형사 판결문 정보추출 데이터셋 구축 방안*
-GPT-3.5 프롬프트 활용을 중심으로-

박예린[O1] 원광재[1] 박노섭*[1]
[1]한림대학교 정보법과학 전공
iisabelle559@gmail.com, lifsai@hallym.ac.kr, rspark@hallym.ac.kr

Building a Dataset for Key Information Extraction from Judgement in Criminal Case: Focusing on the Use of GPT-3.5 Prompt

Yerin Park[O1] Gwang-Jae Won[1] Ro-Seop Park*[1]
[1]Legal Informatics and Forensic Science Major, Hallym University

요 약

판결문의 범죄사실에서 의미 있는 정보를 추출하는 것은 사건을 이해하는 데 중요한 역할을 하며 정보 추출을 위한 인공지능 모델 학습에는 양질의 데이터 셋이 필요하다. 하지만 법률과 같은 특정 도메인에서 학습 데이터 셋을 구축하는 것은 인력, 시간 등 많은 자원이 필요하다. 본 연구에서는 효율적으로 학습데이터 셋을 구축하기 위해 인공지능 모델인 OpneAI GPT-3.5를 활용한다. 특히 GPT-3.5에 Simple 프롬프트를 사용하여 1차 어노테이션을 진행한 후 사람(Human)이 검수하여 더 품질 높은 데이터 셋를 구축할 수 있는 프롬프트를 찾아 나가는 방법(“GPT-3.5&Annotator-in-the-loop”)을 제안한다. 연구 결과, 오차행렬(Confusion Matrix)과 ROUGE-L 지표를 통해 세 가지 유형의 프롬프트 중, 여러 예시를 제시하는 Few-shot 프롬프트의 성능이 가장 좋은 것을 확인하였다. 또한 연구에서 사용한 Human-AI 어노테이션 방식이 데이터 셋 구축에 소모되는 막대한 자원을 90%가량 감축시킬 수 있다는 것을 확인하였다. 본 연구에서 제안한 데이터 셋 구축 방안 및 결과는 법률 분야에서의 효율적인 데이터 셋 확보에 대한 가능성을 보여주었다.

1. 서 론

최근 자연어처리(NLP) 기술의 발전으로 인간이 처리하기 어려운 복잡한 법률 문서를 쉽게 분석하기 위한 연구들이 진행 중이다. 특히 거대 데이터로 학습된 사전학습모델(PLM)의 발전에 따라 도메인 전문가의 정답 데이터 셋으로 미세조정(fine-tuning)하여 다양한 태스크에서 활용하는 연구들이 진행 중이다.

한편, 범죄사건의 범죄사실(crime fact)은 사건의 이해를 돕고 판결의 기초가 되는 중요한 요소이다. 수사관은 증거수집 및 사건 파악 후에 범죄 구성요건에 맞춰 범죄사실을 작성하고, 법관은 쟁점 판단에 있어 범죄사실을 기반으로 최종 판결을 내린다. 따라서 범죄사실에는 사건을 이해하는데 중요한 정보가 포함되어 있다.

이러한 범죄사실에서 의미 있는 정보를 추출한 연구들이 있다. [1]은 법률 문서에서 추출할 수 있는 56개의 개체명 데이터 셋을 만들어 학습하였고, 인식 및 분류에 대해 F-1 micro 98%의 성능을 내었다. [2]는 판결문 데이터에서 주체, 일시, 장소 등과 같은 기본 정보 뿐 아니라 방법, 동기, 행위, 결과 등의 범행 정보를 추출하여 타임라인을 구축하는 방안을 제시하였다. 이러한 연구를 바탕으로 본 연구에서는 형사 1심 살인 판결문의 범죄사실에서 사건 정보를 추출하여 데이터 셋을 구축하는 방안을 제안하고자 한다.

학습 데이터 셋 구축에는 시간, 비용, 인력 등 막대한 자원이 필요하며 법률과 같은 전문 분야는 더 많은 자원이 소모된다. 이 문제를 해결하기 위해 사전학습모델을 사용할 수 있다. [3]은 GPT-3 API를 사용해서 한정된 예산으로 데이터 어노테이션의 어려움을 해결하였다. 이 연구에서는 사람과 GPT-3를 함께 활용하여 효율적으로 데이터 셋을 구축하였다.

