KIPS Trans. Softw. and Data Eng.
Vol.11, No.2 pp.93~100
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.2.93

Deep Learning Based Semantic Similarity for Korean Legal Field

Sung Won Kim† · Gwang Ryeol Park††

ABSTRACT

Keyword-oriented search methods are mainly used as data search methods, but this is not suitable as a search method in the legal field where professional terms are widely used. In response, this paper proposes an effective data search method in the legal field. We describe embedding methods optimized for determining similarities between sentences in the field of natural language processing of legal domains. After embedding legal sentences based on keywords using TF-IDF or semantic embedding using Universal Sentence Encoder, we propose an optimal way to search for data by combining BERT models to check similarities between sentences in the legal field.

Keywords : NLP, LegalTech, Semantic Similarity, BERT, Legal

딥러닝을 이용한 법률 분야 한국어 의미 유사판단에 관한 연구

김 성 원† · 박 광 렬††

요 약

기존의 데이터 검색 방법으로는 키워드 중심의 검색 방법이 주로 사용되나, 이는 전문적인 용어가 많이 쓰이는 법률 분야의 검색 방법으로는 적합하지 않다. 이에 대해 본 논문에서는 법률 분야의 효과적인 데이터 검색 방안을 제안한다. 법률 도메인의 자연어처리 분야에서 문장 간의 유사성을 판단하는 데 최적화된 임베딩 방법에 관하여 서술한다. 법률문장을 TF-IDF를 이용하여 키워드 기반으로 임베딩하거나 Universal Sentence Encoder를 이용하여 의미 기반으로 임베딩을 한 후, BERT모델을 결합하여 법률 분야에서 문장 간 유사성을 검사하여 데이터를 검색하는 최적의 방안을 제안한다.

키워드 : 자연어처리, 리걸테크, Semantic Similarity, BERT, 법률

1. 서 론

최근 자연어처리 분야의 인공지능이 발전을 이루면서 도메인에 경계 없이 다양한 서비스에 접목이 되고 있다. 특히 법률, 특허 등 전문적인 용어가 자주 등장하는 리걸테크(Legal-Tech) 분야에서도 기술개발이 활발히 이루어지고 있다.

본 논문에서는 그중에서도 법률 분야의 문장들을 검색에 특화되도록 효과적으로 임베딩을 하고, BERT 기반으로 태스크를 전이학습 하는 방안에 대해서 제안하고자 한다. 법률 분야의 문장에는 전문적인 용어가 다수 포함되어 있으므로 기존 키워드 중심의 데이터 검색 방법으로는 전문 용어에 익숙하지 않은 일반인들에게는 적합하지 않다. 대부분 일반인은 자신이 처한 상황을 수 개의 단어로 표현하는 것에 어려움을 가지기 때문이다.

※ 이 논문은 2021년도 서울시 산학연 협력사업(IC210005)의 재원으로 서울R&D지원센터의 지원을 받아 수행된 연구임.
† 준 회 원 : 한국과학기술원 지식서비스공학대학원 석사과정
†† 정 회 원 : 인하대학교 법학전문대학원 전문석사과정
Manuscript Received : November 23, 2021
First Revision : January 17, 2022
Accepted : January 28, 2022
* Corresponding Author : Sung Won Kim(swkim@kaist.ac.kr)

본 논문에서는 한국어의 의미, 맥락을 파악하는 BERT(pre-training of Deep Bidirectional Transformers for Language Understanding)[1] 모델을 기존의 TF-IDF (Term Frequency-Inverse Document Frequency)[2], Universal Sentence Encoder[3]와 같은 임베딩 기법과 결합하여, 키워드 중심의 검색 방법에서 맥락 중심의 검색 방법으로 개선하고자 하였다. 이에 따라 법률 용어를 모르는 일반인도 자신의 상황과 비슷한 맥락을 가진 데이터를 검색할 수 있다.

본 논문에서 제시하고자 하는 것은 다음과 같다. 첫 번째로는 법률 분야에서 문장 간의 의미 유사성을 판단하는 태스크를 수행할 수 있는 BERT 모델의 학습 데이터셋의 구축 방법을 제시한다. 이는 BERT 모델뿐만 아니라 전이학습을 할 수 있는 GPT[4-6], ELMo[7]와 같은 자연어처리 모델에게도 사용할 수 있다. 두 번째로는 법률 분야에서 BERT 모델과 결합하였을 때 가장 좋은 성능을 보일 수 있는 임베딩 기법에 대해서 제안하고, 그에 대한 실험결과를 제시한다.

2장에서는 본 논문에서 제시하는 자연어처리 모델에 대한 관련 연구를 살펴보고, 3장에서는 법률 분야의 문장 간의 의미 유사성을 판단하는 학습데이터셋의 구축 방안에 대해서 제시한다. 4장에서는 본 논문에 쓰이는 각 모델에 대하여 구체적인

구축 방법을 제시하며, 5장에서는 그에 대한 실험결과를 분석한다. 마지막으로 6장에서는 결론과 향후 방향을 제시한다.

2. 관련 연구

2.1 법률 분야 자연어처리 동향

법률 분야의 핵심은 대전제-소전제-결론으로 이루어진 삼단논법의 적용이라고 할 수 있다. 사안에 대한 사실관계를 토대로 관련 법 조항과 판례를 적용하여 결론을 도출하는 일련의 과정을 수행하기 위해서는 필연적으로 많은 정보의 수집과 추론 과정이 요구된다. 자연어처리 기술(NLP)은 이러한 반복적이고 고난도가 요구되는 법률 분야에 효율적으로 적용될 수 있다. 변호사 등 법조인에게는 업무의 효율을 비약적으로 향상하게 시켜주고, 일반인에게는 어려운 법률 분야의 진입 장벽을 낮춰줄 수 있기 때문이다. 국내외를 막론하고 많은 연구자가 자연어처리 기반의 법률 인공지능을 발전시키기 위해 노력해왔으며, 딥러닝 기술의 발전으로 법률 인공지능은 실생활에 적용될 수 있을 정도로 성장했다.

