자연언어처리 프로젝트 중간보고

2376250 장수연

1. 개발 진행 개요

1.1. 제안서 대비 변경된 사항

제안 단계에서 계획했던 기본 구조는 유지하되, 실제 데이터 분석 과정과 모델 실험을 거치며 일부 개발 방향이 조정되었다.

- 파인튜닝 모델 검증 방식 고도화: 데이터 편향을 줄이고 객관적인 성능 지표를 확보하기 위해 10-Fold 교차 검증 방식을 도입하였다.
- Retriever 최적화: 초기 계획에서는 KURE-v1 모델을 그대로 활용하여 법률 검색에 사용하는 방식이었으나, 법률 도메인의 특수성을 고려하여 Retriever 자체를 도메인 맞춤으로 파인튜닝하는 방향으로 변경하였다.

1.2. 현재까지의 전체 작업 흐름 요약

1. 데이터 구축: 약관 조항과 법령 데이터를 수집·정제하고, Retriever 학습 및 평가를 위한 데이터셋을 구축하였다.
2. LLM 파인튜닝: Meta-Llama-3.1-8B-Instruct 모델을 기반으로 파인튜닝을 진행하였으며, 데이터의 편차를 보완하기 위해 10개의 Fold에 대해 각각 별도의 모델 학습을 수행했다.
3. RAG 파이프라인 구축: KURE-v1 모델을 재학습하여 Retriever를 고도화하고, 파인튜닝된 LLM 모델이 생성한 '판단 근거'를 검색 쿼리에 포함하는 Query Expansion 방식을 적용하였다.
4. 성능 평가: 파인튜닝된 LLM, 임베딩 모델과 RAG 시스템에 대한 성능 평가를 수행하였다.

2. 데이터 구축

2.1. Llama 학습 데이터

- 소스: AI-Hub 약관 데이터셋 활용.
- 전처리: 원본 데이터의 Train/Validation 분할을 해제하고 전체를 병합하였다. 병합한 데이터를 무작위로 셔플링한 후 9:1 비율로 분할하는 과정을 10회 반복하여, 상호 배타적인 검증셋을 가진 총 10개의 Fold (Train/Val_fold_1~10)를 생성하였다.

2.2. 법령 데이터 구축

- 소스: 국가법령정보센터 '약관의 규제에 관한 법률' 제2장(불공정 약관 조항).
- 스키마: 다음의 구조로 통일하여 law_db를 구축하였다.

```
{
        "Id": "법령 조항 식별자",
        "Law_text": "법령 전문 텍스트"
}
```

2.3. Retriever 학습 및 평가를 위한 데이터셋 구축

- 정답 법령 매핑: AI-Hub 데이터셋의 각 약관 조항에는 해당 조항이 위반한 관련 법령의 정보가 텍스트로 명시되어있다. 이 법령 정보를 파싱하여 앞서 구축한 law_db의 고유 ID와 매칭하는 전처리 작업을 수행하였다. 하나의 약관 조항에 여러 개의 법령이 적용될 수 있으므로, 정답은 리스트 형태로 저장하였다.

  Ex) "약관법 제6조" → ID: 45

- 단순 약관 원문만으로는 법령 검색에 한계가 있어, '판단 근거'를 쿼리에 포함하는 전략을 수립하였다. 이를 위해, 데이터셋의 불공정 조항들을 앞서 학습시킨 SFT 모델에 입력하여, 해당 조항이 왜 불공정한지에 대한 법률적 판단 근거를 생성하게 하였다. 또한 이 결과를 데이터셋에 추가하였다.
- 구축된 데이터를 무작위 셔플링한 후, Train과 Validation을 위해 9:1 비율로 분할하였다.

3. LLM 모델 파인튜닝

3.1. 학습 환경

NVIDIA GeForce RTX 3070 Laptop GPU (8GB VRAM) 환경에서 수행되었다.

3.2. 최적화 전략

GPU 메모리 부족 문제를 해결하고 학습 효율을 높이기 위해 QLoRA방식으로 파인튜닝을 진행하였다. BitsAndBytesConfig를 통해 모델을 4bit로 로드하여 VRAM 사용량을 최소화하였다. 또한, 모델의 모든 파라미터를 업데이트하는 대신, Attention 모듈과 Feed-Forward Network의 선형 레이어에 LoRA 어댑터를 부착하고 해당 부분만 학습함으로써 연산 비용을 절감하였다.

3.3. 학습 방법

시스템 프롬프트를 활용하여 모델에게 법률 전문가라는 역할을 부여하였고, 출력 가이드라인을 제시하여 '분야/불공정여부/근거'의 정해진 형식으로 응답하도록 제약 조건을 설정하였다.