문서 분류 작업에 대해서 훈련된 어노테이터 혹은 크라우드 소싱보다 GPT-3.5가 정확도 측면과 비용적 측면에서 더 우수하다는 것을 확인하였다[4]. [5]은 로스쿨 시험 문제를 사용하여 ChatGPT의 성능을 검증하였고, 좋은 결과를 얻기 위한 프롬프트를 제공하였다.

본 연구에서는 OpeanAI의 GPT-3.5를 활용하여 대량의 정보추출 학습 데이터 셋을 구축하고자 한다. 따라서 여러 유형의 프롬프트를 비교하여 정확도 높은 결과를 도출하는 최적의 프롬프트를 찾아 양질의 학습 데이터 셋을 구축하는 것이 본 연구의 목적이다.

2. 방 법

2.1 범죄사실 정보추출 항목 구축

본 연구는 형사 1심 살인 판결문 210건을 사용하며, 판결문의 범죄사실에서 사건의 기본정보 및 범행정보를 파악할 할 수 있는 17개의 정보를 추출한다.

표 1 범죄사실 정보추출 항목

단어 항목	예시	구절 항목	예시
피고인 성명	피고인A	범행동기	“피해자를 죽여라“는 환청을 듣고 피해자를 살해하기로 마음먹었다.
피고인 직업	일용직 노동자		
피해자 성명	피해자D		
피해자 나이	35세		
피해자 성별	여		
피해자 직업	미용실 운영		
피고인-피해자 관계	부부관계	공격행위	칼로 찔러
범행주소	서울 00시	공격부위	등 부위
범행장소	주거지	공격횟수	1회
범행일시	2021.12.15	피해자상해	다발성 자창 등
범행도구	망치(길이 30cm)	범행결과	미수에 그쳤다.

데이터의 단위(단어, 구절)에 따라 위의 17개 항목에 대한 어노테이션 작업을 분리하여 진행한다.

2.2 프롬프트 구축

GPT-3.5를 활용한 데이터 셋 구축 연구가 증가함에 따라 성능

* 이 논문은 2021년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [내역사업명: AI 기반 범죄수사 지원 / 연구개발과제번호: PR10-02-000-21]

을 높이는 다양한 방법론이 제시되어왔으나, 궁극적으로 최상의 아웃풋을 내기 위해서는 2차적으로 사람이 검수하는 작업이 필요하다. 따라서 본 연구에서는 GPT-3.5(api 사용)로 1차 어노테이션을 거쳐 나온 데이터 셋에 대해 사람이 2차 어노테이션 및 검수[1]를 진행한다. 그 과정에서 발견된 프롬프트의 문제점을 반영하여 새로운 프롬프트를 만들어내는 "GPT-3.5&Annotator-in-the-loop 방식"을 제안한다.

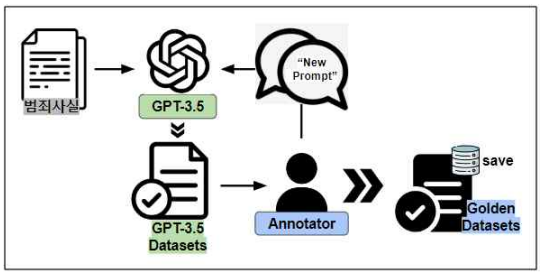

그림 1 "GPT-3.5&Annotator-in-the-loop 방식" 흐름도

본 연구의 핵심은 GPT-3.5의 정확도를 높여 2차 어노테이션 및 검수 시간을 최소화하는 것이다. 따라서 정확도가 가장 높은 프롬프트를 찾기 위해 1) Simple 프롬프트 2) One-shot 프롬프트 3) Few-shot 프롬프트의 성능을 비교한다.

1) Simple 프롬프트: <태스크 정보>와 <태그>가 포함된 프롬프트

표 2 Simple 프롬프트 예시

<태스크 정보>
I want you to extract key information from crime-related text corresponding to tags below:
<태그>
[Tags]
1. motive of crime
2. injury of victim
...(생략)
If you can't find matching information, don't infer anything and simply add "없음".