법률 인공지능 연구는 대표적으로 (1)임베딩 기반 방법론과 (2)기호 기반 방법론이 있다[9]. 표현 학습(Repersentation Learning)이라고도 불리는 임베딩 기반 방법론은 법적 사실과 지식을 임베딩 공간에 표현하는 것을 기본으로 한다. 딥러닝 기법을 활용하여 판결 예측, 법률 문답 등의 작업에 효과적이라는 것이 다수의 연구 결과에서 도출된 바 있다[10-12]. 기호 기반 방법론은 구조화된 예측 모델이라고도 불리며, 이름에서 알 수 있듯 사건의 타임라인, 당사자들의 관계 등의 구조로부터 핵심 정보나 법적 요소를 추출한다[13,14].

해외에서는 법률 인공지능이 다양한 분야에 적용되어 실무에 사용되고 있다. 예컨대 미국의 Judicata는 리걸서치 플랫폼을 통해 법조인이 원하는 조건의 판례문을 구조화된 문서로 제공하며, 중국의 메타소타는 일반인을 대상으로 법률 문서 작성을 보조하는 서비스를 제공하고 있다.

국내에서도 법률 자연어처리 기술에 관한 연구와 개발이 활발히 진행되고 있다. 국내 법률 인공지능은 (1)법률 빅데이터 생태계 구축과 (2)법률 인공지능 서비스 개발이라는 두 축을 토대로 활성화되고 있는데, 대체로 법률 빅데이터는 정부 주도로, 서비스는 민간 주도하에 연구개발이 이루어진다는 특징을 가지고 있다.

법률 인공지능의 핵심이 되는 법령, 판결문 등의 데이터는 공개 범위가 협소하고, 또 열람 방식도 오프라인, PDF 형식으로 제한되어 인공지능 개발에 많은 제약이 있었다. 예컨대 대법원 종합법률정보시스템에는 전체 대법원 판결의 3.2%만 공개되어 있고, 판결문 통합검색·열람시스템은 판결문 사본 제공 신청과 비실명화 작업, 수수료 납부 등의 절차를 거쳐야 비로소 데이터를 열람할 수 있다[15]. 최근에는 판결문 공개의 목소리와 리걸테크 산업 발전의 필요성이 높아지자 정부 주도 법률 데이터 구축 사례가 늘어나고 있다. 2018년 한국지능정보사회진흥원은 AI 통합 플랫폼 AIHub에 국민 생활과 밀접한 분야의 법령, 조문, 판례, 법률상담 빅데이터 27만건을 구축하였다[16]. 2021년 법제처는 핵심 추진 과제로 법령 정보 데이터베이스 플랫폼의 구축을 선정했으며, 법령정보 지식베이스를 무료로 개방하여 리걸테크 사업이 활성화될 수 있도록 지원한다고 밝혔다[17].

본 논문에서는 법률 분야 자연어처리 중 일상 언어로 전문적인 법률 분야의 문장과의 의미 유사도를 파악하고, 데이터를 검색할 방안을 제시하고자 한다. 리걸테크 산업의 발전을 전문 법률 전문가뿐만 아니라, 법을 전문으로 배우지 않은 일반 시민들에게도 진입 장벽을 낮출 수 있도록 하고자 한다.

2.2 BERT

최근 자연어처리 분야에서는 사전학습을 한 후 특정 태스크를 전이학습(fine-tuning) 시키는 연구가 활발하게 이루어지고 있다. GPT, ELMo, BERT와 같은 자연어처리 모델들은 당시 다양한 태스크에서 state-of-art한 성과를 보여주었다.

본 논문에서 적용하는 모델은 BERT 모델로 GPT, ELMo와 같은 언어모델과 달리 양방향으로 문맥을 모두 파악할 수 있는 학습 태스크를 제공한다. 또한, 한국어를 포함하여 103개의 언어에 대하여 처리가 가능한 base Multilingual Cased모델을 오픈소스로 제공하고 있어 리소스가 부족한 환경에서도 연구가 용이하다.

BERT 모델의 사전학습 태스크는 크게 두 가지로 이루어진다. 첫 번째는 마스크 언어모델이다. 학습데이터 한 문장 토큰의 15%를 마스킹하고, 마스킹 대상 토큰 가운데 80%는 실제 빈칸으로 만들며 모델은 그 빈칸을 채운다. 이어서 마스킹 대상 토큰 가운데 10%는 무작위로 다른 토큰으로 대체하며, 모델은 해당 위치의 정답 언어가 무엇일지 맞춘다. 또한, 마스킹 대상 토큰 가운데 10%는 그대로 두며, 모델은 해당 위치의 정답 단어가 무엇일지 맞추게 된다. 이런 태스크를 수행하면 앞뒤 문맥을 파악하여 문장 내 어느 자리에 어떤 단어를 쓰는 것이 자연스러운지 파악할 수 있으며, 문장 안에 문법적으로 비문이 존재하는지를 파악하고, 문장 내 모든 단어 사이의 의미적, 문법적 관계를 파악할 수 있게 된다. 두 번째는 Next Sentence Prediction(NSP)이다. 이는 두 문장을 비교하여 두 문장이 이어진 문장인지 아닌지를 반복적으로 학습한다. 문장 서두나 말미의 단어 일부를 삭제하여 일부 문장 성분이 없어도 전체 의미를 이해할 수 있도록 설계하고, max_sequence_length를 설정하여, 학습에 사용되는 최대 길이를 제한함으로써 학습 데이터에 짧은 문장이 포함되어 있어도 성능에 영향을 미치지 않도록 한다.

