JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 11, pp. 57-66, November 2022
https://doi.org/10.9708/jksci.2022.27.11.057

Legal search method using S-BERT

Gil-sik Park*, Jun-tae Kim**

*research professor, Industry-Academic Cooperation Foundation, KwangWoon University, Seoul, Korea
**Professor, Dept. of Computer Engineering, Dongguk University, Seoul, Korea

[Abstract]

In this paper, we propose a legal document search method that uses the Sentence-BERT model. The general public who wants to use the legal search service has difficulty searching for relevant precedents due to a lack of understanding of legal terms and structures. In addition, the existing keyword and text mining-based legal search methods have their limits in yielding quality search results for two reasons: they lack information on the context of the judgment, and they fail to discern homonyms and polysemies. As a result, the accuracy of the legal document search results is often unsatisfactory or skeptical. To this end, This paper aims to improve the efficacy of the general public's legal search in the Supreme Court precedent and Legal Aid Counseling case database. The Sentence-BERT model embeds contextual information on precedents and counseling data, which better preserves the integrity of relevant meaning in phrases or sentences. Our initial research has shown that the Sentence-BERT search method yields higher accuracy than the Doc2Vec or TF-IDF search methods.

▸ **Key words**: Legal Service, Machine Learning, Deep Learning, BERT, Data Mining

[요 약]

본 논문에서는 Sentence-BERT 모델을 활용한 법률 문서 검색 방법을 제안한다. 법률 검색 서비스를 이용하고자 하는 일반인들은 법률 용어 및 구조에 대한 이해가 부족함에 따라 관련 판례 검색 등에 있어 어려움을 겪고 있다. 기존의 키워드 및 텍스트마이닝 기반 법률 검색 방법은 판결문의 문맥에 대한 정보가 없으며, 동음이의어 및 다의어에 대해 구분하기 어려워 성능을 높이는데 한계가 있었다. 그로 인해 법률 문서 검색 결과에 대한 정확도가 낮아 신뢰하기가 어려웠다. 이를 위해, 대법원 판례 및 법률구조공단 상담사례 데이터에서 일반인의 법률 검색 문장에 대한 성능을 개선하고자 한다. Sentence-BERT 모델은 판례 및 상담 데이터에 대한 문맥 정보가 임베딩되므로, 문장의 의미 손실이 적어 TF-IDF 및 Doc2Vec 검색 방법과 비교했을 때보다 검색 정확도가 개선된 것을 확인할 수 있었다.

▸ **주제어**: 법률 서비스, 머신러닝, 딥러닝, BERT, 데이터마이닝

• First Author: Gil-sik Park, Corresponding Author: Gil-sik Park
*Gil-sik Park (gspark@kw.ac.kr), Industry-Academic Cooperation Foundation, KwangWoon University
**Jun-tae Kim (jkim@dongguk.edu), Dept. of Computer Engineering, Dongguk University
• Received: 2022. 10. 12, Revised: 2022. 11. 07, Accepted: 2022. 11. 07.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 들어 인공지능을 이용한 법률 문서 이해, 문서 검색, 판결 예측, 법률 문서 분류, 요약 등 다양한 형태로 법률과 관련된 태스크가 진행되고 있다[1]. 이 태스크들은 법률 분야에 종사하는 법률 인들에게 복잡하고 반복적인 업무의 최소화, 판결의 정확도 향상 등 도움이 될 것으로 예상된다. 이와 함께 법 관련 전문가가 아닌 일반인 당사자들이 직접 소송을 진행하는 '나 홀로 소송' 또한 많아지는 추세다. 요즈음 진행되고 있는 민사사건의 약 60~70%가 '나 홀로 소송'에 해당한다고 보고 되고 있다[2].

하지만, 나 홀로 소송 과정에서는 법률 관련 전문가의 도움이 없으므로 법률 문서에 대한 정확한 이해와 해석, 관련 판례, 법률 및 판결내용 참조가 필요하다. 이 과정에서 법률 용어에 대한 정확한 의미에 대한 이해 부족과 까다로운 재판 용어 등 난해한 판결문 용어는 나 홀로 소송을 더욱 어렵게 만들고 있다. 이처럼 법률에서 사용되는 용어는 일반인들의 용어와 일치하지 않는 경우가 많아 나 홀로 소송을 위한 사이트 (pro-se.scourt.go.kr)나 혼자 하는 소송 법률지원센터 사이트(support.klac.or.kr)를 이용하는데도 어려움이 따른다.

예컨대 '근로'라는 용어는 '노동'이라는 용어로도 많이 사용되지만, 법률에서 쓰는 용어는 근로시간, 근로계약서, 근로기준법 등 '근로'로만 사용이 되고 있어 일반인의 법률 검색에 어려움을 줄 수밖에 없다. '노동자'는 '노동을 제공하고 얻는 임금으로 생활을 유지하는 사람'이며, '근로자'는 '근로에 의한 소득을 생활하는 사람'으로서 의미는 거의 같지만, 실제 키워드 기반 검색에서는 전혀 다른 의미의 단어처럼 취급되어 풍부하고 정확한 검색 결과를 도출하는데 어려움이 따른다. 이처럼 키워드를 기반으로 하는 법률 정보 검색 방식은 일상 용어와 법률 용어가 서로 일치하지 않아 검색 결과의 정확도가 낮으며, 법률 용어의 다의성과 문맥을 반영하지 못한 검색이기 때문에 효율적이지 못하다.