2) One-shot 프롬프트: <태스크 정보>, <태그 및 예시 한 개>, <조건>이 포함된 프롬프트

표 3 One-shot 프롬프트 예시

<태스크 정보>
I want you to extract key information from crime-related text in Korean. I want you to extract key information Simpled on the criteria I give you. The information extracted should be a phrase/word. Below is Criteria B with some sample data:
<태그 및 예시-한 개>
[Criteria B]
1. motive of crime - e.g., "마음먹고,"
2. injury of victim - e.g., "다발성 절창 등"
...(생략)
<조건>
If you can't find matching information, don't infer anything and simply add "없음.". ✓Tag must be in English and extracted information must be in Korean.✓It is VERY important that you extract the information word by word.✓NEVER summarize, rephrase or translate the given text.
Tag the following text using criteria B:

3) Few-shot 프롬프트: <태스크 정보>, <태그 및 예시 두 개 이상>, <조건>이 포함된 프롬프트

표 4 Few-shot 프롬프트 예시

<태스크 정보>
I want you to extract key information from crime-related text in Korean. I want you to extract key information Simpled on the criteria I give you. The information extracted should be a phrase/word.. Below is Criteria B with some sample data:
<태그 및 예시-두 개 이상>
[Criteria B]
1. motive of crime - e.g., "피해자가 자신을 무시한다는 생각에 격분하여", "피해자가 화를 내면서 피고인의 말을 들어주지 않자 화가 나 살해하기로 마음먹고," 등
2. injury of victim - e.g., "다발성 절창 등", "경부 압박에 의한 질식", "고도의 두부손상" 등
...(생략)
<조건>
If you can't find matching information, don't infer anything and simply add "없음.". ✓Tag must be in English and extracted information must be in Korean.✓It is VERY important that you extract the information word by word.✓NEVER summarize, rephrase or translate the given text.
Tag the following text using criteria B:

위 세 개의 프롬프트를 사용한 결과를 정답 데이터 셋과 비교하여 성능평가를 진행한다. 정답 데이터셋은 형사 1심 살인 판결문 210건을 대상으로 10명의 연구원(경찰대학 소속)이 1차 어노테이션을 진행한 후, 한 명의 검수자가 2차 어노테이션 및 검수 과정을 거쳐 구축한 데이터 셋이다. 성능평가는 항목 별로 ROUGE score 및 F1 score를 계산하는 방식으로 진행된다. 평가 후, 성능이 가장 좋은 프롬프트를 사용하여 1차 어노테이션을 진행하고, 2차 어노테이션 및 검수를 통해 최종 학습 데이터 셋 1,000건(판결문)을 구축한다.

2.3 성능평가

세 개의 프롬프트에 대한 성능지표로는 ROUGE score와 오차행렬(Confusion Matrix)을 사용하였으며, 재현율(recall), 정밀도(precision) 그리고 조화평균인 F1-score를 측정하여 평가하였다.

본 연구에서는 단어 뿐만 아니라 구절 혹은 문장 단위의 정보 또한 추출하기 때문에 텍스트 요약, 기계번역 등 자연어 생성 모델의 성능 평가 지표인 ROUGE와 분류 모델의 성능 평가 지표인 오차행렬(Confusion Matrix)을 모두 사용하였다. 본 연구에서 오차행렬의 통계치를 계산할 때, 모두 정확하게 예측하는 경우는 TP(True Positive) 및 TN(True Negative)으로 카운트하고, 실제값은 없는데 예측값을 생성한 경우는 FP(False Positive), 그리고 실제값의 토큰 중 일부만을 정확하게 예측한 경우나 그 외의 경우는 FN(False Negative)으로 카운트하여 계산하였다.

구절 단위로 추출하는 항목의 경우, 데이터의 단위와 길이를 고려하여 연속적이지 않은 최장 길이의 공통 문자열을 측정하는 방법인 ROUGE-L만을 사용하여 평가를 진행하였다.

각 프롬프트의 성능평가 결과는 다음과 같다.

표 5 단어 단위 정보에 대한 성능평가 결과

Prompt (word unit)	ROUGE-L			Confusion Matrix		
Simple	recall	precision	F1	recall	precision	F1
	59.02%	71.35%	60.13%	25.05%	80.74%	38.23
One-shot	69.89%%	80.64%	71.88%	50.03%	83.48%	62.57%
Few-shot	**75.14%**	**80.45%**	**75.91%**	**50.68%**	**84.42%**	**63.35%**

표 6 구절 단위 정보에 대한 성능평가 결과

Prompt (phrase unit)	ROUGE-L		
Simple	recall	precision	F1
	50.15%	42.87%	40.57%
One-shot	39.81%	62.67%	43.06%
Few-shot	**53.88%**	**69.37%**	**56.50%**

[표 5]와 [표 6]에서 볼 수 있듯 단어 단위 뿐만 아니라 구절 단위의 정보 추출에서 가장 좋은 성능을 보인 프롬프트는 조건과 두 개 이상의 예시를 제공한 Few-shot 프롬프트이며, 가장 낮은 성능을 보인 프롬프트는 조건이나 예시를 전혀 제공하지 않은 Simple 프롬프트인 것을 확인하였다. Few-shot 프롬프트는 단어 추출 태스크에서는 F1-score 75.91%(ROUGE-L), 구절 추출 태스크에서는 56.50%의 성능을 보였다.

