법률 AI의 성능 향상을 위한 DPO 알고리즘 기반 연구 제안

이수린, 김수민, 이흥노*
광주과학기술원

leesurin@gm.gist.ac.kr, smkim6927@gm.gist.ac.kr, heungno@gist.ac.kr*

A study on the Performance Enhancement of Legal AI Using DPO Algorithm

SuRin Lee, Sumin Kim, Heung-No Lee*
Gwangju Institute of Science and Technology (GIST)

요 약

현대의 법률 분야에서의 언어 모델은 데이터 부족, 윤리적 편향, 그리고 공정성 등 다양한 문제에 직면하고 있다. 본 논문은 이러한 문제들을 극복하고 효율적인 Fine-tuning 방법을 제안하기 위해 DPO(Direct Preference Optimization) 알고리즘을 사용하는 방법에 관한 연구를 제안한다. DPO는 기존의 RLHF와는 다른 접근 방식으로, 사용자 선호도에 기반하여 직접 모델 성능을 최적화한다. 법률 분야에 특화된 언어 모델을 구축하기 위해 논문은 Reference-Free Confidence-Based Truthfulness Estimation 방법을 활용하여 선호 데이터셋을 구축한다. 이를 통해 데이터의 다양성과 전문성을 효과적으로 활용해 법률 AI의 성능을 향상시킬 수 있다.

Ⅰ. 서론

법률 분야에서의 언어 모델 개발은 데이터 부족, 윤리적 편향, 그리고 공정성 등의 다양한 문제에 직면하고 있다[1]. 현재까지 발전에도, 기존 언어 모델들은 여전히 RLHF(Reinforcement Learning from Human Feedback) 방법론을 사용하여 Fine-tuning 되어 왔다. 그러나 이 방법은 여전히 구현이 복잡하며 학습이 복잡하며 학습이 불안정하며 전문적인 피드백 데이터 부족 등의 한계를 지니고 있다[2]. 본 연구에서는 이러한 문제들을 극복하고 법률 AI의 성능을 향상시키기 위해 Direct Preference Optimization(DPO) 알고리즘을 활용하여 법률 AI를 선호도 데이터에 기반하여 개발하는 방법을 탐구한다.

이 연구의 목표는 기존의 한계를 극복하고 사람의 개입 없이 선호도 데이터 셋을 구성함으로써, 높은 사실성과 성능을 갖춘 Legal AI를 개발하는 것이다. 제안된 방법을 통해 법률 AI의 개발에 새로운 가능성을 제시하고자 한다.

Ⅱ. 본론

법률 분야에서의 AI는 다양한 모델의 개발이 진행되고 있으며, 특히 대형 언어 모델(LLMs)이 주목받고 있다[1]. 법률 AI는 주로 법률 문서 분석, 관련 문서 및 계약 생성, 법률 상담 제공 등의 다양한 기능을 수행하여 법적 의사 결정을 지원한다. 그러나 현재 법률 AI 분야에서는 발전과 함께 여러 도전 과제가 나타나고 있다. 특히, 데이터셋의 결함과 알고리즘 모델의 단점, 전통적인 법률 산업에 미치는 영향 그리고 특정 사법 실무에서 발생하는 문제들이 주요한 이슈로 드러나고 있다. 이에 대한 고찰이 모델 알고리즘의 복잡한 구조와 가치 중립성의 어려움으로 해석을 잘하지 못하며, 결과적으로 법률 AI 모델의 결정 과정과 근거를 명확하게 이해하기 어렵게 만든다. 추가로, 알고리즘이 부정적인 편견을 포함하거나 특정 그룹에 편향되어 있을 때 이는 윤리적인 문제를 초래할 수 있다. 끝으로, 기존 모델들은 최적성을 얻기 어려워 알고리즘 최적화가 필요하다. 이러한 결함을 극복하기 위해 DPO(Direct Preference Optimization)같은 새로운 접근 방식이 필요해 보인다.

기존의 OpenAI의 GPT-3.5와 Meta의 LLaMa-2-Chat 등의 모델들은 Fine-tuning을 위해 RLHF(Reinforcement Learning from Human Feedback) 방법론을 기반으로 학습되었다. 그러나 RLHF는 구현이 복잡하며 학습이 불안정하고 많은 GPU 자원을 필요로 한다는 점에서 한계가 있다[2]. 또한, 법률 분야에 적용 시 전문적인 피드백 데이터의 부족과 함께 사람의 편향된 피드백이 모델 학습에 영향을 미칠 수 있다는 우려가 있다[1]. 현대의 법률 분야에서의 정확하고 특화된 언어 모델을 만들기 위해서는 더 효율적이고 간편한 Fine-tuning 방법이 필요하다. 이러한 배경 속에서 DPO(Direct Preference Optimization)와 같은 새로운 방법론들이 중요성을 가진다. DPO는 기존의 RLHF와는 다른 접근 방식을 제안한다[3].