2.3 TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency)는 단어(t)가 출현하는 빈도와 특정 단어가 등장한 문서(d)의 빈도의 역수를 이용하여 문장 내 각 단어의 중요도를 계산한다. 이를 통해 각 문장을 벡터로 임베딩할 수 있다. TF(Term Frequency)는 출현 빈도가 높을수록 해당 문서에서 중요하다는 의미를 나타내고, DF(Document Frequency)는 전체 문서에서 출현 되는 빈도를 나타내는데, 전체 문서에서 빈도가 높을수록 이는 공통으로 자주 등장하는 단어일 확률이 높

아 키워드가 될 가능성이 작아지게 된다. 이에 따라 중요하지 않은 단어를 파악하기 위하여 그의 역수(Inverse)를 반영하여 임베딩 벡터를 계산한다.

$$idf(d,t) = log(\frac{n}{1+df(t)}) \quad (1)$$

2.4 Universal Sentence Encoder

USE(Universal Sentence Encoder)는 의미론적인 문장 임베딩을 가능하게 하는 모델이다. 자연어처리 분야에서 태스크를 수행하기 위해 제한적인 양의 학습데이터로 인하여 낮은 성능을 가지게 되는 문제를 해결하기 위하여 Transfer Learning을 사용하며, USE는 Word2Vec, GloVe[9]와 같이 단어 단위의 임베딩이 아니라 문장 단위의 임베딩을 통하여 성능 향상을 입증하였다. USE에서는 Transformer[10] 또는 Deep Averaging Network(DAN)[11] 인코더와 같은 2가지의 인코더를 필요에 따라 사용하는데, 본 논문에서는 더 높은 정확도를 보이는 Transformer의 인코더를 사용하는 모델을 구현하였다. Transformer를 사용하는 USE모델은 문맥의 순서와 문맥에 맞는 의미를 고려하여 문장을 임베딩하기 위하여 sub-graph를 사용한다. 또한 한국어를 포함한 16개국어를 한 공간에 임베딩시킨 Multilingual Universal Sentence Encoder[12,13]를 활용하였다.

3. 법률 분야 의미 유사판단 데이터셋

본 논문에서 구축하고자 하는 데이터셋은 두 그룹의 작업자로 구분되어 구축되었다. 첫 번째로는 일반적인 법률 지식을 가지고 있는 작업자이며, 두 번째는 전문적인 법률적 지식을 가지고 있는 작업자이다.

데이터 수집 단계에서는 법무부 문답 지식, 각 지방법원 자주 묻는 질문, 대한법률구조공단 카테고리별 법률 상담 사례 등 공공기관에서 제공하는 법률 상담 사례를 수집하였다. 데이터는 상담 사례의 카테고리로 분류하고, 조회 수가 높은 데이터를 선별하여 데이터셋을 구성하였다. 데이터의 카테고리는 조회 수가 비중이 높은 임대차, 계약 일반, 손해배상, 노동/인사로 구축하였다.

본 논문의 목적은 법률적인 용어가 다수 포함된 한국어 문장 간의 의미 유사성을 정확히 판단함에 있다. 이런 태스크를 하기 수행하기 위해서는 기본적으로 다운태스크(downtask)를 학습할 수 있는 자연어처리 모델 BERT를 전이학습(finetuning)시킬 수 있는 데이터셋이 필요하다.

1차적으로는 일반적인 법률 지식을 가지고 있는 작업자가 데이터셋 내에서 유사한 사례들을 수작업으로 비교한다. 하나의 사례는 최대 10개의 유사한 사례로 매칭이 된다. 이렇게 매칭된 사례 데이터는 변호사, 또는 로스쿨 재학생들에게 검수 과정을 거친다. 전문적인 법률용어가 포함되었기 때문에, 일반인 작업자는 실제적인 의미를 혼용 및 오용을 할 가능성이 있다. 따라서 전문적인 법률적 지식을 가지고 있는 작업자들로부터 라벨링이 제대로 되었는지 검수하는 과정을 마쳤다.

Table 1. Example of Training Data Set

Type	Contents
id	125
sent1	임대차보호법에 보호받을 수 있는 지 여부
sent2	주택 임대차보호법상 대항력의 취득 요건
category	lease
label	1

Table 2. Statistics of Dataset Category

Category	Count	Ratio (%)	#word (1)	#net word (2)	Avg length
Lease	19,833	49.0	135,732	3,091	60
Contract	10,038	24.8	16,530	2,421	26
Compensation	5,909	14.6	12,885	2,530	25
HR	4,695	11.6	92,428	4,354	49

(1) The number of words including duplicate words.
(2) The number of words except duplicate words.
Avg length : Average length of sentences

데이터셋은 총 40,475개로 이루어져 있으며, 학습 데이터와 시험 데이터 비율은 8:2로 분리하였다. Table 1과 같이 sent1과 sent2 간에 문장의 의미가 유사하거나 동일하다면 1로 라벨링하고, 그렇지 않다면 0으로 라벨링을 하였다.

문장의 의미가 유사 또는 동일하다는 것은 작업자에 따라서 주관적인 판단이 영향을 미치기 때문에 일반화시킬 수 있는 작업이 필요하다. 이에 따라 하나의 작업물에 대하여 5명의 다른 작업자가 의미의 유사 또는 동일의 판단을 진행하였고, 3개 이상의 유사 또는 동일의 판단을 받은 문장은 1로 라벨링을 하였다. 이로써 두 문장 간의 의미 유사 또는 동일의 여부에 대해서 최대한의 객관성을 확보하였다.