[1]의 연구와 결과를 비교했을 때, 해당 연구에서는 개체명 추출 태스크에서 KoELECTRA 모델을 사용한 결과 98%의 성능을 보인 반면, 본 연구에서는 74.91%의 성능을 보였다.

정량적인 지표로 보았을 때는 높은 점수는 아니지만, 해당 프롬프트를 사용하여 1,000건의 범죄사실 정보추출 학습 데이터 셋을 구축함으로써 92%의 시간 절감(65시간 → 5시간) 및 90%의 비용 절감

1) GPT-3.5 결과에 대한 2차 어노테이션 및 검수는 10명의 경찰대학 연구원 및 법률 전문가와 수사 실무자가 진행하였다.

(990,000원 → 99,000원)의 효과를 얻을 수 있었다.

2.4 결과 분석

[표 7]은 Few-shot 프롬프트의 항목 별 성능을 정리한 표이다. 결과적으로 17개의 항목 중 8개의 항목의 ROUGE-L F1-score가 70% 이하인 것을 확인하였고, 해당 항목들에 대한 오추출 분석을 진행하였다.

표 7 Few-shot 프롬프트 결과에 대한 항목 별 성능

Tags	ROUGE-L (f1_avg)	Confusion Matrix (f1_avg)
피고인 성명	87.40%	91.98%
피고인 직업	79.94%	20.00%
피해자 성명	92.91%	92.35%
피해자 나이	95.66%	97.11%
피해자 성별	75.49%	75.60%
피해자 직업	78.56%	17.65%
피고인-피해자 관계	**57.10%**	**42.51%**
범행주소	79.07%	71.29%
범행장소	**41.99%**	**31.67%**
범행일시	70.13%	2.90%
범행도구	76.80%	18.10%
범행동기	**42.95%**	해당없음
공격행위	**27.64%**	해당없음
공격부위	**68.37%**	해당없음
공격횟수	**67.90%**	해당없음
피고인 상해	**63.12%**	해당없음
범행결과	**69.01%**	해당없음

[표 8]은 범행동기, 범행결과, 범행도구 등 8개의 항목에 대한 오추출 유형이다. 5가지의 유형은 1) 정보를 그대로 추출하지 않고 요약하는 “요약” 문제, 2) 말의 어미를 바꾸는 “의역” 문제, 3) 데이터에는 없는 정보를 생성해내는 “생성” 문제, 4) 범위(span)를 잘 못 잡는 “범위” 문제, 그리고 5) 여러 개의 정보를 일부 잡아내지 못 하는 “누락” 문제로 분류할 수 있다. 가장 많이 발생하는 오추출 유형은 발생률 50% 이상을 차지하는 “범위” 문제이고, 그 다음으로는 30% 이상을 차지하는 “누락” 문제이다. 그 외에 “생성” / “요약” / “의역” 문제는 모두 10% 이하로, 오추출 유형 중 상대적으로 낮은 발생률을 보였다.

표 8 오추출에 대한 및 유형 분류

문제 유형	구절 추출 태스크		단어 추출 태스크	
	정답셋	GPT-3.5	정답셋	GPT-3.5
요약 / 의역 (3%)	motive of crime		relation between accused and victim	
	“피해자에게 반말을 하지 말라고 하였고, 피해자가 이에 전혀 응하지 아니하자, 칼로 피해자를 살해할 것을 마음먹고“	*“반말에 대한 불만 “*	*“같은 마을에 살면서 오래전부터 알고 지내던“*	*“알고 지내던 마을 사람“*
	attack act		Nan	
	“찌르는 등“	*“찔렀다.“*	*Nan*	*Nan*
범위 (53%)	injury of victim		relation between accused and victim	
	“경부압박 질식“	*“경부압박 질식으로 사망하게 하였다“*	*“중학교 때부터 알고 지내는 친구 사이“*	*“친구 사이“*
생성 (10%)	motive of crime		place type of crime scene	
	“없음“	*“피해자가 자신을 무시한다는 생각에 격분하여“*	*“없음“*	*“아파트“*
누락 (34%)	attack act		place type of crime scene	
	[’힘껏 내리 찍었고’, ’휘두르고 찔렀다.’]	*“휘두르고 찔렀다“*	*[’”주거지“, “앞마당“]*	*”주거지“*