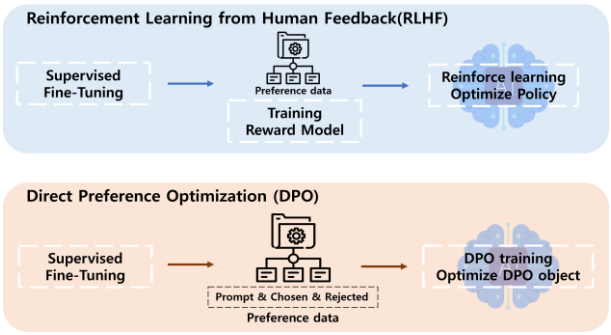

그림 1. RLHF와 DPO의 Pipeline

그림 1과 같이RLHF는 선호 데이터를 리워드 모델이 학습하고, 강화 학습을 통해 최적화하는 단계를 거친다.

반면, DPO는 이 단계를 DPO training으로 간소화한다. 다시 말해, DPO는 RLHF 사용되는 Reward model 학습 및 정책 최적화를 우회하며, 대신 사용자 선호도에 기반하여 직접 성능을 최적화하는 방법을 제공한다. DPO의 선호데이터셋은 사용자의 질문(Prompt), 모델이 선택하는 선호하는 답변(Chosen), 그리고 모델이 비선호하는 답변(Rejected)으로 구성된다.

DPO 알고리즘은 선호 데이터를 직접 학습하므로 선호 데이터셋의 구축은 매우 중요하다. DPO 알고리즘을 적용하기 전, 선호 데이터 셋을 구축하기 위한 효과적인 방법의 하나는 [4]에서 제안한 Reference-Free Confidence-Based Truthfulness Estimation 방법이다. 이 방법은 LLM이 생성한 답에 대한 확신과 해당 답이 올바른지 여부 간의 강력한 상관관계를 활용한다. 구체적으로, GPT-3.5로 생성된 텍스트에서 핵심 주장이나 문장을 추출하고, 각 주장을 특정 사실에 대한 지식을 테스트하는 질문으로 변환한다. 그 후, 각 주장에 대해 몇 차례 응답을 다시 추출하고, 이에 대한 응답을 일치 그룹에 따라 분류하여 가장 큰 그룹에 속하는 응답의 비율을 최종 진실성 점수로 사용한다. 이때, 모델의 확신 점수를 사용하여 응답을 분류한다. 이 방식을 통해서 사람의 개입 없이 선호 데이터셋을 만들 수 있다. 외부 지식을 사용하지 않고, 기존 LLM 모델의 확신을 활용하여 선호 데이터 셋 생성 방법을 제안한다.