데이터셋은 조회 수가 높은 4종류의 카테고리에 해당하는 문장들로 구성하였다. Table 2를 보면 데이터셋의 비율은 각각 임대차 49.0%, 계약 일반 24.8%, 손해배상 14.6%, 노동/인사 11.6%로 구성되어 있다.

임대차 카테고리의 데이터 수가 가장 많고, 평균 문장의 길이도 가장 긴 반면, 노동/인사의 데이터에 속한 단어의 종류가 가장 다양하였다. 이는 노동/인사 데이터셋 내에서도 다양한 종류의 법률 분쟁이 존재하여 나타나는 결과이다.

4. 임베딩 및 실험방법

4.1 TF-IDF를 이용한 키워드 기반 임베딩

TF-IDF 임베딩은 단어의 빈도를 기반으로 키워드의 중요성을 산정하고, 이를 기반으로 문장을 벡터 공간으로 임베딩시킨다. 따라서 TF-IDF를 이용한 문장의 임베딩은 키워드 기반 임베딩이며, 두 문장의 벡터 공간상 거리는 각 문장 속 키워드 구성의 유사성을 나타내는 지표가 된다.

Table 3. Example of Preprocessed data for TF-IDF

Origin	Preprocessed Data
상가 계약 만료로 원상복구에 대해	상가 계약 만료 원상복구
계약서작성 없는 부동산 가계약금반환	계약서작성 부동산 계약금 반환
상가 계약 후 사정이 생겨서 못 들어갈 경우 법적 책임 여부	상가 계약 후 사정 경우 법 책임 여부
임대인 명의 상수도 요금 체납 관련하여 임차인의 권리구제 방법	임대인 명 상수도 요금 체납 관련 임차인 권리구제 방법

단어의 빈도를 계산하기 위해서 Table 3과 같이 한국어 형태소 분석기인 MeCab[14]을 활용하여 문장 중 명사만 남기는 방식으로 전처리하였다.

Fig. 1에서 전체 데이터셋 내 문장과 비교 대상인 입력 문장을 TF-IDF 인코더로 임베딩하고, 입력 문장의 임베딩 벡터값과 전체 데이터셋 임베딩 벡터의 코사인 유사도를 산정하여 similarity를 계산한다. 이때 Inverse Document Frequency (IDF)의 Document는 전체 법률 데이터셋 40,475개를 기준으로 한다.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (2)$$

이때 similarity는 0에서 1까지의 값을 가지고, 0은 서로 독립적인 경우를 의미한다. 값이 1인 경우는 완전히 같은 경우로, similarity가 1에 가까운 값일수록 입력 문장과 유사도가 높다.

최종적으로 입력 문장과 유사도가 높은 순서대로 10개의 데이터를 결괏값으로 반환한다.

4.2 USE를 이용한 의미 기반 임베딩

Universal Sentence Encoder(USE)는 단어가 아닌 문장 자체를 의미를 기반으로 임베딩을 하는 모델이다. 다양한 NLP task에 활용될 수 있도록 데이터셋 내 문장들을 임베딩 벡터로 변환한다.

키워드 중심의 단어가 아닌 문장 단위의 임베딩은 문장의 맥락 또는 의미를 기반으로 각 문장 간의 의미상 유사성을 벡터 간의 차이로 도출할 수 있게 한다. 문장 단위로 임베딩을 하기 때문에 TF-IDF와 같은 형태소 분석은 필요로 하지 않는다.

Fig. 2에서 TF-IDF와 같이 전체 데이터셋 내 문장과 비교 대상인 입력 문장을 USE 인코더로 임베딩한다. 임베딩한 벡터로 코사인 유사도를 산정하는 것도 동일하나, 이렇게 산정한 각 거리(angular distance)로 변환하기 위해서 arccos을 취한다.

$$\text{sim}(u, v) = \left(1 - \arccos\left(\frac{u \cdot v}{\|u\| \|v\|}\right)/\pi\right) \quad (3)$$

이는 USE 모델에서 문장 간 코사인 유사도를 산정하여 비교하는 것보다 각 거리(angular distance)로 비교하는 것이 더 좋은 성능이 구현되기 때문이다. similarity는 1에 가까울수록 높은 유사도를 의미한다.

USE 모델 또한 10개의 유사도가 높은 데이터를 유사도를 기준으로 내림차순하여 정렬한 후 결괏값으로 반환한다.

TF-IDF Algorithm

```
Encoder: TF-IDF()
Dataset_Embedding = TF-IDF(dataset)
Input_Embedding = TF-IDF(input)

for data in Dataset_Embedding :
    calculate a similarity between the
    data & Input_Embedding with Cosine Similarity

Return Top 10 high similarity data with the order of high similarity.
(IF similarity is close to 1-> top similarity)

def TF-IDF(INPUT) :
    Vector = Embedding INPUT into vector space
    return Vector
```

Fig. 1. Pseudo Code of TF-IDF Algorithm

USE Algorithm

```
Encoder: USE()
Dataset_Embedding = USE(dataset)
Input_Embedding = USE(input)

for data in Dataset_Embedding :
    calculate a similarity between the
    data & Input_Embedding with angular distance

Return Top 10 high similarity data with the order of high similarity.
(IF similarity is close to 1 -> top similarity)

def USE(INPUT) :
    Vector = Embedding INPUT into vector space
    return Vector
```

Fig. 2. Pseudo Code of USE Algorithm

4.3 실험방법

상기 키워드 기반 임베딩 또는 의미 기반 임베딩을 기반으로 필터링한 유사도가 높은 후보 문장을 BERT 모델에 전달하여 맥락까지 유사한 문장을 필터링하기 위해서는 그에 관한 판단을 수행하는 BERT 모델이 필요하다. BERT 모델이 두 문장의 유사판단을 하기 위해서는 이에 대응하는 태스크를 수행할 수 있는 전이학습이 필요하다.

