LLM 법률 데이터셋의 품질검증 방안 고찰

김민진, 이상복*
한국정보통신기술협회
minjin19@tta.or.kr, *jangpo@tta.or.kr

A Study on the Quality Evaluation Method of LLM Legal Datasets

Kim Min Jin, Lee Sang Bok*
Telecommunications Technology Association.

요 약

전 세계적으로 리걸테크(LegalTech)에 대한 관심이 증가하고 있으며, 대형 언어모델(LLM)의 발전으로 법률 산업에서 인공지능 기술의 도입이 가속화되고 있다. 그러나 한국의 리걸테크 발전은 고품질의 데이터셋 및 한국어 법률 LLM 모델의 부족으로 인해 여전히 초기 단계에 머물러 있다. 본 논문에서는 법률 데이터셋의 특성과 이를 반영한 품질검증 방안을 제시하고, 법률 LLM의 성능 향상을 위한 새로운 기법을 제안한다. 법률 데이터셋의 정확성, 최신성, 편향성, 일관성, 복잡성을 보장하기 위한 검증 방법을 제시하고 RAG(Retrieval-Augmented Generation)와 Instruction Tuning 데이터셋 활용으로 법률 LLM의 최신성과 정확성을 높이는 방안을 제안한다. 이러한 방안은 신뢰할 수 있는 법률 서비스 제공으로 향후 한국 리걸테크 시장 발전을 기대한다.

Ⅰ. 서 론

전 세계적으로 리걸테크(LegalTech)에 대한 관심이 커지고 있으며, 대형 언어모델(LLM)의 등장으로 법률 산업의 인공지능 기술 발전이 가속화되고 있다. 그러나 해외 리걸테크의 빠른 성장세에 비해 한국의 리걸테크는 저조한 성장을 하고 있다[1]. 한국 법률산업의 시장 규모가 1조 8,000억원에 달하는 것으로 추산되고 있음에도 불구하고 데이터셋 부족, 언어 모델 부족으로 인해 인공지능 기술 도입의 초기 단계에 있다[2]. 국내 리걸테크 시장 확대를 위해서는 고품질의 대규모 데이터셋 구축 및 한국어 법률 LLM 모델 개발이 필수적이다.

법률 데이터셋은 법적 문서, 판례, 법령 등의 복잡한 정보를 포함하고 있다. 이러한 데이터의 품질은 대형 언어 모델의 성능에 직접적인 영향을 미친다. 품질이 보장되지 않은 데이터셋으로 모델을 학습 할 경우 모델이 잘못된 법적 판단을 내릴 위험이 크기 때문이다. 법률 LLM이 효과적으로 활용되기 위해서는 법률 도메인이 가지는 몇 가지 특성을 보장해야 한다. 따라서 본 논문에서는 법률 데이터셋의 특성과 그에 맞는 품질검증 방안을 제시하고 품질 향상을 위한 새로운 기법을 제안하고자 한다.

Ⅱ. 본론

1. 법률 데이터셋의 특성

법률 데이터셋은 정확성, 최신성, 편향성, 일관성, 복잡성을 만족해야 한다[3]. 일반적인 LLM 모델의 데이터셋이 가지는 특성을 모두 만족하면서 높은 수준의 정확성과 최신성을 요구한다. 이에 대한 설명은 표1과 같다.

