LLM 반복적 개선 기법을 활용한 법률 문서 개인정보 비식별화 연구

김수연, 민병석
세종대학교 데이터 사이언스학과, 세종대학교 인공지능 데이터사이언스학과
rlatndusgu@sju.ac.kr, bmin@sejong.ac.kr

Personal Information De-Identification in Legal Documents using LLM-based Iterative Refinement

Soo yeon Kim, Byungseok Min
Sejong University.

요 약

본 논문은 대규모 언어모델(LLM)을 활용한 법률 문서 개인정보 비식별화 방법을 제안한다. 기존 규칙 기반 및 NER 모델의 문맥적 한계를 극복하기 위해 프롬프트 엔지니어링, LoRA 기반 파인튜닝, 그리고 Dual-Refine 반복적 개선 기법을 결합한 프레임워크를 구축하였다. Llama3, Qwen2.5 두 가지 모델에 대한 실험 결과, 템플릿 구조 프롬프트와 few-shot 학습의 조합이 가장 효과적이었으며, LoRA 파인튜닝을 통해 WER을 약 12% 감소시키고 BERTScore를 약 7% 향상시켰다. 제안된 Dual-Refine 기법은 단일 모델의 편향을 완화하여 더욱 견고한 비식별화 성능을 달성하였다.

Ⅰ. 서 론

법률 문서는 판결문, 계약서, 소송 기록 등 다양한 형태로 작성되며, 개인의 신원과 관련된 민감한 정보를 다수 포함하고 있다. 이러한 문서가 공개되거나 연구 및 분석 목적으로 활용될 경우, 개인정보 보호법 및 GDPR과 같은 관련 법적 규제를 준수하기 위해 개인정보 비식별화(de-identification)는 필수적인 전처리 과정이다. 법률 문서 비식별화는 이름, 주소, 주민등록번호, 전화번호 등 개인식별정보(Personally Identifiable Information, PII)를 탐지하여 제거하거나 대체하는 작업으로, 단순한 패턴 인식만으로는 해결이 어려운 문맥적 이해를 요구한다.

초기 비식별화 연구는 주로 정규표현식 및 사전 기반 매칭을 활용한 규칙 기반(rule-based) 접근법에 의존하였다. 이러한 방법은 전화번호나 주민등록번호와 같이 형식이 명확한 개인정보에 대해서는 효과적이나, “피고인 김철수는”과 같이 문맥에 따라 의미가 결정되는 인명 정보나 역할 기반 표현을 정확히 식별하는 데에는 한계가 있다.[1]

이러한 한계를 보완하기 위해, 이후에는 딥러닝 기반 Named Entity Recognition(NER) 모델을 활용한 비식별화 연구가 활발히 진행되었다. 특히 BERT 계열의 사전학습 언어모델을 기반으로 한 NER 접근법은 문맥 정보를 활용하여 개인정보를 탐지함으로써 기존 규칙 기반 방법 대비 향상된 성능을 보였다. Thunder-DeID 프레임워크[2]는 한국어 법원 판결문에 특화된 NER 기반 비식별화 시스템을 제안하여 일정 수준의 성능을 입증하였다. 그러나 이러한 방식은 대규모 도메인 특화 학습 데이터 구축에 상당한 비용이 요구되며, 학습 데이터에 포함되지 않은 새로운 유형의 개인정보 표현에 대해서는 일반화 성능이 제한되는 문제를 가진다.

최근 대규모 언어모델(Large Language Model, LLM)의 발전으로 자연어 처리 전반에서 획기적인 성능 향상이 이루어지고 있으며, 개인정보 탐지 및 비식별화 분야에서도 그 활용 가능성이 주목받고 있다. LLM은 방대한 사전학습을 통해 다양한 문맥과 표현을 이해할 수 있으나, 범용 모델을 법률 도메인에 직접 적용할 경우 법률 특유의 용어, 문장 구조, 그리고 의미적 미묘함을 충분히 반영하지 못하는 한계가 존재한다.[1]

이에 본 연구에서는 프롬프트 엔지니어링[3], LoRA 기반 경량 파인튜닝[4], 그리고 Self-Refine 기법 [5]을 확장한 Dual-Refine 반복적 개선 기법을 결합한 법률 문서 비식별화 방법을 제안한다. 제안하는 방법은 제한된 학습 자원 환경에서도 효과적으로 적용 가능하도록 설계되었으며, 반복적 자기 검증과 모델 간 상호 검증을 통해 비식별화 결과의 정확성과 견고성을 동시에 향상시키는 것을 목표로 한다.