전이학습의 기초가 되는 모델은 BERT-base Multilingual Cased 모델(기반 모델)로 한국어 Wikipedia를 포함한 다국어 언어처리가 되는 모델이다. 보통 BERT와 같은 언어 자연어처리 분야의 모델은 사전학습(pretrain)을 한 후 각종 전이학습을 진행한다. 한국어 Wikipedia를 사전학습 시켜놓은 기반 모델은 한국어의 의미적, 문법적 관계를 파악할 수 있다. 다만 법률 분야의 전문적인 어휘, 특이적인 어순을 파악

BERT Algorithm

```
IF ENCODER: TF-IDF()
    Dataset_Embedding = TF-IDF(dataset)
    Input_Embedding = TF-IDF(input)

    Calculate a similarity between the
    data & Input_Embedding with Cosine Similarity

    data_index <- Extract the index of data with the
order of high similarity. (Top 10)

IF ENCODER: USE()
    Dataset_Embedding = USE(dataset)
    Input_Embedding = USE(input)

    Calculate a similarity between the
    data & Input_Embedding with angular distance

    data_index <- Extract the index of data with the
order of high similarity. (Top 10)

for data in dataset[data_index]:
    BERT(input, data)
    Extract the data which is TRUE.
    Reorder the data (BERT(TRUE) > BERT(FALSE))

def TF-IDF(INPUT) :
    Vector = Embedding INPUT into vector space
    return Vector

def USE(INPUT) :
    Vector = Embedding INPUT into vector space
    return Vector

def BERT(sent1, sent2):
    IF the meaning of sent1 & sent2 is simlilar or the
same, return TRUE
    ELSE, return FALSE
```

Fig. 3. Pseudo Code of BERT Algorithm

하기 위해서 추가적인 임베딩이 필요하다.

하지만 본 논문에서는 BERT모델 자체의 성능이 아니라 키워드 기반 임베딩 또는 의미 기반 임베딩과 조합하여 사용하였을 때 법률 분야의 검색 성능향상에 미치는 영향을 객관적으로 파악하기 위하여 기반 모델을 기준으로 전이학습을 진행하여 성능을 평가하였다.

전이학습을 통하여 수행하고자 하는 태스크는 2개의 입력 문장이 주어졌을 때 그 문장의 의미가 동일 또는 유사한가(1), 아니면 유사하지 않은가(0)의 이진 분류를 판단하는 것이다.

Fig. 3에서 태스크를 수행하기 위해서 입력 데이터(sent1)와 비교 대상의 문자열 데이터(sent2)가 전달되어야 한다. 본 연구에서 실험한 sent1, sent2의 평균 어절 수는 약 5어절이다. sent2는 TF-IDF 인코더, 또는 USE 인코더를 이용하여 sent1과 유사도가 높은 10개의 문장을 데이터셋에서 검색한다. 10개의 sent2 데이터는 TF-IDF 또는 USE 인코더가 판단한 유사도가 높은 순서대로 정렬되어 BERT 모델로 전달된다.

전달된 sent2 데이터들은 BERT 모델로 각각 유사 여부가 판단되며, BERT 모델이 유사하다고 판단한 데이터는 상기 기재한 TF-IDF, 또는 USE 인코더가 판단한 유사도에 관계없이 높은 우선순위로 재정렬된다. 즉, BERT가 유사하다고 판단한 문장이 먼저 상위 rank에 위치하게 되고, 그다음으로 남은 문장들은 TF-IDF, 또는 USE 인코더가 판단한 유사도가 높은 순서대로 정렬된다.

5. 실험 결과

키워드 기반, 의미 기반의 임베딩 모델을 조합하여 성능평가를 진행하였다. 기존 성능평가의 방법으로는 법률 분야에서 의미의 유사성을 판단하는 성능을 측정할 수 없기 때문에 자체적으로 성능 평가의 지표를 수립하였다.

성능을 나타내는 score는 모델이 제시하는 similarity가 높은 문장 Top-n개 중에서 유사 문장으로 라벨링(1)이 된 문장의 여부를 평가하였다. 1로 라벨링이 된 문장을 포함하고 있는 경우 Hit로 판단하여, 전체 실험갯수 중 Hit의 비율을 산정한다. 1로 라벨링이 된 문장은 일반인과 법률적 지식을 가지고 있는 작업자들이 두 문장의 의미가 동일 또는 유사하다고 판단한 것이다.

Table 5를 보면, Top 7을 사용한 결과 키워드 기반의 임베딩을 수행하는 TF-IDF모델과 BERT모델을 결합한 모델이 가장 좋은 성능을 보였다. 이는 TF-IDF 단독모델보다 상승한 결과이다. 이에 반해 의미 기반의 임베딩을 수행하는 USE 모델과 BERT모델을 결합한 모델은 USE 단독으로 수행한 결과보다 성능이 하락하였다.

데이터를 검색하는 것은 실제 서비스에도 밀접하게 연관되는 태스크이기 때문에, 각 모델이 실행되는 데 걸리는 시간을 비교하였다.

