한국어 법률 텍스트 처리를 위한 언어 모델링 연구

강예지○, 비립, 장연지◇, 강혜린, 박서윤, 김한샘†

연세대학교 언어정보연구원†, 국립국어원◇

{yjkang5009○, feili0820, hyerink, seoyoon.park, khss†}@yonsei.ac.kr

yeonji3547◇@korea.kr

A Study on Language Modeling for Korean Legal Text Processing

Ye-Jee Kang○, Fei Li, Yeon-Ji Jang◇, Hye-Rin Kang, Seo-Yoon Park, Han-Saem Kim†

Institute of Language and Information Studies, Yonsei University†

National Institute of Korean Language◇

요 약

본 논문은 한국어 법률 텍스트 처리를 위해 세 가지 서로 다른 사전 학습 모델을 미세 조정하여 그 성능을 평가하였다. 성능을 평가하기 위해 타겟 판결 요지에 대한 판결 요지 후보를 추출하여 판결 요지 간의 유사도를 계산하였다. 또한 유사도를 바탕으로 추출된 판결 요지가 실제 법률 전문가와 일반 언어학자의 직관에 부합하는지 판단하기 위해 정성적 평가를 진행하였다. 그 결과 법률 전문가가 법률 전문 지식이 없는 일반 언어학자에 비해 판결 요지 간 유사도를 낮게 평가하였는데 법률 전문가가 법률 텍스트의 유사성을 판단하는 기준이 기계와 일반 언어학자와는 달라 전문가 자문에 기반한 한국어 법률 AI 모델 개발의 필요성을 확인하였다. 최종 연구 결과로 한국어 법률 AI 프레임워크를 제안하였다.

주제어: Legal AI, 법률 텍스트 처리, 법률 AI 프레임워크

1. 서론

법률 문서의 접근이 용이해짐에 따라 일반인은 물론 법률 전문가가 온라인상에서 법률 문서를 검색하고 이를 활용할 수 있게 되었다. 하지만 한국의 현 대법원 판례 검색 시스템은 키워드 기반으로 검색 결과를 제공하기 때문에 법률 전문 지식이 없는 일반인의 경우 검색 의도에 부합하는 결과를 찾는 데 한계가 있다. 또한 법률 문서는 많은 법률 전문 용어와 법률적 사실 관계가 포함되어 있어 일반인들이 이를 이해하기 어려울 뿐만 아니라 법률 전문가 역시 대법원 시스템에서 법률 문서를 검색하는 데 많은 시간과 비용이 든다는 점에 문제가 있다.

해외에서는 이러한 문제를 해결하기 위하여 유사 케이스 매칭 데이터셋 구축에 힘을 기울이고 있다. 반면, 국내에서는 법률 도메인 내 유사 판례 데이터셋이 존재하지 않을 뿐만 아니라 법률 텍스트를 처리하기 위한 대규모의 법률 텍스트로 학습된 도메인 특화 언어 모델 개발에 대한 연구가 미진하다.

따라서 본 논문에서는 한국어 법률 텍스트 처리를 위해 기존의 사전 학습(pre-trained) 모델을 미세 조정(fine-tuning)하여 3가지의 서로 다른 모델을 구축하였다. 각 모델의 성능을 확인하기 위한 다운스트림 태스크로 판례에 나타나는 요지와 비슷한 요지를 찾아내도록 하였다. 또한 모델에 의해 추출된 판결 요지가 실제로 타겟 요지와 비슷한지 검증하는 단계를 거쳐 모델의 성능이 합당한지 확인하였으며, 본 연구를 통해 한국어 텍스트 처리를 위한 언어 모델링의 중요성을 밝히고 모델 설계를 위해 고려해야 할 요인 역시 살펴보았다.