이에 따라 최근 딥러닝 기반 문서 검색이 연구되고 있다. 하지만 딥러닝 기반 모델은 일반인의 자연스러운 다양한 질문이 입력될 때마다 모델 내에서 매번 연산을 수행해야 하므로 법률 문서와 같은 대량의 문서에 대해서는 검색하는데 많은 연산이 필요하다는 단점이 있다. 이러한 한계를 극복하기 위해 임베딩 모델이 연구되었는데, 임베딩 기법은 대량의 법률 문서를 벡터 공간에 벡터로써 나타낸 다음, 일반인의 법률 질의를 입력받아 독립적인 벡터 공간에 벡터로 나타내고, 두 벡터 간 유사도를 구하여 법률 질의에 가장 유사한 법률 문서를 검색할 수 있게 한다[3,4,5]. 임베딩 방식은 사전에 법률 문서에 대한 임베딩을 수행하게 되므로 속도는 빠르지만, 임베딩 과정에서 문서가 가지고 있는 중요한 정보 등이 손실될 수 있다는 단점이 있다.

위와 같은 문제점들을 극복하기 위해 본 논문이 제안하는 법률 검색 방법은 S-BERT를 활용한 새로운 법률 문서 검색 모델이다. 본 논문에서 제안하는 방법은 다음과 같다.

첫 번째로, 법률 상담사례를 수집하기 위해 대한법률구조공단(https://www.klac.or.kr/legalinfo/counsel.do) 사이트에서 약 520여 건의 노동 관련 상담사례에 관한 질문, 답변, 제목 등을 수집한다.

두 번째로, 노동 관련 상담사례 법률 문장에 대한 형태소 분석, 띄어쓰기, 불용어 제거, 맞춤법 검사 및 수정 등 텍스트 전처리를 수행한다.

세 번째로, 국가법령정보센터(http://www.law.go.kr/)로부터 1947년에서 현재까지 진행된 대법원판결 관련 약 84,470여 건의 판결문을 스크래핑하여 데이터 셋으로 저장한다. 그런 다음, 전처리 된 노동 관련 상담사례 법률 문장과 함께 각각 S-BERT 임베딩한다.

네 번째로, S-BERT 임베딩 된 노동 관련 상담사례 법률 문장과 대법원판결 문장 간에 코사인 유사도를 구한 후, 유사도가 가장 높은 판결문과 판결 정보를 추출하여 질의자에게 제공한다.

제안된 법률 검색 방법은 법률 용어가 아닌 일상 용어로 작성한 법률 상담사례 문서에 대해서도 유사도가 높고 풍부한 판례 및 판결 정보를 검색할 수 있다. 따라서 법률 관련 지식이 없는 일반인들도 본인에게 필요한 판례 정보를 제공 받을 수 있도록 하는데 본 논문은 의의가 있다.

본 논문의 구성은 다음과 같다. 2장에서는 법률 문서 검색 방법 및 BERT 임베딩과 관련한 선행 연구를 분석한다. 3장에서는 S-BERT 기반 법률 검색 방법에 대한 전체적인 시스템의 동작과 관련 기능 등을 설명한다. 4장에서는 제안하는 모델의 실험 결과를 분석하고, 5장에서는 결론 및 향후 연구 방향에 대해 논의한다.

II. Related Work

1. Rule-based legal search system

해외의 경우, 인공지능 기술을 이용한 법률 검색 방법은 1970년대부터 시작하여 연구가 진행됐으며, 초기에는 주로 법률 규정에 따른 인공지능 추론 기반의 자동 계산 시스템이 주류였다[6][7].

국내의 경우, 규칙 기반 생활 법령 안내 시스템을 의사결정나무(decision tree)를 이용하여 제작하였다. 이것은 법률 분야에 관한 질문 및 답변을 미리 준비해 두고, 사용자의 질문이 입력되면 '예' 또는 '아니오'로 답변하도록 설계하여 적절한 답변을 찾아가도록 하는 규칙 기반 법률 자문 방식이다[6][8].

규칙 기반 시스템은 사전에 법률 전문가의 지식을 바탕으로 설계되기 때문에 사용자의 다양한 질의에 대한 적절한 답변을 모두 출력하기에는 상당한 한계가 따를 수밖에 없었다.