Table 4를 보면 실행시간이 가장 짧은 모델은 TF-IDF 단독모델을 사용하였을 때로, 0.66s가 걸렸다. 이에 반해 가장 긴 모델은 USE 단독모델을 사용하였을 때이다. USE 모델은

Table 4. Time Comparison between Models

Model	Time
TF-IDF	**0.66s**
USE	**6.01s**
TF-IDF + BERT	0.66s
USE + BERT	5.97s

Table 5. Scores of Model for Legal Fields(Top n)

n	Model	Score
3	TF-IDF	67.14
	TF-IDF + BERT	69.05
	USE	**39.52**
	USE + BERT	44.29
5	TF-IDF	75.24
	TF-IDF + BERT	78.57
	USE	59.52
	USE + BERT	57.14
7	TF-IDF	79.05
	TF-IDF + BERT	**80.95**
	USE	65.24
	USE + BERT	65.71

전체 데이터셋에 대하여 의미 기반으로 임베딩을 하는 과정에서 많은 시간을 소요한다. 이에 따라 USE가 포함된 모델은 포함되지 않은 모델들에 비해 긴 실행시간을 보인다.

6. 실험 분석

6.1 TF-IDF vs USE

TF-IDF를 인코더로 사용하였을 때 USE 단독으로 사용하는 모델에 비해 13.81점의 성능 차이로 더 좋은 성능을 보여주고 있다. 이는 실제 실험결과를 바탕으로 각 모델이 어디에 초점을 두어 유사도를 판단하는지 짐작할 수 있다.

입력 데이터로 '무단퇴사를 했는데, 손해배상 청구할 수 있나요?'를 전달하였을 때 각각의 모델이 판단한 Top 10 문장은 Table 6, 7과 같다. 작업자에 의하여 유사하다고 판단된 문장은 1로 라벨링이 되었다.

Table 6의 결과를 보면, TF-IDF 모델은 키워드 기반으로 유사도를 판단하기 때문에 '무단퇴사'와 '손해배상'에 대한 가중치를 높게 판단하는 것으로 보인다. 이에 반해 Table 7을 보면 USE 모델은 '무단퇴사'에 대한 가중치를 TF-IDF 모델에 비해서 적게 판단하나, '손해배상'이라는 단어가 아닌 '손해보상'이라는 동음이의어에 대한 데이터도 검색하는 것으로 보인다.

두 모델 모두 '손해배상'에 대한 문장에 대해 유사도를 높게 판단하나, TF-IDF 모델은 수 개의 손해배상 관련 문장 중에서 '무단퇴사'라는 키워드와 관련된 문장을 유사도 높게 판단하기 때문에, 인간에게 더 실효성 있는 검색결과를 도출한다.

Table 6. Top 10 High Similarity Sentence with TF-IDF
sent1 : '무단퇴사를 했는데, 손해배상 청구할 수 있나요?'

Priority	Sent2	Label
1	무단퇴사를 했는데 손해배상 청구 가능한지	1
2	무단퇴사(손해배상 청구)	1
3	알바가 무단 퇴사 했는데 손해 배상 청구할 수 있나요?	1
4	손해배상 청구 가능한가요	1
5	무단퇴사 직원의 손해배상 청구	1
6	무단퇴사한 직원 손해배상청구가 가능한가요?	1
7	전세계약해지에 대한 손해배상 청구가 가능한지 문의	0
8	손해배상청구는 어떻게 하나요?	0
9	손해배상청구를 할 수 있나요?	1
10	손해배상청구를 받을 수 있나요?	0

Table 7. Top 10 High Similarity Sentence with USE
sent1 : '무단퇴사를 했는데, 손해배상 청구할 수 있나요?'

Priority	Sent2	Label
1	무단퇴사를 했는데 손해배상 청구 가능한지	1
2	무단퇴사 손해배상청구가 가능한가요?	1
3	손해배상 청구 가능한가요	1
4	손해배상청구를 당했을 때, 청구할 수 있는 게 있나요?	0
5	손해배상 청구하는 방법	1
6	손해배상청구를 할 수 있나요?	1
7	의료사고에 대한 손해배상 받을 수 있나요?	0
8	손해배상 청구를 하려고 하는데, 어디까지 받을 수 있을까요?	1
9	손해보상 받을 수 있을까요?	1
10	업무방해로 손해배상을 청구 가능한가요?	0

6.2 TF-IDF vs TF-IDF + BERT

TF-IDF 단독모델이 키워드 기반으로 유사한 문장을 검색할 수 있으나, 문장이 짧은 경우 정확도가 떨어질 수 있다는 문제점을 가지고 있다. 입력 데이터로 '월세집 곰팡이 수선에 대한 책임'이라는 문장을 줬을 때 TF-IDF 단독모델을 검색한 결과는 Table 8과 같다.

임차인의 곰팡이에 대한 수선을 묻고 있는 경우, 1,2,4,7번 째 문장이 높은 우선순위를 가져야 할 것이다. 하지만 TF-IDF의 경우 3번 문장처럼 문장의 길이가 짧은 경우, '월세'에 대한 가중치를 높게 측정하여 우선순위가 높아지는 문제를 가지고 있다.

이런 문제를 보완하기 위하여, 문장의 맥락적 유사도를 판단할 수 있도록 전이학습 시킨 BERT 모델을 결합한다. BERT 모델은 TF-IDF 모델이 판단한 유사한 문장 각각을 2차적으로 유사 여부를 판단하여 우선순위를 재정렬할 수 있다. BERT 모델을 결합하였을 때 우선순위가 재정렬된 검색결과는 Table 9와 같다.