2. 관련 연구

2.1 Legal AI

'법률 AI'는 인공지능 기술을 적용하여 법률 문서 검색, 판결 예측, 판결문 요약, 유사한 판결문 도출, 법률 Q&A 등 여러 법률 태스크를 수행하기 위한 연구이다. 법률 AI는 기존에 존재하는 데이터셋을 이용하여 세부 법률 태스크를 수행하고자 하는 연구와 세부 태스크 수행을 위해 태스크 목적에 맞는 데이터셋을 구축하는 연구로 나눌 수 있다. 먼저 세부 법률 태스크를 위한 연구로는 자연어 처리(NLP)가 여러 법률 태스크를 수행하는 데 사용됨으로써 법률 시스템을 어떻게 개선할 수 있을 것인가에 대한 연구가 있다[1]. [2]에서는 법률 문서 자동 요약과 관련한 태스크를 수행하기 위해 법률 문서의 요약문의 구조에서부터 기타 다른 업스트림 자연어 처리 작업을 수행하여 요약문의 품질을 향상시킬 수 있는지 등과 같은 법률 요약문 자동 생성과 관련한 전반적인 연구를 진행하였다. COLIEE 는 법률 텍스트를 사용하여 정보 검색 및 함의에 대한 최신 기술을 평가하기 위한 경진대회로, 성문법과 판례법 각각에서 정보 검색, 기존의

케이스와 새로운 케이스 사이의 함의 관계 도출, QA를 수행하도록 하였다[3].

법률 태스크 수행을 위한 다양한 데이터셋 구축 연구는활발히 진행되고 있다[4, 5, 6, 7]. [4]는 판결 예측의 성능을 높이기 위해 이혼, 노동, 대출을 주제로 서로 다른 케이스에서 법률적 요소를 추출하여 데이터셋을 구축한 후 판결 결과를 예측하도록 하는 멀티 라벨 분류 문제를 수행하였다.

[5]는 중국어 판결 예측의 벤치마크 데이터셋 C-LJP를 구축하였으며, [6]은 유사 케이스 매칭 데이터셋인 CAIL2019-SCM을 구축하여 법률 전문가로 하여금 판결문을 찾는 데 들어가는 시간과 비용을 줄이게 하고 법률 IR(Information Retrieve)에 기여하였다는 점에서 의의가 있다. 한편, [7]은 법률 Q&A의 성능을 향상시키기 위해 중국어 법률 QA 벤치마크를 구축하였으며, [8]은 영어 법률 NLU(Natural Language Understanding) 태스크 처리를 위한 벤치마크 데이터셋인 LexGLUE를 구축하였다. 이상의 연구들과 같이 국외에서는 법률 AI를 위한 다양한 태스크 및 데이터셋 구축 관련 연구가 활발하게 진행되고 있으나 국내에서는 이에 대한 연구가 미비한 실정이다. 아주 최근에 들어서 국내 최초로 한국어 법률 AI 벤치마크 데이터셋 Lbox Open 및 한국어 법률 언어 모델 LCUBE가 공개되었다[9]. 이는 국내 최초로 한국어 법률 평가 데이터셋을 구축하였다는 점에서 의의가 있으며, 향후 Lbox Open에서 다루는 태스크 외 다양한 세부 법률 태스크 관련 연구가 활발하게 진행될 수 있다는 가능성을 제시해 주었다.

2.2 법률 문서 처리를 위한 언어 모델

법률 문서를 처리하기 위해 다양한 신경망 모델들이 사용된다. 본 논문의 목적인 유사한 판례 요지를 검출하기 위해 사용할 수 있는 모델은 Doc2Vec과 트랜스포머 기반의 모델 등이 있다. Doc2Vec은 문서 간의 유사도를 계산할 수 있도록 하여 문서 분류의 성능을 높이며, 단어의 층위에서 더 나아가 문장, 단락 그리고 문서 등을 벡터로 표현할 수 있게 한다[10]. [11]은 판결 문서가 유사한지의 여부를 확인하기 위해 Doc2Vec을 이용하여 문서 간 유사도를 계산하였으며, 전문가가 판단한 결과와 이를 비교하였다. 한편, 트랜스포머 기반의 모델이 등장하게 되면서 법률 문서를 처리하기 위해 트렌스포머 기반 모델을 이용한 연구들을 확인할 수 있다. [12]에서는 법률 텍스트를 처리하기 위해 트랜스포머 기반 모델에 미세 조정(fine-tuning)을 거친 후 5가지 법률 세부 태스크를 수행하였다.

BERT는 Google AI팀에서 개발한 트랜스포머 기반 언어 모델로 문장 단위 임베딩 및 문맥 기반 임베딩 모델이다[13]. BERT를 한국어에 적용하기 위해서는 BERT-multilingual 모델을 사용할 수 있으며, 기존의 BERT 모델에서 한국어 성능 한계를 극복하기 위해 개발된 SKTBrain의 KoBERT[1] 등을 사용할 수 있다. DistilBERT는 복잡한 모델로부터 증류한 지식을 작은 모델로 전달한다는 의미를 가지는 지식 증류(knowledge distillation)를 기존의 BERT에 적용해 모델의 크기가 작음에도 불구하고 빠른 속도로 기존의 모델과 비슷한 성능을 가진다는 점이 강점이다[14].