2. Legal search system based on text mining

법률 문서 검색 시 키워드 기반 검색으로 가장 일반적으로 활용되는 텍스트 마이닝 기법으로 Term Frequency (TF)와 Inverse Document Frequency(IDF)를 이용하는 방법이 있다. TF는 문서 내에서 특정 단어가 언급된 빈도수를 나타내며, 빈도수가 클수록 문서 내에서 특정 단어가 중요한 역할을 한다고 할 수 있다. DF는 전체 문서에서 특정 단어가 언급된 문서의 개수를 나타내며, 빈도수가 클수록 여러 개의 문서에서 공통으로 흔하게 등장한 단어임을 나타내며 오히려 특정 단어의 중요도가 떨어지게 된다. TF-IDF 값은 TF와 IDF의 역수를 서로 곱하여 구하게 되는 수치로서, 값이 클수록 문서 내에서 특정 단어가 어느 정도로 중요하게 사용되고 있는지를 나타낸다(수식 1 참조).

$$tfidf_{i,j} = tf_{i,j} * \log(\frac{N}{df_i}) \quad (1)$$

TF-IDF 기반 검색은 검색어가 문서 내에 존재하지 않을 때는 검색을 수행할 수가 없다. 예컨대 '근로, '노동'과 같이 단어의 의미는 유사하나 표현이 다른 경우에, '근로'를 검색했을 때는 '노동'이 포함된 문서를 검색하기가 어렵다는 단점이 있다. 이처럼 동의어, 다의어에 취약한 방법이 키워드 기반 검색이라고 할 수 있다.

다른 텍스트마이닝 기법으로 LDA(Latent Dirichlet Allocation)는 토픽 모델링의 대표적인 알고리즘이라고 할 수 있다. LDA 알고리즘에서 문서는 여러 개의 토픽이 모여 구성되어 있다고 가정하고, 각각의 토픽은 확률 분포에 따라 단어들이 생성되는 것으로 간주한다. 따라서, 임의의 문서가 입력되면 LDA 모델은 문서가 생성되는 과정을 역추적하여 토픽을 추출할 수 있게 된다.

LDA 토픽 모델링을 이용한 판례 검색 및 분류 방법의 연구에서 저자들은 LDA 알고리즘을 바탕으로 생성한 확률 모델로 주제를 추출하고, 코사인 유사도를 계산하여 유사 판례 그룹을 분류하였다[9]. 위 방법은 판례별로 하나의 토픽으로만 분류할 수 있다는 한계점을 가지고 있다. 또한, 법률 용어와 일반용어 간 매칭을 위해 포털사이트의 블로그 글을 활용한 연구가 있었다[10]. 저자들은 일반인 질의문에 대한 단어와 법령 대응 확률을 학습하여 적절한 법령을 매핑해서 제공하는 시스템을 제안하였다. 그러나 이 시스템은 일반인들이 사용하는 용어에 대한 적절한 전처리 및 정제되지 않은 단어들로 매핑하도록 함으로서 한계가 있었다.

3. Embedding-based legal search system

텍스트 분석을 위해 자연어 처리 분야에서는 단어 또는 문서 등을 연산할 수 있도록 수치로 변환하게 된다. 이 작업을 임베딩이라고 하는데, 임베딩하는 알고리즘에 따라 단어 또는 문서가 표현되는 수치가 다양하게 된다. 사람들의 언어를 기계가 분석할 수 있도록 벡터 표현(vector representation) 방법에 관한 다양한 연구가 진행 중이다[11]. 단어를 벡터로 표현하는 방법으로 초기 연구 모델에는 n 자리의 단어 벡터를 만들고, 특정 단어가 해당하는 위치에 1을 저장하고, 나머지 위치에는 0을 저장하는 one-hot encoding 방식이 있다. 이 방식은 특정 단어가 본질적으로 다른 단어와 어떤 점에서 차이를 갖는지 표현할 수 없다는 점에서 문제가 있다. 이러한 문제를 해결하기 위해 단어가 갖는 의미를 다차원 벡터 공간에다가 벡터로써 표현하는 방식을 사용하게 된다. '비슷한 분포를 가진 단어들은 비슷한 의미를 가진다'라는 언어학의 가정에 따라 1990년대부터 여러 모델이 제안되었는데 2000년대에 와서 neural network의 학습 원리에 기반을 두는 'NNLM(Neural Network Language Model)' 이 만들어졌다[12]. NNLM은 n-gram의 언어 모델을 신경망으로 구현한 후 목표 단어의 앞/뒤 단어들을 입력받아 목표 단어들과 의미상으로 연관성을 갖도록 학습시키는 방식이다. NNLM은 학습 과정에서 각 계층을 잇는 가중치 변수들에 대해 연산하는 과정에서 많은 시간이 소요되는 문제가 있었다[13].

3-1 Word2Vec & Doc2Vec

2013년에 제안된 Word2Vec[3]은 neural network 구조에서 hidden layer를 줄이므로 학습 속도와 정확도를 높일 수 있는 순전파 및 역전파 개념을 도입하였다. Word2Vec으로 학습된 단어의 벡터 표현은 다양한 자연어 관련 태스크에서 우수한 성능을 보여주었다.

Word2Vec은 CBOW와 skip-gram 방식 알고리즘으로 구현된다. CBOW는 문장을 구성하는 여러 단어 중, 주변 단어들의 중심에 들어갈 적절한 단어를 예측하는 모델링 알고리즘이고, skip-gram은 CBOW와는 반대로 문장을 구성하는 여러 단어 중, 중심 단어를 입력받은 다음에 주변 단어를 예상하는 모델이다(Fig. 1. 참조). Word2Vec과 같은 단어 기반 임베딩 모델은 단어들을 벡터로 나타낸 다음, 단어 벡터들 사이의 관계 추론 및 유사성 계산이 가능해졌다는 점에서 의미가 있다. 단어 기반 임베딩 모델은 다양한 분야에서 활용이 되고 있으며, 특히 문서 분류, 감성 분석 등에서 활발하게 이루어지고 있다.

