ChatGPT 를 이용한 형사사건 양형 예측 연구

조민한, 한진영
성균관대학교 인공지능융합학과

zvezda@g.skku.edu, jinyounghan@skku.edu

Term of Penalty Prediction using ChatGPT

Minhan Cho, Jinyoung Han
Dept. of Applied Artificial Intelligence, Sungkyunkwan University

요 약

형량 예측 연구는 법률 인공지능에서 가장 활발히 연구되고 있는 분야 중 하나이며, 비법률전문가의 사법 신뢰도 상승과 법률전문가의 업무 부담 완화에 긍정적 영향을 줄 수 있다. 본 연구는 형사사건의 양형 예측에 ChatGPT 를 접목하여 입력된 사실관계와 유사한 선행 판례를 검색함으로써 형량 예측에 필요한 모델의 훈련 시간과 비용을 절감하는 접근법을 제안한다. 본 모델의 weighted F1-score 는 0.53 으로, 미세조정된 BERT 모델과 유사한 성능을 기록하였다.

1. 서론

법률 인공지능은 과거 수십년 간 '법률 QA', '법률 개체명 인식', '판결서 검색' 등 다양한 태스크에서 발전을 거듭하고 있다. 이 중 '양형 예측'은 법률 인공지능에서 가장 중요하게 여겨지는 분야 중 하나로, 양형 예측에 인공지능이 개입될 시 판사의 개인적 성향과 편견 등이 형량 결정에 영향을 끼침으로써 비슷한 범죄라도 판사에 따라 처벌 수위가 크게 차이나는 현상을 줄일 수 있으며,[1] 2021 년 기준 판사 1 명 당 담당사건이 464 건에 달하는 등 고질적인 판사의 과로 문제를 완화할 수 있을 것으로 기대된다.[2]

한편, 2022 년 이후 ChatGPT 를 시작으로 거대언어모델(Large Language Model, LLM)이 등장함에 따라, 거대언어모델을 법률 인공지능 분야에 접목하는 연구 또한 증가하고 있다. 광범위한 텍스트 데이터를 학습한 거대언어모델의 추론 능력을 법률 인공지능 분야에 이용할 경우, 학습 데이터가 부족한 경우에도 거대언어모델이 사전학습 시 학습한 정보를 바탕으로 적절한 법률적 판단을 추론할 수 있기 때문이다.

따라서 본 연구는 거대언어모델인 ChatGPT 를 이용해 모델의 학습을 최소화하여 국내 형사 사건 7 개 분야에 대해 양형 예측을 수행하는 연구를 제안한다.

2. 관련 연구

LLM 등장 이후, LLM 을 법률에 적용하여 모델의 추가적인 파라미터 학습 없이 'Chain-of-Thought', '법률적 삼단논법(Legal Syllogism)' 접근법을 이용하여 사건의 사실관계를 분석하고, 법리에 적용하는 연구가 진행되었다.[3][4] 양형 예측은 이전 유사 판례의 사실관계와 법리판단을 참조하는 것이 필수적이므로, LLM 만을 사용하기보다는, 판례 데이터베이스를 구축하여 입력된 사실관계와 유사한 선행 판례를 검색해 LLM 이 In-Context Learning 을 통해 학습하는 연구가 이루어졌다.[5][6]

3. 양형 예측 모델 설계

본 연구는 huggingface 에서 제공하는 한국 형사 사건 양형 예측 데이터셋인 'lbox/lbox-open ljp-criminal' 데이터셋을 사용한다. 본 데이터셋은 "강제추행", "공무집행방해", "교통사고처리특례법(치상)", "도로교통법 위반(음주운전)", "사기", "상해", "폭행" 등 7 개 죄목에 해당하는 판례의 '사실관계'와 '양형의 이유', 13 개 구간별로 나뉜 '양형 레이블'로 구성되어있다. 학습 데이터셋 8400 건은 FAISS[1]를 이용하여 유사 판례 검색을 위한 판례 벡터 데이터베이스로 구축하고, 테스트 데이터 928 건에 대해 모델의 결과를 비교한다.

모델은 크게 입력 사실관계의 죄목 예측, 유사 선행 판례 검색 및 반환, 양형 예측으로 이루어진다. 우선, 양형 예측의 대상이 되는 사실관계에 해당하는 죄목을 판단하기 위해, 형사 사건 데이터셋을 대상으로 죄목 예측을 학습시켜 미세조정한 BERT 모델을

[1] https://faiss.ai/index.html

이용하여 입력된 사실관계에 적용될 수 있는 죄목을 예측한다. 이후 예측된 죄목과 동일한 죄목의 선행 판례를 판례 데이터베이스에서 검색하여 반환받는다. 최종적으로, 반환된 선행 판례의 사실관계, 양형의 이유, 양형 레이블을 양형 예측을 수행할 사실관계와 함께 ChatGPT 에 프롬프트로 제시하여 ChatGPT 가 In-Context Learning 으로 모델 파라미터 조정 없이 선행 판례를 학습하고, 주어진 사실관계에 해당하는 양형을 예측하도록 한다.