또한, 데이터가 모델에 입력되는 방식을 구조화하기 위하여 토크나이저의 apply_chat_template 기능을 적용하였다. 이를 통해 시스템 지시문과 사용자 입력이 모델 내부에서 구분되도록 처리하였다.

결과적으로 모델에 입력되는 학습 데이터는 다음과 같다

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

당신은 약관의 공정성을 분석하는 법률 전문가입니다.

문맥상 주체 (고객/ 사업자) 를 명확히 구분하세요.

반드시 아래 한 줄 포맷만 출력하세요:

분야: <정수> / 불공정여부: <유리|불리> / 근거: <간결한 문장 또는 '해당 없음'><|eot_id|><|start_header_id|>user<|end_header_id|>

다음 약관 조항의 문맥을 이해하여 분야 분류, 불공정 여부 판단, 판단 근거를 요약하시오.

입력:

제18조 (배상책임)

제1항 불의의 사고에 의해 을이 받은 손해 등 재난에 대하여 갑은 일체의 배상책임을 지지 아니한다. 다만, 갑의 고의로 을에게 불의의 사고가 생겼을 시에는 갑은 배상책임을 진다<|eot_id|><|start_header_id|>assistant<|end_header_id|>

분야: 6 / 불공정여부: 불리 / 근거: 해당 약관조항은 사업자인 시설 측의 면책권한을 인정하면서 사고 발생에 대하여 사업자의 고의가 있는 경우에만 배상책임을 지는 것으로 한정하고 있어 사업자의 과실 또는 부주의에 의해 초래될 수 되는 사건,사고에 대하여는 책임을 배제하여 사업자에게 지나치게 폭넓은 면책권한을 인정하게 된다.<|eot_id|>

3.4. 성능 평가

10-Fold 교차 검증 방식을 채택하여 모델의 성능을 객관적으로 검증하였다. 객관적인 성능 향상 입증을 위해, 파인튜닝 전후 모델을 동일한 조건에서 비교하였으며, 전체 데이터셋을 10개의 Fold로 분할하여 총 10회 반복 평가를 수행하였다. 평가의 공정성을 위해 각 Fold는 해당 회차의 모델 학습에 포함되지 않은 검증 데이터만을 사용하여 수행하였으며, Base 모델과 SFT 모델 모두 동일한 검증 데이터셋을 사용하여 성능을 측정하였다. 모델의 성능은 다음 두 과제에 대해 정량적 지표로 검증하였다.

(1) 유불리 판단

- 약관 조항의 유/불리 여부를 이진 분류한다
- unfair = 1, fair =0으로 라벨링하였다
- 평가 지표: Accuracy, Precision, Recall, F1

(2) 분야 분류

- 입력된 조항이 어떤 계약 분야에 속하는지 분류한다
- 총 43개 카테고리를 대상으로 한다
- 평가 지표: Accuracy, Macro F1

구분		유불리 판단				분야 분류	
Fold	**Model**	**Recall**	**Precision**	**Accuracy**	**F1**	**Accuracy**	**Macro F1**
Fold1	Base	0.3123	0.5427	0.6989	0.3964	0.2811	0.2890
	SFT	0.9649	0.9752	0.9811	0.9700	0.6500	0.6674
Fold2	Base	0.2882	0.5155	0.6856	0.3697	0.2967	0.2982
	SFT	0.9306	0.9889	0.9744	0.9589	0.4544	0.4495
Fold3	Base	0.2932	0.4756	0.6956	0.3628	0.3022	0.2903
	SFT	0.9098	0.9878	0.9700	0.9472	0.6956	0.6825
Fold4	Base	0.3364	0.5000	0.7556	0.4022	0.3044	0.3055
	SFT	0.9182	0.9902	0.9778	0.9528	0.6233	0.5939
Fold5	Base	0.3431	0.5165	0.7022	0.4123	0.2833	0.2884
	SFT	0.9380	0.9772	0.9744	0.9572	0.6978	0.6955
Fold6	Base	0.3058	0.4885	0.6867	0.3761	0.2956	0.2972
	SFT	0.9640	0.9640	0.9778	0.9640	0.6922	0.6589
Fold7	Base	0.2571	0.4675	0.6778	0.3318	0.3000	0.3061
	SFT	0.9500	0.9500	0.9689	0.9500	0.6467	0.6463
Fold8	Base	0.3014	0.5484	0.7033	0.3890	0.3022	0.2934
	SFT	0.9504	0.9817	0.9789	0.9658	0.5356	0.5305
Fold9	Base	0.2594	0.4276	0.7111	0.3229	0.2922	0.2792
	SFT	0.9331	0.9738	0.9756	0.9530	0.6522	0.6242
Fold10	Base	0.3403	0.5568	0.7922	0.4224	0.3100	0.2812
	SFT	0.9444	0.9891	0.9789	0.9663	0.4322	0.4507