Table 8. Top 10 High Similarity Sentence with TF-IDF
sent1 : '월셋집 곰팡이 수선에 대한 책임'

Priority	Sent2	Label
1	**월세집 곰팡이 수선에 관하여**	**1**
2	**곰팡이 월세 계약 중도해지**	**1**
3	못받은 월세 받는 방법	0
4	**월세 집 곰팡이로 인한 책임 문제**	**1**
5	월세집에 살기 힘든데 보증금과 월세를 돌려받을 수 있나요?	1
6	올바른 수선이 아닌 방법으로 수선할 시 계약해지가 가능한지	0
7	**월세 곰팡이 때문에 이사비용청구**	**1**
8	월세 보증금을 덜 받았는데 다 받는 방법	0
9	월세 보증금 1000만 원 중 800만 원만 돌려받았습니다.	0
10	월세 보증금 돌려받을 수 있을까요?	0

Table 9. Top 10 High Similarity Sentence with TF-IDF + BERT
sent1 : '월세집 곰팡이 수선에 대한 책임'

Priority	Sent2	Label	BERT result
1	**월세집 곰팡이 수선에 관하여**	**1**	**1**
2	**곰팡이 월세 계약 중도해지**	**1**	**1**
3	**월세 집 곰팡이로 인한 책임 문제**	**1**	**1**
4	**월세 곰팡이 때문에 이사비용청구**	**1**	**1**
5	못 받은 월세 받는 방법	0	0
6	월세집에 살기 힘든데 보증금과 월세를 돌려받을 수 있나요?	1	0
7	올바른 수선이 아닌 방법으로 수선할 시 계약해지가 가능한지	0	0
8	월세 보증금을 덜 받았는데 다 받는 방법	0	0
9	월세 보증금 1000만 원 중 800만 원만 돌려받았습니다.	0	0
10	월세 보증금 돌려받을 수 있을까요?	0	0

Table 10. Top 10 High Similarity Sentence with USE
sent1 : '부당해고를 당한 근로자가 회사에 손해배상을 청구할 수 있는가요?'

Priority	Sent2	Label
1	**부당해고를 당한 근로자가 회사에 손해배상을 청구할 수도 있는가요?**	**1**
2	**부당해고를 당한 근로자가 회사에 위자료를 청구할 수도 있는가요?**	**1**
3	회사에서 업무상 과실 직원에게 손해배상 청구 가능한가요?	0
4	업무방해로 손해배상을 청구할 수 있을까요?	0
5	근로자 간 폭행의 경우에도 사용자에게 손해배상을 청구할 수 있나요?	0
6	해고를 당한 직원이 무효확인소송을 제기하는 외에 사용자에게 손해배상을 청구할 수도 있는가요?	1
7	**부당해고를 당한 근로자가 회사에 정신적 손해를 배상하라고 할 수도 있는가요?**	**1**
8	회사에 손해배상을 해주어야 하더라도 일단은 퇴직금을 전부 받을 수 있는가요?	0
9	해고를 당한 근로자가 회사를 상대로 임금지급가처분을 신청할 수도 있는가요?	1
10	**명백한 부당해고를 당해 억울한데 회사를 상대로 위자료를 받을 수 있나요?**	**1**

BERT 모델을 결합하여 의미가 유사한 문장을 상위 rank에 위치시키어 재정렬한 것을 확인할 수 있다. 또한, BERT 모델은 의미가 유사하지 않은 문장들에 대해서도 0으로 판단하여 필요에 따라 검색결과에서 제외할 수 있다. 실제 검색결과라고 가정했을 시 BERT 모델을 결합한 모델이 실제로 더 양질의 검색결과를 제공하는 것을 알 수 있다.

6.3 USE vs USE + BERT

USE 모델은 맥락기반으로 키워드 자체가 아닌 문장유사도를 판단하는 모델이다. 따라서 단어 자체가 똑같지 않더라도 비슷한 의미가 있는 문장을 비슷한 위치로 임베딩 시킬 수 있다. 그러나 이런 경우 특정 키워드에 대한 검색 성능이 떨어질 수 있다는 단점이 있다. 입력 데이터로 '부당해고를 당한 근로자가 회사에 손해배상을 청구할 수 있는가요?'라고 전달하였을 때, USE 단독모델로 검색한 결과는 Table 10과 같다.

Table 10의 검색결과를 보면, '손해배상'과 의미가 유사한 '위자료', '정신적 손해를 배상'이라는 구절뿐만 아니라 어순이 비슷한 문장들을 검색한다는 것을 유추할 수 있다. 그러나 부당해고에 대한 특정 키워드가 중요한 경우임에도 7,10과 같이 인간이 판단하였을 때 실제로 의미 유사도가 높은 문장들을 상대적으로 후순위에 배치하는 것을 확인할 수 있다.

상기 서술한 TF-IDF + BERT 모델과 같이 sent2를 BERT로 다시 판단하여 좀 더 구체적인 의미가 유사한 문장을 상위 우선순위를 가질 수 있도록 재정렬할 수 있다. USE 모델에 BERT 모델을 결합한 검색결과는 Table 11과 같다.

Table 11의 5번 문장을 제외하고 실제 의미가 유사한 문장들이 상위 rank로 재정렬되었다. 실제 유사도가 높은 데이터를 높은 순위로 보여주는 검색결과 방식은 양질의 검색 서비스를 제공하는 데 적합하다.

Table 11. Top 10 High Similarity Sentence with USE+BERT
sent1 : '부당해고를 당한 근로자가 회사에 손해배상을 청구할 수 있는가요?'