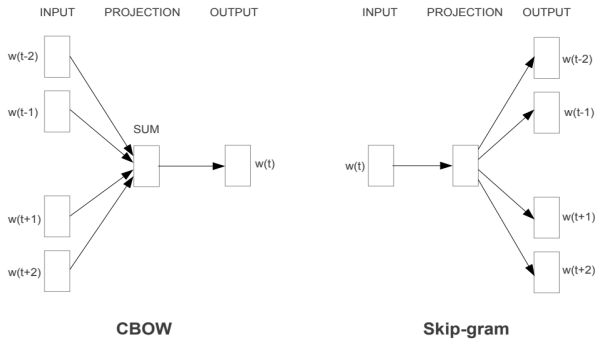


Fig. 1. Word2Vec Architecture

Word2Vec이 단어 단위로 임베딩하는 방법이라면, Doc2Vec은 문서 단위로 임베딩하는 방법이다. Doc2Vec은 감정 분류, 문서 분류 등 자연어 처리와 관련한 다양한 작업에서 높은 성능을 가져왔다[14]. Doc2Vec은 Word2Vec에서 문장이나 문서 단위로 임베딩한 벡터 간의 거리를 계산하여 문서 간 유사도를 구할 수 있도록 하는 알고리즘이다. 이를 위해 Doc2Vec은 문서 또는 문장을 한 개의 id 값을 갖는 단어처럼 취급한 다음 학습을 수행하게 된다. 그에 따라 문서 id는 일반적인 단어들처럼 임베딩 벡터 공간에 하나의 위치 좌표값을 갖게 되는 방식이다(Fig. 2. 참조). 즉, 문서에 등장하는 단어와 함께 문서 id도 함께 임베딩이 되는 방식이다. 이를 통해 Doc2Vec은 문서 간 유사도를 계산하는 데 있어 성능을 높일 수 있었다[14]. 그러나, 법률 검색 알고리즘으로 문서 간 유사도를 측정했을 때 Doc2Vec은 다른 알고리즘에 비해 대체로 낮은 정확도를 가진 것을 확인할 수 있었다. 이는 문서와 단어의 개수가 늘어날수록 성능이 저하되고, 특히 동음이의어 및 다의어에 대한 좀 더 상세한 정보가 함께 임베딩되지 않은 문제에 기인한다고 볼 수 있다.

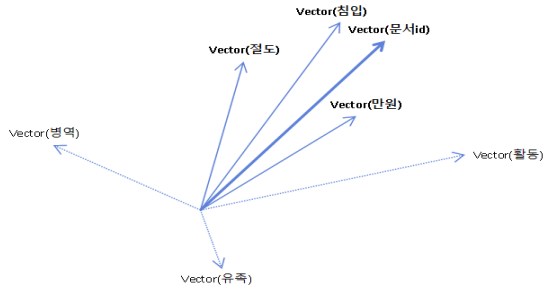


Fig. 2. Document id vector embedding example in Doc2Vec

3-2 BERT

최근 자연어 처리 분야에서는 사전 훈련을 한 다음 특정 태스크에 대해 fine-tuning 하는 연구가 활발하게 이루어지고 있다. 기존의 RNN 및 LSTM과 같은 모델은 문장 학습 시 단어의 순서에 따라 단방향으로 읽어서 학습하므로, 이전 단어에 대해서만 참조하여 모델링하는 단점이 있었다. 이것을 해결한 방식이 BERT와 같은 모델이며 ELMo, GPT 등 다양한 모델들이 자연어 처리 태스크에서 SOTA를 보여주었다. 키워드 검색의 단점을 극복하기 위해 구글에서 공개한 BERT(Bidirectional Encoder Representations from Transformer) 모델[15]은 전 세계 104개 언어에 대한 대용량 위키백과 데이터로 생성된 모델이다. BERT는 레이블이 없는 데이터를 바탕으로 사전학습(pre-trained) 모델을 생성한 다음, 레이블을 가지고 있는 다양한 작업에서 추가로 학습을 진행하면서 가중치를 수정하는 방식으로 모델을 생성하게 된다(Fig. 3. 참조).

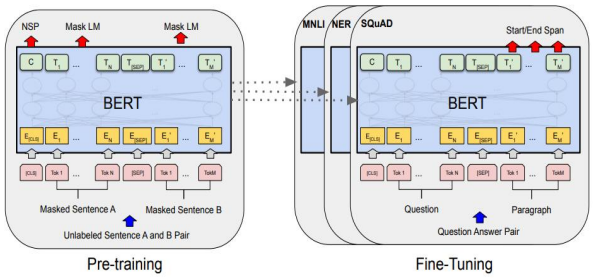


Fig. 3. BERT learning method

BERT 모델은 마스크드 언어 모델(Masked LM)과 다음 문장 예측 모델(Next Sentence Prediction, NSP)로 구성된다. 마스크드 언어 모델은 학습하고자 하는 문장에서 랜덤하게 15%에 해당하는 임의의 단어들을 선택하여 마스킹한 다음, 마스킹 된 위치에 적절한 단어를 모델이 예측할 수 있도록 학습한다. 다음 문장 예측 모델은 질문/응답을 위한 2개의 문장으로, 두 문장이 서로 이어지는 문장인

지 아닌지를 분류하는 모델을 학습하게 된다. 이처럼 2가지 방식으로 pre-trained(사전학습) 된 BERT 모델을 자연어 관련 다양한 태스크에 적용하기 위해 각각의 태스크에 따른 데이터들을 추가로 학습하여 모델을 개선하게 된다. 이 과정을 fine-tuning(파인튜닝)이라고 하는데, BERT 저자들은 논문에서 크게 4가지 학습 방법을 소개하였다(Fig. 4. 참조).