(1) 유불리 판단

Base 모델의 경우 Recall이 0.3수준으로 '불리' 약관을 '유리'로 잘못 판단하는 비율이 높았다. 그러나 SFT 모델은 약 95% 수준의 Recall을 보여 '불리' 약관을 거의 빠짐없이 잡아낼 수 있게 되었다. 동시에 Precision 또한 대폭 상승했는데, 모델이 불리하다고 지적했을 때 그것이 오답일 확률이 매우 낮아졌음을 의미한다. 결과적으로 F1-Score가 상승하며, SFT적용 이후 모델의 유불리 판단 성능이 유의미하게 상승하였음을 확인하였다.

(2) 분야 분류

SFT 적용 후 성능이 향상되었으나, 정확도가 약 60%대 구간에서 머물렀다. 이는 모델의 성능 문제라기보다, 입력 정보의 한계에 기인한것으로 분석하였다. 약관 조항이 여러 데이터에 공통으로 등장하는 범용적인 텍스트로 구성된 경우가 많고, 분류해야 할 타겟 클래스가 43개로 매우 세분화되어 있기 때문에 분류 작업이 어렵다. 현재 모델은 약관 원문만을 입력으로 받고 있는데, 분야 분류의 경우, 약관 원문 조항뿐만 아니라 분쟁 경위 등을 함께 입력받으면 성능을 극대화 할 수 있을 것으로 사료된다.

4. 관련 법률 RAG 검색 성능 고도화

본 프로젝트의 초기 단계에서는 법률 도메인에 특화된 것으로 알려진 nlpai-lab/KURE-v1 모델을 그대로 사용하였다. 약관 원문만으로 법령을 검색하는 실험 결과, Recall@1 지표가 0.0889에 그치는 저조한 성능을 보였다. 이를 해결하기 위해 2단계 성능 고도화 전략을 수립하고 실행하였다.

(1) 임베딩 모델 파인튜닝: 근본적인 모델 성능 향상을 위해 약관과 정답 법령을 매칭시킨 데이터셋으로 Retriever 모델 자체를 재학습시켰다.

(2) LLM이 생성한 '법률적 판단 근거'를 활용한 쿼리 확장: 단순 약관 텍스트만으로는 정보가 부족하다고 판단하여, SFT 모델이 생성한 법률적 판단 근거를 검색어에 포함시켰다. LLM이 생성한 텍스트에는 "약관법 제6조", "신의성실의 원칙" 등 법령과 유사한 전문 용어가 포함되어 있어, 검색 모델이 정답 법령과 매칭될 확률을 높여줄 것으로 가설을 세웠다.

구분	Recall@1	Recall@5	Recall@10	MRR	비고
Baseline	0.0889	0.2926	0.4556	0.1847	쿼리: 약관 원문 모델: Base KURE-v1
Step 1	0.3778	0.6926	0.8370	0.5235	쿼리: 약관 원문 모델: Tuned KURE-v1
Step 2	0.3926	0.7148	0.8519	0.5332	쿼리: 약관 원문 + 근거 모델: Tuned KURE-v1

우선 쿼리는 '약관 원문'으로 고정한 상태에서, 모델만 Base에서 Tuned 모델로 교체하여 성능을 비교하였다. Baseline 대비 Step 1의 Recall@1은 0.0889에서 0.3778로 약 325% 대폭 향상되었다.

다음으로 파인튜닝된 모델을 고정한 상태에서, 쿼리에 SFT 모델이 생성한 '판단 근거'를 추가했을 때의 변화를 측정하였다. 약관 원문만 사용한 Step 1 대비, 판단 근거를 포함한 Step 2의 Recall@1은 0.3778에서 0.3926으로 약 3.9% 추가 상승하였다. 또한 MRR 지표가 0.5235에서 0.5332로 개선되었는데, 이는 쿼리 확장이 정답 법령을 검색 결과 상위권으로 끌어올리는 데 실질적인 기여를 했음을 시사한다.

5. 추후 고도화 계획

5.1. 외부 데이터 기반 실용성 검증

- 데이터 출처: 한국공정거래조정원 분쟁조정사례
- 실제 분쟁 사례 데이터를 수집·정제하여, 현재 AI-hub 데이터 기반으로 구축된 모델이 현실 세계의 불공정 약관을 얼마나 잘 탐지하고 법적 근거를 제시하는지 일반화 성능을 최종 검증할 예정이다.