선행 판례 검색 모델은 sBERT 임베딩, openai 의 'text-embedding-3-small' 임베딩, BM25 알고리즘의 결과를 비교하여 상대적으로 준수한 결과를 기록한 BM25 를 사용하였다. 사용한 GPT api 의 버전은 'GPT-3.5-turbo-0125'이며, 재현성을 위해 temperature 하이퍼파라미터를 0 으로 고정하였다.

4. 실험결과

	3-shot	5-shot	3-shot (facts only)	Finetuned BERT
강제추행	0.57	0.57	0.34	0.55
공무집행방해	0.67	0.63	0.49	0.65
교통사고처리 특례법(치상)	0.40	0.32	0.27	0.46
도로교통법 위반(음주운전)	0.63	0.6	0.57	0.61
사기	0.42	0.33	0.3	0.47
상해	0.37	0.29	0.28	0.56
폭행	0.57	0.41	0.35	0.6
전체	0.53	0.46	0.38	0.58

<표 1> 양형 예측 모델의 평가결과 (weighted F1 score)

표 1 은 형사 사건 양형 예측 모델의 죄목별 weighted F1 score 의 결과를 정리한 것이다. 'shot'은 ChatGPT 에 제시한 유사 선행 판례의 개수를 의미하고, 'facts only'은 선행 판례의 내용 중 '양형의 이유'를 제시하지 않고 오직 '사실관계'만을 ChatGPT 에 입력했음을 의미한다. Finetuned BERT 는 'klue/bert-base' 모델을 전술한 8400 건의 학습데이터에 학습하고, 928 건의 테스트 데이터셋에 예측한 결과이다.

양형 예측 모델의 결과 중 3 건의 유사 선행 판례를 제시한 결과는 weighted F1-score 기준 0.53 이며, 이는 5 건의 유사 선행 판례를 제시한 결과, 3 건의 유사 선행 판례를 제시하되, '양형의 이유'를 생략한 결과보다 높은 수치이다. 이는 선행 판례를 5 건 제시하였을 때 모델이 선행 판례의 노이즈를 효과적으로 배제하지 못하였고, '양형의 이유'를 생략하였을 때 모델이 형량 예측에 필요한 충분한 정보를 제공받지 못했기 때문인 것으로 추측된다. ChatGPT 를 사용한 형량 예측 모델의 결과는 Finetuned BERT 의 결과인 0.58 보다 낮으나, '강제추행', '공무집행방해', '도로교통법위반(음주운전)' 등 항목에서 더 높은 성능을 기록하였으며, 딥러닝 모델의 미세조정에 요구되는 시간과 비용, 데이터량을 고려하였을 때 ChatGPT 를 이용한 형량 예측 접근법은 충분한 경쟁력이 있다고 판단된다.

5. 결론

본 연구에서는 모델의 학습과 비용을 최소화하면서 미세조정을 수행한 딥러닝 모델과 비견되는 성능을 기록한 ChatGPT 이용 형량 예측 모델의 결과를 제시하였으며, 상대적으로 많은 선례를 ChatGPT 에 제시하는 것이 오히려 성능에 악영향을 주고, 형량 예측에 '사실 관계' 뿐만 아니라 '양형의 이유'도 중요한 정보임을 확인하였다. 향후 연구에서는 데이터를 충분히 학습한 딥러닝 모델보다 우수한 성능을 기록하고, 더 다양한 죄목에 대한 형량 예측이 가능한 거대 언어모델 활용 양형 예측 모델 연구를 수행할 것이다.

ACKNOWLEDGEMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단과 (No. 2023R1A2C2007625) 2022 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022S1A5A8054322).

참고문헌

[1] "'재범가능성, 징역 6 년" AI 가 내린 판결 받아들일 수 있을까', https://www.chosun.com/national/court_law/2023/06/26/VULGBNBBJFB5PHLRTZJNXYMLMA/

[2] "판사 1 명당 연간 담당사건 464 건… 독일의 5.17 배", https://www.yna.co.kr/view/AKR20210923083100004

[3] Yu et al., Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13582-13596).

[4] Deng, et al., Syllogistic Reasoning for Legal Judgment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13997-14009).

[5] Shui et al., A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.

[6] Wu et al., Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.