Ⅱ. 본론

1) 제안 비식별화 시스템

본 연구에서는 LLM 기반 법률 문서 비식별화 성능을 향상시키기 위해 단계별 접근법을 제안한다. 우선, 모델이 법률 문서의 특수성을 이해하고 비식별화 규칙을 준수할 수 있도록 프롬프트 구조 최적화[3]를 수행하였다.일반적/구조적 프롬프트별 Zero-shot, 1-shot, Few-shot 등 총 6가지 변형 실험을 진행한 결과, 자연스러운 지시어와 예시를 조합한 '일반적 프롬프트 + Few-shot' 및 '모델별 템플릿 구조 + Few-shot' 조합을 통해 최적 성능을 도출한다. 특히 Few-shot 방식은 모델에게 구체적인 비식별화 사례와 출력 형태 예시를 제공함으로써 프롬프트 엔지니어링 측면에서 Zero-shot 대비 유의미한 성능 향상을 달성하도록 한다.

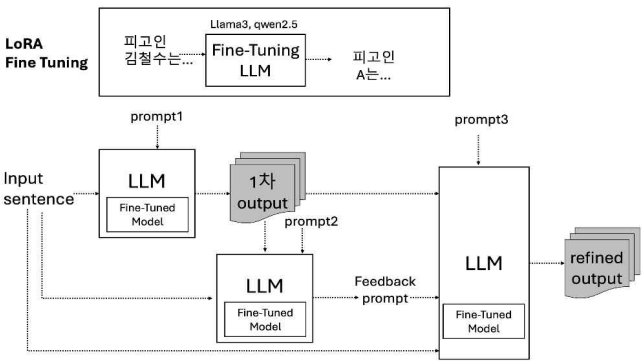

그림 1 제안 비식별화 시스템

프롬프트 엔지니어링을 통한 기초 성능 확보 이후, 법률 도메인에 특화된 비식별화 지식을 모델 내부에 직접 주입하기 위해 모델 파인튜닝을 수행하였다. 수십억 개의 파라미터를 가진 모델을 전체 파인튜닝하는 것은 막대한 컴퓨팅 자원을 요구하며 모델의 범용적 지식을 훼손할 위험이 있으므로, 본 연구에서는 LoRA(Low-Rank Adaptation) 기법[4]을 적용하여 효율적인 학습을 도모하였다. 모델의 가중치를 고정시킨 상태에서 저차원의 어댑터 행렬만을 학습함으로써, 법률 문서 특유의 비식별화 규칙을 정교하게 반영하는 동시에 베이스 모델의 추론 능력을 온전히 보존하고자 하였다. 공개된 약 4,000개의 법률 문장 데이터[2]를 활용하였으며, 각 데이터는 개인정보가 포함된 원문과 비식별화 처리가 완료된 결과 쌍으로 구성되었다. 본 연구에서는 단일 모델 추론의 한계를 극복하고 성능을 추가적으로 개선하기 위해 Dual-Refine 기법을 제안한다. 이는 모델이 자신의 출력을 스스로 평가하고 개선하는 Self-Refine 방법론[5]을 확장한 것으로, 두 모델이 생성한 비식별화 결과를 상호 교차 검증하고 수정하는 프로세스를 포함한다. 이를 통해 단일 모델이 가질 수 있는 편향성을 완화하고, 더욱 견고하며 신뢰도 높은 비식별화 결과물을 도출하고자 하였다.