본 연구에서는 트랜스포머 기반의 BERT 모델을 미세 조정하여 한국어 법률 도메인에 특화된 모델을 개발하고 그 성능을 평가하고자 한다.

3. 유사 판결 요지 도출 방법

3.1. 데이터 수집 및 전처리

본 연구에서는 국가법령정보센터의 open api2[2]를 통해 판례 총 83,340건을 수집하였다. 그중 판결 요지가 있는 판례만 추려 63,126건을 별도로 저장하였다.

BERT 모델은 입력 시퀀스 문장의 최대 길이를 512로 제한하고 있다. 법률 문서의 경우 문장의 길이가 길다는 도메인의 특수성이 있는데, 판결 요지를 BERT 모델에 임베딩하기 위해 각각의 판례 요지를 구분자인 마침표 단위로 분절하였다. 또한 분절된 문장에서 불필요한 특수문자는 제거하였다.

3.2. 모델링 방법

유사 판결 요지를 추출하기 위해 세 가지 모델을 사용하였으며, 세 가지 모델은 모두 BERT 언어 모델을 미세 조정(fine-tuning)한 것이다. 각 모델은 huggingface에 공개되어 있는 사전 학습 모델로 BERT-multilingual 모델[3]과 Distilbert의 multilingual 모델[4], 그리고 KoBERT[5]를 사용하였다.

각 모델의 파라미터는 아래와 같으며 학습 데이터셋과 평가용 데이터셋은 9:1의 비율로 분할하였다.

표 1. 각 모델의 파라미터

	Bert-multi	Distilbert-multi	KoBERT
epoch	5	5	10
Learning rate	1e-4	1e-4	1e-4
Batch size	32	32	32
Max length	128	128	128

3.3. 유사 판결 요지 추출

세 가지 서로 다른 BERT 기반의 fine-tuning 모델의 성능을 평가하기 위해 각 모델에 특정 판결 요지를 넣었

[1] https://github.com/SKTBrain/KoBERT
[2] https://open.law.go.kr/LSO/main.do
[3] https://huggingface.co/bert-base-multilingual-cased
[4] https://huggingface.co/distilbert-base-multilingual-cased
[5] https://huggingface.co/monologg/kobert

을 때 가장 유사한 판결 요지가 추출되도록 하는 태스크를 수행하였다. 각 모델에서 테스트셋 6,313건을 대상으로 각 판결 요지가 서로에게 얼마나 유사한지를 코사인 유사도를 통해 계산하도록 하였다.

각 판결 요지 간의 유사도를 바탕으로 특정 타겟 요지와 가장 유사한 판결 요지 9개를 유사 판결 요지 후보(candidate)로 선정하였다. 아래 표는 타겟 요지와 가장 유사한 판결 요지 후보 9개 중 일부이다.

표 2 타겟 요지에 해당하는 후보 판결 요지

target	candidate_1	candidate_2
사실혼관계에 있어서는 당사자 사이에 부부로서의 실체가 없어지면 달리 어떠한 형식적 절차를 받을 필요도 없이 그 관계는 해소되는 것이므로 당사자가 관례에 따른 혼례식을 치루고 잠시 함께 동거한 사실이 있다 하더라도 부가 시가에서 쫓겨나 4년이 훨씬 넘는 세월 동안 부부로서의 일체의 교섭을 끊고 서로 따로 떨어져 지내왔고 더욱이 부가 부와 함께 지내기를 굳이 마다하고 있다면 그 책임이 누구에게 있든간에 그들 사이에 부부로서의 실체가 없어져 그 사실혼의 관계는 이미 해소되었다고 볼 것이다.	중매결혼의 경우 당사자 일방이 과거에 이미 혼인하여 자식까지 출산한 전력이 있었다는 사실은 상대방 당사자가 그와의 결혼여부를 결정함에 있어 중요한 조건이 되므로, 당사자 일방이 자신의 혼인 전력을 감추고 노총각이라고 속여 상대방과 결혼하게 되었고, 뒤늦게 이러한 사실을 알게 된 상대방이 그에 충격을 받아 친정으로 돌아가 별거하고 있다면 그들 사이의 사실혼관계는 위 기망행위 등으로 인하여 파탄되었다고 할 것이다.	개발제한구역내의 토지를 소유한 자가 부정한 방법을 써서 증축허가신청을 받아 건물을 신축한후 관계공무원을 기망하여 준공검사까지 받았다면 도시의 무질서한 확산을 방지하고 도시주변의 건전한 생활환경을 확보하기 위한 개발제한구역의 지정취지에 비추어 개발제한지역내의 위 불법건축물의 철거불이행을 방치함은 심히 공익을 해한다 할 것이며, 또 건물이 완성되고 등기까지 마치었다 하여 건축허가나 준공처리의 취소를 못할 바도 아니다.