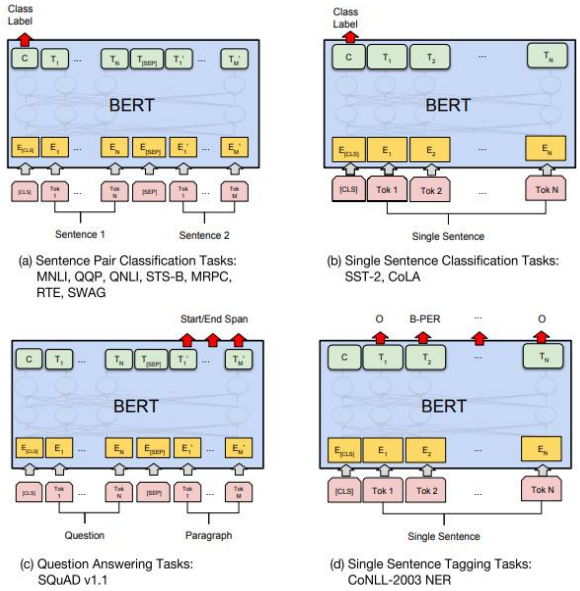


Fig. 4. BERT Model Classification

본 논문에서는, 입력되는 법률 상담 질문에 대해 가장 유사한 대법원 판결문을 검색하기 위해, 다음 문장 예측 모델(NSP)로 pre-trained 된 BERT 모델에 fine-tuning (파인튜닝)을 수행하여 유사 문서 검색 모델(STS, Semantic textual Similarity)을 할 수 있는 모델을 개발한다. BERT 기반 법률 문서 검색 방식은 일반인의 질의 문장과 법률 문서를 BERT 모델로 같은 벡터 공간에 벡터로 표현하게 된다. 그런 다음, 코사인 유사도를 이용하여 질의 문장 벡터와 법률 문서 벡터 간 유사도를 측정하여 가장 높은 유사도에 해당하는 법률 문서 벡터를 리턴하는 방식으로 동작한다. 하지만 BERT를 이용한 두 문장 사이의 유사성을 계산하기 위해, 두 문장 모두가 네트워크에 입력되어야 하므로 막대한 계산이 발생하는 문제점이 있다. 예를 들어, 10,000개의 문장 집합에서 가장 유사한 문장 쌍을 찾기 위해 BERT는 약 5천만 개의((n*n-1)/2) 문장간 유사도를 계산하는 시간이 필요하다[15]. 본 논문에서 수집한 법률 판결문 8만여 건과, 사용자 질의문 300여 개 사이에서는 약 2,400여만 개의 문장간 유사도를 구하는 시간이 필요하므로, 계산하는 과정에서 오버헤드가 발생하게 되어 많은 시간이 소요되게 된다.

3-3 S-BERT

S-BERT는 BERT 모델에서 문장 임베딩의 성능을 개선한 모델로서 BERT의 문장 임베딩을 응용하여 BERT를 파인 튜닝하는 모델링 기법이다[16]. BERT는 10,000개의 문장 집합에서 가장 유사한 문장 쌍을 찾기 위해 약 65시간이 소요되지만, BERT 네트워크를 수정한 S-BERT는 가장 유사한 문장들을 검색하는 시간을 약 5초로 줄이는 획기적인 성능을 가져왔다. S-BERT는 NLI(Natural Language Inferencing) 문장 쌍 분류 태스크 문제를 다루는 데 주로 사용된다. NLI는 두 개의 문장을 입력받아, 두 문장의 관계가 수반, 중립, 모순 중 어디에 속하는지를 예측하는 문제이다. S-BERT는 BERT/RoBERTa를 사용하여 가장 유사한 문장을 찾는 시간을 수 초 내에 가능하게 하였으며, 다른 임베딩 방법보다 성능이 매우 뛰어난 것으로 확인되었다[16].

S-BERT는 Sentence A와 Sentence B를 각각의 BERT에 입력하고, 풀링을 수행한 다음 각각에 대한 문장 임베딩 벡터 u와 v를 구한다(Fig. 5. 참조). u와 v 두 벡터 간의 코사인 유사도를 구한 다음 레이블 유사도와의 평균 제곱 오차(Mean Squared Error)를 최소화하는 방식으로 학습하게 된다.

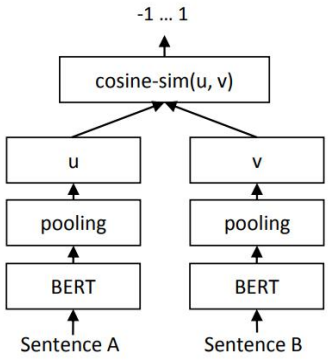


Fig. 5. SBERT architecture

III. S-BERT based legal search

1. Legal Search System Process

본 논문에서 제안하고자 하는 법률 검색시스템의 프로세스는 다음과 같다(Fig. 6. 참조).