2) 실험 구성 및 평가 메트릭

본 연구의 실험은 NVIDIA RTX A6000 GPU 환경에서 수행되었으며, AdamW 옵티마이저와 2e-4의 학습률을 적용하였다. 학습의 효율성 및 최적의 일반화 성능 확보를 위해 Callback 함수 기반의 자동 조기 종료(Early Stopping) 메커니즘을 도입하여 Validation Loss가 3회 연속 개선되지 않을 경우 학습을 중단함으로써 과적합을 방지하였다. 연산은 Bfloat16 정밀도로 수행되었으며, 실질적으로 16의 유효 배치 사이즈를 확보하여 학습의 안정성을 달성하였다.실험 대상으로는 모델 파라미터 크기에 따른 성능 변화를 분석하기 위해 Llama-3.2-3B, Llama-3.1-8B, Qwen2.5-3B, Qwen2.5-7B 모델을 활용하였다. 그리고 모델의 비식별화 성능은 다음의 세 가지 객관적 지표를 통해 검증되었다.

$$WER = \frac{\text{대체된 단어 수} + \text{삭제된 단어 수} + \text{삽입된 단어 수}}{\text{정답 문장의 총 단어 수}}$$

***WER* (Word Error Rate):** 모델 응답과 정답(GT) 간의 단어 수준 차이를 측정하며, 값이 낮을수록 원문 구조 보존력이 우수함을 의미한다.

$$R_{BERT} = \frac{1}{|Y|}\sum_{y \in Y} \max_{\hat{y} \in \hat{Y}} v_y \top v_{\hat{y}}$$

***BERTScore*:** 정답 문장(Y)과 생성 문장($\hat{Y}$) 간의 토큰 임베딩 벡터($v_y, v_{\hat{y}}$) 유사도를 측정하여 의미적·문맥적 유사성을 평가한다.

$$Recall_{leakage} = \frac{\left| e \in L_{ppi} \mid e \notin \hat{Y} \right|}{\left| L_{ppi} \right|}$$

Leakage Recall: 입력 문장에는 존재하나 정답에서는 기호화된 개인정보 요소(L_{ppi})가 생성 문장($\hat{Y}$)에서 얼마나 제거되었는지 측정하여 유출 위험을 평가한다.

다만, WER이나 BERTScore 수치가 낮아 문장의 구조적·의미적 정밀도가 훼손된 상태에서의 높은 Leakage Recall은 비식별화의 성공보다는 단순 정보 소실에 가깝다고 볼 수 있다. 따라서 문서의 실질적 활용 가치와 정보 보안성이라는 두 가지 목표를 모두 충족했는지 판단하기 위해, 위 세 지표를 유기적으로 연계하여 고려하는 통합적 성능 평가가 필수적이다.

3) 실험 결과 분석

본 연구에서는 Llama-3와 Qwen-2.5 모델을 대상으로 모델 체급 및 학습 전략에 따른 비식별화 성능을 분석하였다. 실험 결과, 사전 학습된 베이스 모델을 그대로 사용하는 방식보다 파인튜닝을 적용했을 때 구조적 보존력과 문맥 유사도 면에서 뚜렷한 성능 향상이 관찰되었으며, 모델의 파라미터 규모가 클수록 법률 문서 비식별화 작업에 최적화된 결과가 도출됨을 확인하였다.

8B 모델에서의 템플릿 구조 및 Few-shot 최적화

본격적인 학습에 앞서, 동일한 모델 환경에서 프롬프트의 구조적 차이가 비식별화 정밀도에 미치는 영향을 분석하였다. [표 1]에서 확인할 수 있듯이, 단순 지시문만을 활용한 Zero-shot 방식보다 구체적인 예시를 포함한 Few-shot 형태에서 전반적인 지표의 개선이 나타났다.

Few-shot 프롬프팅은 모델에게 비식별화의 대상과 출력 형식을 명확히 가이드함으로써, 법률 문서 특유의 복잡한 문맥 속에서도 일관된 결과물을 생성하는 데 기여하였다. 특히 모델이 임의로 문장을 요약하거나 형식을 파괴하는 현상을 억제하여, 원문의 언어적 틀을 유지하면서도 필요한 정보만을 정확히 치환하는 기초적인 성능을 확보할 수 있었다.