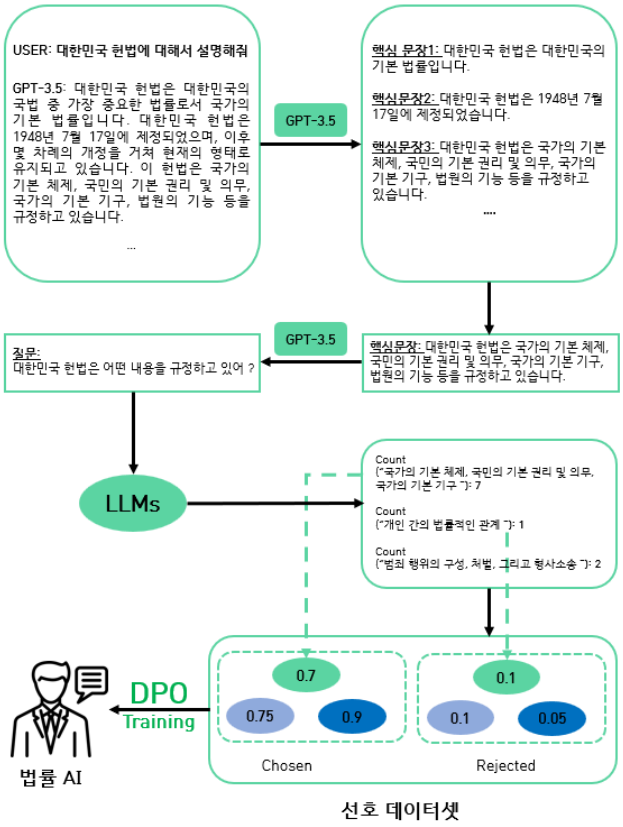

그림 2. 법률 선호 데이터셋 구축 기반 DPO 방법

본 연구는 앞서 언급한 방법들을 통해 사용자의 개입 없이도 법률 AI의 성능을 효과적으로 향상시킬 수 있는 방법을 제시한다. 법률 AI 모델의 성능 향상을 위해 Fine-tuning 단계에서 DPO 알고리즘과 Reference-Free Confidence-Based Truthfulness Estimation 방법을 도입한다. 위의 그림과 같이 GPT-3.5 활용하여 특정 법률 관련 질문 또는 프롬프트에 대한 답변을 생성한다. 예를 들어, “대한민국 헌법에 대해서 설명해봐” 라는 프롬프트에 대해 여러 후보 응답을 생성한 후, 이를 다시 질문의 형태로 바꾼다. LLM이 해당 질문에 대한 응답을 생성하고, 몇 차례 응답 후의 결과를 선호 데이터 형성에 사용한다. Chosen 데이터 셋은 기존 모델이 선택한 답변 중 가장 큰 그룹을 선택하여 형성되며, Rejected 데이터 셋은 모델이 선택하지 않은 다양한 답변 중 가장 확률이 낮은 그룹을 선택한다. 예를 들어, “국가의 기본 체계, 국민의 기본 권리 및 의무, 국가의 기본 기구”라는 답변이 Chosen 데이터 셋에 속하고, “개인 간의 법률적인 관계” 라는 답변이 Rejected 데이터 셋에 속한다. DPO는 구축된 선호 데이터 셋을 바탕으로 사용자가 더 선호하는 응답을 선택하는 방식으로 데이터를 형성하며, 모델이 Fine-tuning 되면서 선호도가 높은 응답을 더 많이 생성한다. 결과적으로 DPO를 통해 법률 AI는 선호 데이터에 기반하여 학습하고 성능이 향상된다. 또한, 데이터셋에 편향이나 차별이 없도록 지속적으로 모니터링하고 조정함으로써 DPO는 선호 데이터의 공정성을 보장한다. 이를 통해 DPO는 간소한 구현과 높은 학습 효율성을 가지며, 특히 법률 관련 언어 모델에 적용 시 유용한 대안이 될 것으로 기대된다.

Ⅲ. 결론

본 논문에서는 Direct Preference Optimization(DPO) 알고리즘을 도입하여 법률 AI의 성능 향상을 탐구하였다. 이러한 연구를 통해 제안된 DPO 알고리즘이 RLHF 알고리즘의 한계와 법률 AI 분야의 도전 과제에 대한 혁신적인 해결책을 제시함으로써, 새로운 시각에서의 법률 언어 모델 개발의 가능성을 제시한다. 본 연구는 사람의 개입 없이 명시적이고 효율적인 선호도 데이터셋을 생성하고, 이를 활용하여 DPO 알고리즘을 통한 모델 학습을 수행함으로써 기존 언어 모델의 전문성을 효과적으로 활용하고 법률 AI의 성능을 향상시킬 것으로 기대한다. 이 방식을 통해 새로운 법률 AI의 성능 향상에 대한 통찰을 얻을 것으로 기대되며, 실제 연구를 통해 법률 AI 모델의 성능을 측정하고, 다른 특정 도메인에서의 응용 가능성을 검토할 예정이다.

ACKNOWLEDGMENT

This work was supported by the MSIT, Korea, under the ITRC (Information Technology Research Center) support Program (IITP-2024-2021-0-01835) supervised by the IITP (Institute for Information & Communications Technology Planning Evaluation.)

참 고 문 헌

[1] Lai, Jinqi, et al. "Large Language Models in Law: A Survey." ArXiv, 2023,/abs/2312.03718. Accessed 9 Jan. 2024.

[2] Santacroce, Michael, et al. "Efficient rlhf: Reducing the memory usage of ppo." arXiv preprint arXiv:2309.00754 (2023).

[3] Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." arXiv preprint arXiv:2305.18290 (2023).

[4] Tian, Katherine, et al. "Fine-tuning Language Models for Factuality." NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following. 2023.