4. 평가

3.3장의 모델에 의해 나온 유사도를 바탕으로 도출된유사 판결 요지 후보를 5개의 도메인(가사, 일반 행정, 형사, 민사, 세무)에서 일부 샘플링하였다. 그리고 샘플링한 타겟 판결 요지와 9개의 후보 판결 요지가 실제 유사한지 정성적 평가를 진행하였다. [15]에서는 법률 전문가 한 명, 언어학자 한 명, 일반인 한 명, 총 세 명의 판단자가 [0, 10]의 척도로 케이스가 유사한지 그렇지 않은지를 판단하였다. 이때 두 케이스가 전혀 유사하지 않을 경우 0을, 매우 유사할 경우에는 10을 부여하였다. 본 연구는 [15]의 방법을 따라 법률 전문가 2명, 법률 전문적 지식이 없는 일반 언어학자 2명이 판단자가 되어 판결 요지 간 유사도를 정성 평가하였으며, 평가 척도는 0 또는 1로만 부여하도록 하였다. 그리고 판단 결과를 법률 전문가와 일반 언어학자로 나누어 각 그룹의 최종 점수를 부여하고, 모델에 의해 도출된 유사도와 비교하여 기계가 도출한 유사도가 인간의 직관과 얼마나 부합하는지 살펴봄으로써 모델 성능의 신뢰성을 평가하였다. 표 3은 타겟 판결 요지와 이에 대한 판결 요지 후보의 유사도 평가를 진행한 결과 중 일부이다.

표 3 타겟 판결 요지-판결 요지 후보에 대한 유사도 평가 예시

판결 요지 (target)	candidate	평가자	평가
이혼당사자 사이나 그의 배우자의 친족 특히 직계존속과의 사이에 행동이 수반하지 않는 단순한 감정의 갈등, 균열 방지 내지대립이 생겼다는 것을 본조 제6호 소정의 혼인을 계속하기 어려운 중대한 사유에 해당된다고는 할 수 없다.	친권자가 3살과 1살밖에 안되는 자식들을 남겨 놓고 무단가출한 이래 10여년간 아무런 소식 없었다면 이는 위 자식들에 대한 친권을 행사시킬 수 없는 중대한 사유가 있는 때에 해당한다.	일반인	1
		전문가	0
	민법 제840조 제6호에 해당하는 혼인을 계속하기 어려운 중대한 사유가 이혼청구 당시까지 계속되고 있는 경우에는 민법 제842조가 적용될 여지가 없다.	일반인	1
		전문가	1
불법점유를 당한 소유자로서는 불법점유자에 대하여 그로 인한 임료상당의 배상이나 부당이득의 반환을 구할 수 있을 것이나 불법점유라는 사실이 발생한 바 없었다고 하더라도 부동산소유자에게 임료상당 이익이나 기타 소득이 발생할 여지가 없는 특별한 사정이 있는 때에는 손해배상이나 부당이득반환청구를 할 수 없다.	피고 시가 원고 소유인 대지를 법률상 원인없이 도로로 조성하여 점용 사용하고 있음을 이유로 원고가 피고에 대하여 위 대지에 대한 임료상당액의 반환을 구하는 경우에는 원고는 피고가 받고 있는 이득인 도로로서의 임료상당액 이외에 민법 제748조 제2항에 의하여 원고가 입고 있는 손해배상까지를 구하고 있는 것이라고 볼 수도 있으므로 피고시는 대지로서의 임료상당액을 원고에게 반환하여야 한다.	일반인	0
		전문가	1