첫째, 대한법률구조공단 사이트에서 법률 상담 데이터를 스크래핑한 다음 문서화한다. 또한, 대법원 종합법률정

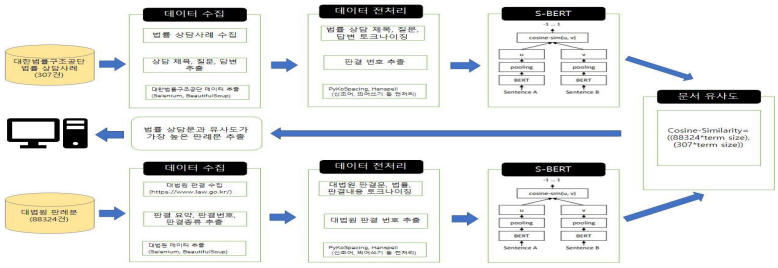


Fig. 6. Legal Search System Process (S-BERT)

보 사이트의 판례 데이터를 스크래핑하여 문서로 만든다.

둘째, PyKoSpacing 패키지 등을 활용하여 띄어쓰기를 체크 및 교정하고, Hanspell 패키지를 활용하여 오타를 정정한다.

셋째, 데이터 전처리를 수행한 법령 데이터 셋에 대해 판결내용, 판결번호, 법률 상담 질문, 제목, 답변 등을 추출하고 토크나이징을 수행한다.

넷째, 토크나이징을 수행한 법률 상담문과 대법원 판결문에 대해 TF-IDF 행렬(DTM) 생성 및 BERT, S-BERT 모델로 각각 임베딩한다.

다섯째, 사용자 법률 상담 질문에 대한 S-BERT 임베딩 벡터 공간에 표현된 법률 상담 질문 벡터와 대법원 판결문 벡터에 대해 코사인 유사도를 구하고 유사도가 가장 높은 대법원 판결문을 추출한다.

마지막으로, 판결번호 및 판례 정보를 사용자에게 제공한다.

2. Legal Data Set and Development Environment

본 논문에서 활용하고자 하는 데이터셋은 두 그룹으로 나누어 구축되었다. 첫 번째로는 법률 지식이 없는 일반인의 법률 질의문이며, 두 번째는 법률적으로 전문적인 지식을 갖추고 있는 법률인의 판결문이다. 데이터 수집 단계에서는 대한법률구조공단에서 제공하는 법률 상담사례를 수집하였고, 다양한 카테고리 중에 노동 카테고리에 해당하는 상담사례 전체를 수집하였다.

본 논문에서는 한국어로 작성된 법률 질의 문장과의 의미 유사성을 정확히 판단하여 법률 질의에 가장 유사한 대법원 판결문들을 검색하는 데 있다. 이를 위해, 실험에서는 하나의 법률 상담사례에 대해 유사도가 가장 높은 10개~50개의 대법원판결 사례를 각각 검색 및 추출하였다. 이렇게 추출된 대법원판결 검색 결과의 판결 번호와 법률 상담사례 데이터의 각 질의문에 관한 법률 전문가의 답변글에 포함된 참조 대법원 판결번호가 서로 비교하여 같은지를 판단하여 분류 결과의 정확도를 평가하였다.

실험은 AWS 클라우드 및 구글 코랩(Colab)환경에서 진행되었으며, 하드웨어 및 스크랩 툴은 아래와 같다.

· HW - NVIDIA T4 Tensor 코어 급 GPU(단일 GPU VM) - 1vGPU / 4vCPU / 16GB RAM / 16GB GPU 급 RAM / 1000GB SSD, Ubuntu 18.04 OS / NVIDIA CUDA, cuDNN / Anaconda3 / python 3.8 - R 3.6.3

· Tool - KoNLPy / gensim / Okt / Hanspell / PyKoSpacing / Selenium / Beautifulsoup

3. Implementation Procedures

3-1. Legal counseling case collection

판례검색시스템 구현에 필요한 법률 상담사례 데이터를 수집하기 위해 웹 스크래핑을 수행하였고, 상담자의 질문 제목과 질문 내용, 공단 측의 답변을 수집하였다(Fig. 7. 참조).

Fig. 7. Legal Consultation Case of the Korea Legal Aid Corporation

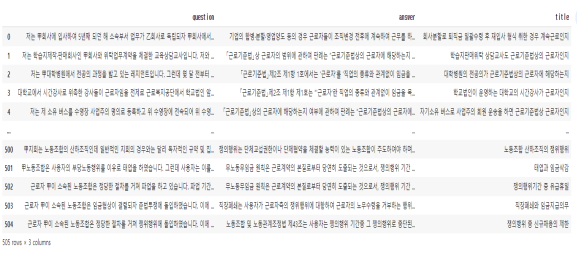


Fig. 8. Legal Consultation Case Document (Title, Content, Answer)