표 1. 법률 데이터셋의 특성

데이터 특성	설명
정확성	법률 데이터셋에 포함된 정보가 실제 사실에 기반하며 오류가 없어야 한다. 법률 데이터셋에서 정확성은 가장 중요한 요소이다. 정확한 사실이나 법률 정보를 반영하지 못한다면 법적 결정에 큰 영향을 미칠 수 있다.
최신성	법률은 시간이 지남에 따라 개정되거나 새로운 판례가 추가 된다. 따라서 데이터셋이 최신 정보를 반영할 수 있도록 지속적인 업데이트가 필요하다. 최신 정보에 기반한 법적 분석은 변화하는 법적 환경 대응에 필수적이다. 최신성을 확보하지 못한 데이터셋은 법적 효력이 없거나 최신 법률 기준에 부합하지 않는 잘못된 결과를 초래할 수 있다.
편향성	법률 데이터셋이 특정 지역, 분야, 사건 유형에 집중 될 경우 다른 법률 분야의 사건을 처리하는 데 어려움을 겪을 수 있다. 또한, 특정한 법적 입장이나 관점으로 불균형을 이루지 않도록 중립적인 관점으로 구축되어야 한다. 편향된 데이터는 법률 해석이나 판례 분석에 있어서 공정성을 해칠 수 있다.
일관성	일관성은 데이터셋 내에서 동일한 기준 및 형식이 유지되는 지를 의미한다. 기준은 개념이나 용어가 일관되게 사용 되는지에 대한 내용적 측면을 의미한다. 형식은 데이터 포맷과 같은 구조적인 측면을 의미한다. 법률 데이터셋에서 일관성이 유지되지 않을 경우 동일한 문제에 대해 다른 결과를 도출하여 신뢰성에 문제가 생길 수 있다.
복잡성	법률 용어와 문장은 다의적인 특성을 가지고 있다. 따라서 문맥에 따라 해석이 달라질 수 있다. 다의어를 제대로 처리하지 않아서 복잡성 관리가 되지 않는다면 잘못된 의미 분석으로 인해 법적 판단이 왜곡될 수 있다.

2. 법률 데이터셋의 품질검증 방안

2.1 구문 정확성

법률 데이터셋의 특성 중 일관성을 확인하기 위한 검증 항목이다. 데이터가 일정한 형태와 구조로 구축되어 있는지를 확인한다. 구조 정확성과 형식 정확성으로 나눌 수 있다. 구조 정확성은 데이터의 포맷에 따른 구조를 확인한다. 데이터 정의서의 구조에 따라 '필수 항목의 여부', '정의 되지 않은 항목의 포함 여부'를 확인한다. 형식 정확성은 정의된 구조 안의 '값의 누락'이나 '범위' 등을 확인한다.

2.2 의미 정확성

법률 데이터셋의 특성 중 정확성, 편향성, 일관성, 복잡성을 확인하기 위한 검증 항목이다. 가장 많은 데이터의 특성을 확인할 수 있으며, 고품질 데이터를 구축하기 위해 가장 중요하게 검증 해야하는 항목이다. 법률 도메인 특성상 전문가 검사가 필수적이다. 데이터의 오류가 신뢰성에 큰 영향을 미치기 때문이다. 모델의 임무에 따라 다르지만 '질의-응답 적정성', '요약 적정성', '판결 예측 정확성'을 확인한다. 구축된 데이터가 정확한지, 데이터의 의미가 문맥에 맞게 정확히 표현되었으며 올바른 단어로 작성되어 있는지 확인한다. 이 때, 데이터 자체의 정확성 뿐만 아니라 데이터 구축시 사용된 기준서도 함께 확인한다. 기준서가 일관적이고 편향적인 기준을 가지지 않는지, 정성적인 기준으로 인해 평가자 마다 다른 의견을 가지지 않는지 확인한다.

2.3 다양성

법률 데이터셋의 특성 중 최신성과 편향성을 확인하기 위한 검증 항목이다. 판결문의 경우 '선고 일자'를 법령의 경우 '시행 일자', '공포 일자'를 확인할 수 있다. 시기별 분포 확인을 통하여 최신화된 데이터가 구축 되었는지, 특정 시기의 데이터에 편향되어 있지 않은지를 확인 할 수 있다. 또한, '법률 문서의 출처', '법원명', '사건명' 등의 분포 확인을 통해 데이터가 특정 지역 혹은 사건에 편향되어 있지 않은지 확인한다. 해당 항목을 통해 다양한 법적 문제에 대응할 수 있는 범용성을 확보할 수 있다.

2.4 중복성

데이터셋에 같은 데이터가 반복적으로 들어가 있지 않은지 확인하기 위한 항목이다. 데이터의 중복이 많을 경우 데이터의 반복 생성을 유도하여 모델을 편향시킬 수 있다. 하지만 법률 데이터의 특성상 판결문의 주문과 같이 같은 문장이 반복되어야 하는 경우가 있다. 따라서 문장 단위의 중복성 검사 보다 문단 단위, 문서 단위 혹은 코사인 유사도를 통한 중복성을 확인한다.