5. 결과 및 분석

2명의 법률 전문가와 2명의 언어학자가 타겟 요지에 대한 유사 판결 요지 후보 9개를 정성 평가한 결과는 다음 표 4, 5와 같다. 각 모델에서 도출한 판결 요지 후보에 대해서 모델별, 도메인별로 결과가 다르게 나타났으며, 법률 전문가와 언어학자 그룹 사이에서도 상이한 결과를 나타냈다. 먼저 KoBERT 모델에서 가장 낮은 성능을 보임을 알 수 있다. 또한 법률 전문가 그룹과 법률 전문 지식이 없는 일반 언어학자 그룹 간 결과를 비교해 보면 법률 전문가가 일반 언어학자에 비해 더 낮은 점수를 준 것을 확인할 수 있다. 이러한 결과는 표 5에서도 나타난다. 표 5는 각 도메인별로 법률 전문가 집단과 언어학자 집단의 정확도를 비교한 것으로, 근소한 차이가 나는 '형사' 도메인을 제외하고는 나머지 4개의 도메인에서 모두 법률 전문가가 모델의 성능을 낮게 평가하였음을 알 수 있다. 즉, 법률 전문가는 타겟 요지에 대한 9개의 판결 요지 후보에 대해서 이들 간 유사도가 낮다고 판단하여 언어학자에 비해 0을 더 많이 부여한 것이다. 이는 다시 말해 언어학자는 법률 전문가에 비해 법적 지식이 부족하기 때문에 타겟 요지에 나타나는 키워드를 기반으로 판결 요지 후보와의 유사도를 평가하였으며, 법률 문서의 유사도를 판단함에 있어서 기계의 직관과 더 유사하다는 것을 확인할 수 있었다. 더 나아가 법률 전문가는 기계 또는 일반 언어학자가 법률 텍스트의 유사도를 판단하는 것과는 기준이 다름을 알 수 있다.

표 4 각 도메인별 모델 결과에 따른 법률 전문가 및 언어학자 평가 결과 (E = 법률 전문가, L = 언어학자)

Model	E/L	도메인				
		가사	일반 행정	형사	민사	세무
BERT-multi	E	13.33%	16.67%	21.67%	**28.3%**	13.3%
	L	21.67%	15.00%	16.67&	51.6%	25.0%
Distil Bert-multi	E	5.00%	8.33%	21.67%	15.00%	3.33%
	L	21.67%	23.33%	20.00%	**40.00%**	18.33%
KoBERT	E	1.67%	3.33%	5.00%	11.67%	1.67%
	L	1.67%	16.67%	5.00%	3.33%	3.33%

표 5 도메인별 법률 전문가와 언어학자 Accuracy 결과

	Accuracy	
도메인	Expert	Linguist
가사	6.67%	15.00%
일반 행정	9.44%	17.78%
형사	15.00%	14.62%
민사	17.78%	31.67%
세무	5.00%	16.11%

법률 텍스트는 문장의 길이가 길며, 일반인들이 이해하기 어려운 법률 용어로 구성되어 있으며 법리를 바탕으로 한 논조가 존재한다. 따라서 두 판결 요지가 비슷하다고 판단할 수 있는 근거는 법리에 있으며, 논조가 되는 부분이 유사해야 한다. 또한 판결 요지에서 텍스트 상 겉으로 드러나는 내용으로 유사성을 판단하기보다 조항과 관련하여 판결이 내려진 관념적인 내용 역시 고려해야 하는 사항이다.

6. 법률 텍스트 모델링에 대한 향후 연구

본 연구에서 사용한 언어 모델은 법률 전문가의 판단에 부합할 만큼 아직 충분한 성능을 보여주지 못하고 있음을 확인하였다. 한국어 특화 법률 AI 모델을 개발하기 위해서는 정규화된 큰 모델이 필요하며, 이를 위해서는 언어학 분야와 법률 분야의 학문적 융합이 필수적이다. 따라서 향후 법률 텍스트 처리를 위한 프레임워크를 그림 1과 같이 제안한다. R1은 언어학적 측면에서 의미 또는 통사적 구성 요소이며, R2는 법률 텍스트에 나타나는 사건(event), R3은 법률 조항, 법리 등이 될 수 있다. 이와 같이 세 개의 R 요소들을 포함하는 태스크는 함께 순환적으로 진행되어야 할 것이다.