8B	Base 템플릿 구조+zero shot			Base 템플릿 구조+few shot		
	WER	Bert	recall	WER	Bert	recall
llama3	0.996	0.206	0.999	0.324	0.87	0.887
qwen2.5	0.368	0.800	0.804	0.208	0.917	0.800

표 1 few shot 예시 유무에 따른 성능 비교

베이스 모델 대비 파인튜닝의 유의성

모든 실험군에서 베이스 모델에 Few-shot 프롬프팅만을 적용한 것보다 파인튜닝을 거친 모델이 전반적으로 우수한 지표를 기록하는 경향을 보였다. 특히 3B 체급의 Llama-3 모델은 파인튜닝 후 BERT Score가 0.884에서 0.952로 유의미하게 상승하며 문맥 보존 능력이 강화되었음을 입증하였다. 비록 일부 모델의 경우 파인튜닝 과정에서 Recall 수치가 소폭 하락하거나 정체되는 현상이 관찰되기도 하였으나, 모델의 개인정보 식별 변별력이 고도화되는 과정에서 나타난 수렴 결과로 분석된다. 베이스 모델이 문맥과 관계없이 특정 패턴을 과도하게 치환하여 재현율을 높였던 방식(Over-detection)에서 벗어나, 학습된 모델은 법률적 문맥 내에서 식별 정보와 일반 용어를 정교하게 구분하기 시작했다. 이는 보호 대상이 아닌 정보의 오치환을 방지함으로써 문서의 실질적 가치를 보존하는 성과를 거두었으나, 지표상으로는 재현율의 보수적인 수치를 기록하는 원인이 되었다.

8B	Base 템플릿 구조+few shot			LoRA 템플릿 구조+few shot		
	WER	Bert	recall	WER	Bert	recall
llama	0.324	0.87	0.887	0.068	0.986	0.851
qwen	0.208	0.917	0.800	0.082	0.986	0.853

표 2 템플릿 구조 및 파인튜닝 여부에 따른 모델별 성능 비교

모델 파라미터 크기에 따른 성능 차이 (3B vs 8B)

모델의 파라미터 규모가 커짐에 따라 비식별화의 정밀도와 구조적 유지 능력이 동반 상승하는 경향이 나타났다. 8B급 모델들은 3B 모델들보다 일관되게 낮은 WER을 기록하며 원문의 언어적 틀을 효과적으로 보존하였다. 이는 대규모 언어 모델이 보유한 방대한 지식량이 법률 용어의 복잡한 문맥을 파악하고, 불필요한 단어 변형 없이 식별 정보만을 정확히 치환하는 데 긍정적인 영향을 미친 것으로 판단된다.

		Base 일반 구조+few shot			LoRA 일반 구조+few shot		
		WER	Bert	recall	WER	Bert	recall
3B	Llama	0.292	0.884	0.814	0.209	0.952	0.802
	Qwen	0.265	0.891	0.79	0.214	0.929	0.865
8B	Llama	0.321	0.871	0.895	0.200	0.937	0.856
	Qwen	0.225	0.905	0.843	0.188	0.945	0.868

표 3. 모델 종류 및 크기에 따른 기본 모델/파인튜닝 모델 성능 비교

Ⅲ. 결론

본 연구는 LLM 기반 법률 문서 비식별화에서 프롬프트 엔지니어링과 파인튜닝의 효과를 검증하였다. 실험 결과 Few-shot 학습은 Zero-shot 대비 wer 평균 38%의 성능 향상을 보였으며, 파인튜닝은 제한된 데이터(3,600 문장)로도 WER을 최대 12% 감소시켰다. 향후 연구에서는 더 많은 법률 도메인 데이터셋 확보, 실시간 처리를 위한 모델 경량화 연구가 필요하다. 또한 다양한 법률 시스템과 언어에 대한 적용 가능성을 탐구하여 실무 환경에서의 활용도를 높일 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다.

참 고 문 헌

[1] “LEGAL-BERT: The Muppets straight out of Law School,” Chalkidis, I. et al.,EMNLP, 2020.

[2] “Thunder-DeID:Accurate and Efficient De-identification Frame work for Korean Court Judgments,” Hahm, S. et al., EMNLP, 2025.

[3] “Language Models are Few-Shot Learners,” Brown,T.et al., NeurIPS, 2020.

[4] “LoRA: Low-Rank Adaptation of Large Language Models,” Hu, E. J. et al., ICLR, 2022.

[5] “Self-Refine: Iterative Refinement with Self-Feedback,” Madaan, A. et al., NeurIPS, 2023.