법률구조공단에서 법률상담사례 데이터 중에 노동과 관련된 상담사례 307건을 크롤링하였고, 대법원 판례 데이터에서는 약 88,324건을 크롤링하였다(Fig. 8-11. 참조).

```
        판례일련번호 = node.find('판례일련번호').text
        사건명 = node.find('사건명').text
        사건번호 = node.find('사건번호').text
        선고일자 = node.find('선고일자').text
        법원명 = node.find('법원명').text
        사건종류명 = node.find('사건종류명').text
        사건종류코드 = node.find('사건종류코드').text
        판결유형 = node.find('판결유형').text
        선고 = node.find('선고').text
        판례상세링크 = node.find('판례상세링크').text

        rows.append({'판례일련번호' : 판례일련번호,
                     '사건명' : 사건명,
                     '사건번호' : 사건번호,
                     '선고일자' : 선고일자,
                     '법원명' : 법원명,
                     '사건종류명' : 사건종류명,
                     '사건종류코드' : 사건종류코드,
                     '판결유형' : 판결유형,
                     '선고' : 선고,
                     '판례상세링크' : 판례상세링크})
    cnt += 1
    url = "https://www.law.go.kr/DRF/lawSearch.do?OC=egohimdrura&target=prec&type=XML&page={}".format(page)
    response = urlopen(url).read()
    xtree = ET.fromstring(response)
```

```
cases = pd.DataFrame(rows)
cases.to_csv('../cases.csv', index=False)
```

Fig. 9. Example of Web Scraping of Supreme Court Case Data

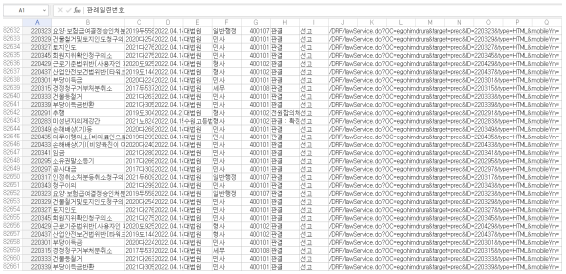
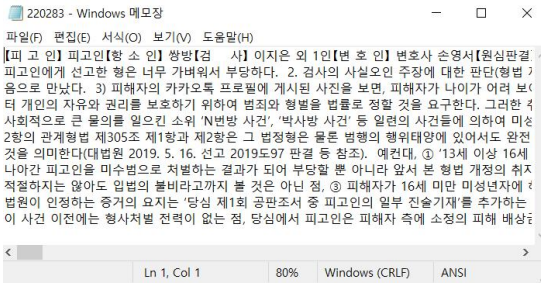


Fig. 10. Supreme Court case data scraping results



220283 - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

【피 고 인】 피고인【항 소 인】 쌍방【검 사】 이지은 외 1인【변 호 인】 변호사 손영서【원심판결

피고인에게 선고한 형은 너무 가벼워서 부당하다. 2. 검사의 사실오인 주장에 대한 판단(형법

음으로 만났다. 3) 피해자의 카카오톡 프로필에 게시된 사진을 보면, 피해자가 나이가 어려 보

터 개인의 자유와 권리를 보호하기 위하여 범죄와 형벌을 법률로 정할 것을 요구한다. 그러한

사회적으로 큰 물의를 일으킨 소위 'N번방 사건', '박사방 사건' 등 일련의 사건들에 의하여 미

2항의 관계형법 제305조 제1항과 제2항은 그 법정형은 물론 범행의 행위태양에 있어서도 완전

것을 의미한다(대법원 2019. 5. 16. 선고 2019도97 판결 등 참조). 예컨대, ① '13세 이상 16세

나아간 피고인을 미수범으로 처벌하는 결과가 되어 부당할 뿐 아니라 앞서 본 형벌 개정의 취지

적절하지는 않아도 입법의 불비라고까지 볼 것은 아닌 점, ③ 피해자가 16세 미만 미성년자에

법원이 인정하는 증거의 요지는 '당심 제1회 공판조서 중 피고인의 일부 진술기재'를 추가하는

이 사건 이전에는 형사처벌 전력이 없는 점, 당심에서 피고인은 피해자 측에 소정의 피해 배상

Ln 1, Col 1 | 80% | Windows (CRLF) | ANSI

Fig. 11. Examples of Supreme Court precedents

3-2. Data Preprocessing

법령 용어를 잘 모르는 일반인의 질문이 입력되었을 때, 질문에 대해 먼저 pykospacing, pyspell 패키지를 사용하여 띄어쓰기 및 맞춤법 등 교정을 수행하였고, 동의어 추출을 위해 konlpy 패키지 내의 Okt 형태소 분석기를 이용하여 형태소 단위로 분할하였다(Fig. 12., 13. 참조)..

```
['제가 골프장에서 캐디로 근무하고 있는데요 근로자라고 할 수 있나요?',
 '알바로커피숍에서일하는데요.휴무일에수당받을수있나요?',
 '제가 공장에서 일하는 노동자인데요. 회사가 휴업중입니다. 임금은 어떻게 될까요?',
 '알바로커피숍에서일하는데요. 휴무일에 수당 받을 수 있나요?',
 '제가공장에서일하는노동자인데요. 회사가 휴업중입니다. 임금은 어떻게 될까요?',
 '제가 골프장에서 캐디로 군무하고 있는데요 근로자라고 할 수 있나요?',
 '알바로 코피샵에서 일하는데요.휴무일에 수당받을 수 있나요?']
```