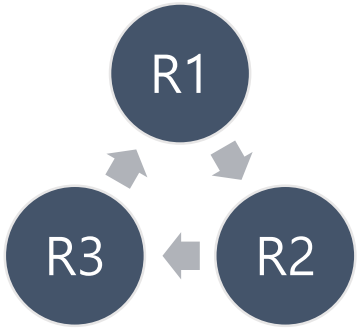

그림 1 한국어 법률 AI 프레임워크 제안

7. 결론

본 연구는 법률 데이터를 활용한 자연어 처리에 근본적으로 활용할 수 있는 한국어 법률 텍스트 특화 모델을 개발하기 위한 기초 실험으로 사전 훈련된 모델을 미세 조정하여 유사 판결 요지를 추출하고 이에 대한 성능을 평가하였다. BERT-multilingual, Distilbert-multilingual, KoBERT를 사용하여 타겟 판결 요지에 대한 판결 요지 후보를 추출하였으며, 법률 전문가 집단과 일반 언어학자 집단으로 나누어 판결 요지 간의 유사도를

측정하고 이에 대한 정성적 평가를 진행하였다. 그 결과 언어 모델의 성능이 법률 전문가의 판단과 부합하지 않으며, 법률 전문가는 기계 또는 법률 전문 지식이 없는 일반인이 법률 텍스트의 유사도를 판단하는 것과 기준이 다름을 확인하였다. 이에 본 연구에서는 법률 태스크 처리를 위해 한국어 법률 벤치마크 데이터셋 구축뿐만 아니라 언어학 분야와 법률 분야가 융합되어 구성된 한국어 법률 AI 프레임워크를 제안한다.

한편, 본 연구는 언어 모델의 결과에 대해 기계와 인간, 일반 언어학자와 법률 전문가가 법률 텍스트를 이해하는 데 차이가 있다는 것을 정성 평가를 통해 확인하였으며 모델 자체에 대한 평가를 시행하지 않았다는 점에서 한계가 있다. 이에 향후 연구로는 모델의 fine-tuning 이후 같은 문장에 대한 masked token 변화를 확인하여 정성 평가에 그쳤던 본 연구에서 나아가 모델 자체에 대한 성능을 비교 분석할 것이다. 또한 본 연구에서는 단순히 유사도에 기반하여 바이너리(binary)로 정성 평가를 진행하였다면, 향후 연구로는 제안한 프레임워크의 요소를 포함하는 정성적 평가 메트릭을 개발하고 이에 따른 정성평가 방법을 고안할 예정이다.

참고문헌

[1] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M., How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. arXiv:2004. 12158v5. 2020.

[2] Deepali Jain, Malaya Dutta Borah, Anupam Biswas, Summarization of legal documents: Where are we now and the way forward. Computer Science Review, Volume 40, 2021.

[3] Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In Proceedings of JSAI, 177-192. Springer. 2018.

[4] Yi Shu, Yao Zhao, Xianghui Zeng, and Qingli Ma. Cail2019-fe. Technical report, Gridsum. 2019.

[5] Xiao, Chaojun et al. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. ArXiv abs/1807.02478. 2018.

[6] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Heng Wang, Jianfeng Xu, et al. Cail2019-scm: A dataset of similar case matching in legal domain. arXiv preprint arXiv:1911.08962. 2019.

[7] Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. “Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension”. In Proceedings of CCL. Springer. 2019.

[8] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland. 2022.

[9] Hwang, W., Lee, D., Cho, K., Lee, H., & Seo, M. A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction. arXiv preprint arXiv:2206.05224.2022.

[10] Quoc Le ,Tomas Mikolov. Distributed Representations of Sentences and Documents. ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning Volume 32, 1188-1196(9) 2014.

[11] Mandal, Arpan & Chaki, Raktim & Saha, Sarbajit & Ghosh, Kripabandhu & Pal, Arindam & Ghosh, Saptarshi. Measuring Similarity among Legal Court Case Documents. 1-9. 10.1145/3140107.3140119. 2017.

[12] Nguyen, HT., Nguyen, MP., Vuong, THY. et al. Transformer-Based Approaches for Legal Text Processing. Rev Socionetwork Strat 16, 135-155 2022.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT (1), Association for Computational Linguistics, pp.4171-4186, 2019.

[14] Sanh, Victor & Debut, Lysandre & Chaumond, Julien & Wolf, Thomas. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019.

[15] Bhattacharya, Paheli & Ghosh, Kripabandhu & Pal, Arindam & Ghosh, Saptarshi. Methods for Computing Legal Document Similarity: A Comparative Study. arXiv:2004.12307. 2020.