이와 같이 GPT-3.5가 정보추출 태스크를 수행할 때, 흔히 발생하는 “요약”, “의역”, “생성”과 같은 문제는 [6] 연구에서 제시한 방법을 적용하여 해결할 수 있을 것이다. 이 방법은 1) 문장에 번호를 부여 2) 질문에 해당하는 문장의 번호를 선택 3) 선택된 문장에서 정답을 추출하는 세 가지 순서로 태스크를 수행하도록 하여 언어모델이 정보를 왜곡하지 못 하도록 하는 방법이다. 따라서 모든 범죄사실 문장에 번호를 부여한 후, 태그에 해당하는 문장을 선택하고, 선택된 문장에서 구절 및 단어를 추출하는 방식을 적용할 수 있을 것이다. 특히 구절을 추출할 때에는 주어, 서술어, 목적어를 모두 포함하는 조건을 제공하고, 단어를 추출 할 때에는 특정 키워드를 포함하는 조건을 제공한다면 “범위” 문제 또한 해결할 수 있을 것이다.

3. 결 론

본 연구에서는 OpenAI의 GPT-3.5를 활용하여 범죄사실 정보추출 데이터 셋을 효율적으로 구축할 수 있는 “GPT-3.5&Annotator-in-the-loop 방식”을 제안하였다.

이 방식을 통해 Simple 프롬프트를 발전시켜 가장 성능이 좋은 Few-shot 프롬프트를 구축하였고, 해당 프롬프트를 사용하여 최종 학습 데이터 셋을 구축함으로써 90% 이상의 비용과 시간 절감의 효과를 보았다.

이러한 결과는 법률과 같은 전문 분야에서 전문가가 직접 데이터를 생성하는 것이 아닌 검수만 함으로써 시간, 비용 등의 자원을 줄일 수 있는 가능성을 보여준다.

또한 GPT-3.5가 정보 추출 태스크를 수행할 때, 조건과 예시의 유무, 예시의 개수에 영향을 많이 받기 때문에 항목 별로 조건과 예시를 세분화하여 부여하는 방향으로 프롬프트의 정확도를 향상시킬 수 있음을 시사하였다.

더 나아가 개체명 추출 태스크에서 높은 성능을 보인 [1]의 연구와 비교했을 때, [1]에서는 범죄 도메인에 특화된 약 3,000여 건의 대량의 데이터셋을 사전학습모델에 미세조정(fine-tuning)하였기 때문에 성능이 좋았던 것으로 보이므로, 본 연구에서도 사전학습모델에 GPT-3.5와 전문가 검수를 통해 구축된 1,000건의 데이터 셋을 미세조정한다면 더욱 더 정확도 높은 결과를 낼 수 있을 것으로 보인다.

따라서 본 연구에서 구축된 1,000건의 정답셋을 활용하여 범죄 도메인에 특화된 모델을 구축할 예정이며, 향후 해당 모델을 활용하여 범죄사실 정보추출 작업을 자동화할 수 있을 것으로 기대한다.

참 고 문 헌

[1] H. D. Kim, “A Named Entity Recognition Model in Criminal Investigation Domain using Pretrained Language Model”, Journal of The Korea Convergence Society, *13(2)*, pp.13-20, 2022.

[2] Y. N. Lee, “A Study on Extraction of Criminal Information Using Machine Reading Comprehension from Judgment in a Criminal Case”, Master’s Degree, thesis, Dept. International Studies., Hallym Univ., Chuncheon, Gangwon, KR, 2021.

[3] S. Wang, “Want To Reduce Labeling Cost? GPT-3 Can Help”, *arXiv preprint arXiv:2108.13487*, 2021.

[4] Gilardi, F., Alizadeh, M., and Kubli, M. “Chatgpt outperforms crowd-workers for text-annotation tasks.“ *arXiv preprint arXiv:2303.15056,* 2023.

[5] Choi, Jonathan H., et al. “ChatGPT goes to law school”, *Available at SSRN*, 2023.

[6] Creswell. A and Shanahan. M, “Faithful Reasoning Using Large Language Models”, *arXiv preprint arXiv:2208.14271*, 2022.