2.5 유효성

법률 데이터셋의 특성 중 정확성, 일관성을 확인하기 위한 검증 항목이다. 구축한 데이터가 LLM에서 목표한 성능을 보이는지, 동일한 문제에 대해 동일한 결과를 도출하는지 확인하기 위한 검증 항목이다. 위의 모든 검증 항목을 만족한다고 해도 LLM이 올바른 결과를 출력하지 못하면 고품질의 데이터셋을 구축했다고 볼 수 없다. 데이터셋의 목적에 따라 달라질 수 있지만, '질의-응답 성능', '요약 성능', '판결 예측 성능' 항목을 통해 LLM이 학습한 데이터가 적절하고 신뢰할 수 있는지 검증한다.

3. 법률 데이터셋의 품질 향상을 위한 기법 제안

법률 LLM의 가장 큰 문제는 환각 문제(Hallucination)이다. 환각은 법률 분야의 인공지능 사용에 있어서 정확성을 떨어뜨리고 잘못된 법적 판단을 유도해 큰 문제를 일으킬 수 있다. 따라서 본 장에서는 법률 LLM의 환각 문제를 해결하고 정확성, 최신성을 보장하기 위한 기법을 제안한다.

3.1 Instruction Tuning 데이터

법률 분야에서 Instruction Tuning 데이터의 사용은 LLM이 단순히 법적 사실을 나열하는 것이 아니라 사용자의 지시에 맞는 적절한 법적 절차나 지침을 구체적으로 안내한다. 따라서 LLM의 목적에 맞는 Instruction Tunning 데이터를 추가로 구축한다면 법적 절차의 단계적 안내, 다양한 법적 시나리오에 대한 대응, 상황 맞춤형 답변과 같이 사용자의 상황에 맞는 구체적이고 정확한 답변이 가능하다.

3.2 RAG

RAG(Retrieval-Augmented Generation) 기법은 기존 데이터셋에서 유사한 사례나 정보를 검색하여 이를 기반으로 새로운 텍스트를 생성하는 방식이다. RAG 기법의 사용은 법률 데이터셋의 정확성 및 최신성을 보완할 수 있다. 관련 정보를 검색해서 사용할 수 있으므로 최신 법률 문서나 판례를 검색하여 업데이트된 정보를 제공할 수 있다. 하지만 RAG 기법의 사용에도 한계가 있다. 환각 문제를 줄여줄 수는 있으나 완전히 해결하지는 못한다는 것이다[5]. 또한, 검색된 정보가 항상 정확하거나 최신 정보일 수 없으므로 수집처의 신뢰성을 확인하거나, 레퍼런스 데이터셋을 사전에 구성하는 방법이 함께 고려되어야 한다[6].

Ⅲ. 결론

현재 한국의 법률 LLM은 판례 검색, 문서 초안 작성 등의 기초적인 기능만을 수행할 수 있는 단계에 머물러 있다. 그러나 고품질의 법률 데이터셋이 구축되고 데이터 품질 향상을 위한 다양한 기법이 적용된다면 보다 효과적이고 신뢰할 수 있는 법률 서비스를 제공할 수 있을 것이다. 향후, Instruction Tuning 데이터셋과 RAG 기법의 활용으로 인한 성능 향상 및 세부적인 품질 검증 결과를 제시하고자 한다. 본 논문에서 제안한 품질검증 방안 및 새로운 기법으로 고품질 법률 데이터셋을 구축하여 한국 리걸테크 시장이 성장하기를 기대한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 초거대AI 확산 생태계 조성 사업 (2100-2131-305, 2024년 초거대AI 확산 생태계 조성 사업)에 의해서 수행되었습니다.

참 고 문 헌

[1] 이지형, "해외 리걸테크 동향과 시사점." KISDI AI Outlook, vol. 16, 2024, pp. 1-22.

[2] Wonseok Hwang, et al. "A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction.", 2022.

[3] ISO/IEC DIS 5259-2(en), Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures, (https://www.iso.org).

[4] 과학기술정보통신부, 한국지능정보사회진흥원, 한국정보통신기술협회. "인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서 v3.1 - 제1권 품질관리 가이드라인 v3.1." 2024.

[5] Varun Magesh, et al. "Hallucination-Free? Assessing the Reliability ofLeading AI Legal Research Tools", 2024.

[6] 박정원 외, "사전학습단계 초거대 인공지능 학습용 데이터의 품질 요소 및 검증방안 고찰", 한국통신학회, 2023.09, pp. 123-124

[7] Junjie Huang, Qiang Li, et al. "Multi-modal Legal Judgment Prediction via Fact-specific Knowledge Injection", 2022.