Priority	Sent2	Label	BERT result
1	**부당해고를 당한 근로자가 회사에 손해배상을 청구할 수도 있는가요?**	**1**	**1**
2	**부당해고를 당한 근로자가 회사에 위자료를 청구할 수도 있는가요?**	**1**	**1**
3	**부당해고를 당한 근로자가 회사에 정신적 손해를 배상하라고 할 수도 있는가요?**	**1**	**1**
4	**명백한 부당해고를 당해 억울한데 회사를 상대로 위자료를 받을 수 있나요?**	**1**	**1**
5	회사에서 업무상 과실 직원에게 손해배상 청구 가능한가요?	0	1
6	업무방해로 손해배상을 청구할 수 있을까요?	0	0
7	근로자 간 폭행의 경우에도 사용자에게 손해배상을 청구할 수 있나요?	0	0
8	해고를 당한 직원이 무효확인소송을 제기하는 외에 사용자에게 손해배상을 청구할 수도 있는가요?	1	0
9	회사에 손해배상을 해주어야 하더라도 일단은 퇴직금을 전부 받을 수 있는가요?	0	0
10	해고를 당한 근로자가 회사를 상대로 임금지급가처분을 신청할 수도 있는가요?	1	0

7. 결 론

본 논문에서는 법률 분야의 문장 간 의미 유사성을 판단하는 학습 데이터셋을 구축하였다. 아울러 법률 분야의 문장 간 유사성을 바탕으로 데이터를 검색하는 데 최적화된 임베딩 기법을 실험하였다. 키워드 기반 임베딩, 의미 기반 임베딩과 더불어 BERT와 같은 전이학습이 가능한 자연어처리 모델을 조합하여 실험하였고, 그 결과 키워드 기반 임베딩 기법인 TF-IDF 모델과 BERT모델을 결합한 모델이 가장 좋은 성능을 보였다.

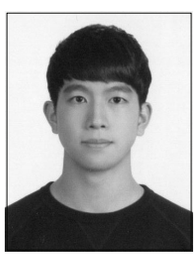
본 논문에 쓰인 BERT 모델은 Wikipedia를 사전학습시킨 기반 모델을 바탕으로 실험을 진행하였다. 하지만 법률 분야에 쓰이는 용어, 그리고 어순은 일상용어와는 다소 차이가 있으므로 법률 분야의 대용량 말뭉치로 사전학습을 한 후 전이학습을 진행하는 것이 성능향상에 도움이 된다. 이에 대한 추가적인 실험은 추후 계획되어 있다.

우리는 본 논문을 통하여 법률 분야에서 자연어처리를 어떠한 방식으로 접근하여야 하는지 방향성을 제시하고자 하였다. 전문적인 분야더라도, 양질의 학습 데이터셋을 구축하고, 기존의 모델들을 조합하여 실제 서비스에 적용할 수 있을 만큼의 충분한 성능을 이끌어 낼 수 있다. 본 논문이 제시한 방향성이 리걸테크의 자연어처리 분야에서 다양한 연구가 이루어지길 바란다.

References

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] J. Ramos, "Using tf-idf to determine word relevance in document queries," In *Proceedings of the First Instructional Conference on Machine Learning*, Vol.242, No.1, pp.29-48, 2003.
[3] Cer, D. et al., "Universal sentence encoder," arXiv preprint arXiv:1803.11175, 2018.
[4] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.
[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, Vol.1, No.8, pp.9, 2019.
[6] T. B. Brown, et al., "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.
[7] M. E. Peters, et al., "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018
[8] T. M. Mitchell, "Bayesian Learning," Machine Learning, McGraw-Hill, pp.154-200, 1997.
[9] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How Does NLP benefit legal system: A summary of legal artificial intelligence," *arXiv: 2004.12158*, 2020.
[10] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext. zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.
[11] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.
[12] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation," In *Proceedings of Empirical Methods in Natural Language Processing*, pp.1532-1543, 2014.
[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," In *Proceedings of Northern American Chapter of the Association for Computational Linguistics*, 2019.
[14] A. Akbik, T. Bergmann, and R. Vollgraf, "Pooled contextualized embeddings for namedentity recognition," In *Proceedings of Northern American Chapter of the Association for Computational Linguistics*, 2019.
[15] "판결문 공개 과감히 확대하라," 법률신문, 2019.10.28. [Internet], https://m.lawtimes.co.kr/Content/Article?serial=156740
[16] "법률 지식 베이스 소개" AI허브, 2018.01.02. [Internet], https://aihub.or.kr/aidata/29
[17] "법제처 보도자료," pp.4-5, 2021.12.29.
[18] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp.1532-1543, 2014.
[19] A. Vaswani, et al., "Attention is all you need," In *Proceedings of Neural Information Processing Systems*, 2017.
[20] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III. "Deep unordered composition rivals syntactic methods for text classification," In *Proceedings of Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2015.
[21] Y. Yang, et al., "Multilingual universal sentence encoder for semantic retrieval," *arXiv preprint arXiv:1907.04307*, 2019.
[22] "universal-sentence-encoder-multilingual," TensorFlow hub. last modified Jan 05, 2022. accessed Oct. 21, 2021. [Internet], https:// tfhub.dev/google/universal-sentence-encoder-multilingual/3
[23] "mecab-ko-dic," bitbucket. last modified Jul. 20, 2018. accessed Oct. 21, 2021. [Internet], https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/



김 성 원
https://orcid.org/0000-0001-8605-2618
e-mail : swkim@kaist.ac.kr
2021년 고려대학교 건축사회환경공학부 (학사)
2022년 ~ 현 재 한국과학기술원 지식서비스공학대학원 석사과정
관심분야 : Machine Learning, Graph Neural Network

박 광 렬
https://orcid.org/0000-0001-6924-3666
e-mail : jamespark0418@gmail.com
2021년 고려대학교 건축사회환경공학부 (학사)
2022년 ~ 현 재 인하대학교 법학전문대학원 전문석사과정
2022년 ~ 현 재 한국AI소프트(주) 대표이사
관심분야 : LegalTech, Natural Language Processing