from clinical notes,” NPJ Digital Medicine, vol. 7, no. 1, p. 106, 2024.

[4] LG AI Research, “EXAONE 3.5-7.8B-Instruct,” Hugging Face, 2024. [Online]. Available: https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct

[5] S. M. Bsharat, A. Myrzakhan, and Z. Shen, “Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4,” arXiv preprint, arXiv:2312.16171, Jan. 2024.