Fig. 12. Before performing text preprocessing

```
['제', '가', '골프장', '에서', '캐디', '로', '근무', '하고', '있는데요', '근로자', '라고', '할', '수', '있나요', '?']
['알바로', '커피숍', '에서', '일', '하는데요', ',', '휴무', '일', '에', '수당', '받을', '수', '있나요', '?']
['제', '가', '공장', '에서', '일', '하는', '노동자', '인데', '요', ',', '회사', '가', '휴업', '중', '입니다', '.', '임금', '은', '어떻게', '될까', '요', '?']
['알바로', '커피숍', '에서', '일', '하는데요', ',', '휴무', '일', '에', '수당', '받을', '수', '있나요', '?']
['제', '가', '공장', '에서', '일', '하는', '노동자', '인데', '요', ',', '회사', '가', '휴업', '중', '입니다', '.', '임금', '은', '어떻게', '될까', '요', '?']
['제', '가', '골프장', '에서', '캐디', '로', '근무', '하고', '있는데요', '근로자', '라고', '할', '수', '있나요', '?']
['알바로', '커피숍', '에서', '일', '하는데요', ',', '휴무', '일', '에', '수당', '받을', '수', '있나요', '?']
```

Fig. 13. After performing text preprocessing

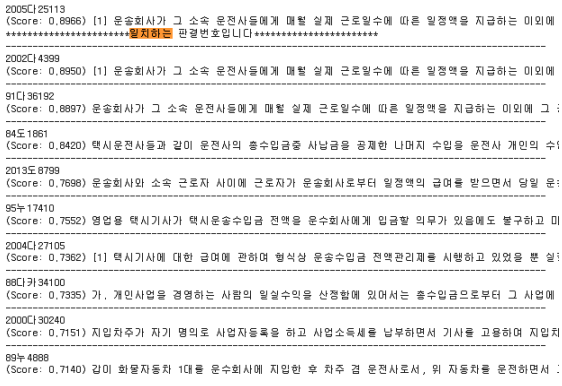
Fig. 12에서 텍스트 전처리 수행 이전의 문장에서는 띄어쓰기가 전혀 안 되어 있거나, 맞춤법이 맞지 않는 단어가 다수 있지만, Fig. 13에서 텍스트 전처리 수행 이후의 문장에서는 교정 및 정정된 것을 확인할 수 있다.

3-3. S-BERT embedding vector and similarity

본 논문에서는 S-BERT 임베딩을 수행하기 위해 한국어 사전학습(pre-trained) 모델 ko-sroberta-multitask를 KorNLU 데이터셋으로 fine-tuning한 모델을 이용하였다[17]. 먼저, 법률구조공단의 법률 상담 사례 데이터와 웹 스크래핑으로 수집한 대법원 판결문에 대해 각각 임베딩을 수용하였다. 이 과정에서, 법률 질의문 및 대법원 판결문 전체 문장과 문단을 각각 768 차원의 dense vector로 임베딩 수행하였으며, 임베딩 된 사용자 질문 벡터와 대법원 판결문 벡터 간 코사인 유사도를 계산하여 유사도가 가장 높은 판결문을 추출하였다.

예컨대, 법률 상담인의 질의문으로 '운송회사의 운전사들이 운송수입금 중 사납금을 공제한 잔액을 운전사 개인의 수입으로 하여 온 경우, 그 사납금 초과 수입금이 임금에 해당하는지 여부'에 대한 법률가의 답변(판결번호)이 ['대법원 2007. 7. 12. 선고 2005다25113 판결']이라고 했을 때, S-BERT 임베딩 벡터 간 코사인 유사도 결과는 다음과 같이 출력되었다(Fig. 14. 참조).

두 벡터 간 코사인 유사도(score)는 0.8966으로 가장 높은 대법원 판결문은 '운송회사가 그 소속 운전사들에게 매월 실제 근로일수에 따른 일정액을 지급하는 ...중략... 근로기준법 제34조에 위반되어 무효이다.' 이었으며, 높은 코사인 유사도 값과 함께 정답과 일치한 것을 확인할 수 있었다.



2005다25113
(Score: 0.8966) [1] 운송회사가 그 소속 운전사들에게 매월 실제 근로일수에 따른 일정액을 지급하는 이외에
**********일치하는 판결번호입니다**********

2002다4399
(Score: 0.8950) [1] 운송회사가 그 소속 운전사들에게 매월 실제 근로일수에 따른 일정액을 지급하는 이외에

91다36192
(Score: 0.8897) 운송회사가 그 소속 운전사들에게 매월 실제 근로일수에 따른 일정액을 지급하는 이외에 그

84도1861
(Score: 0.8420) 택시운전사들과 같이 운전사의 총수입금중 사납금을 공제한 나머지 수입을 운전사 개인의 수

2013도8799
(Score: 0.7698) 운송회사와 소속 근로자 사이에 근로자가 운송회사로부터 일정액의 급여를 받으면서 당일 운

95누17410
(Score: 0.7552) 영업을 택시기사가 택시운송수입금 전액을 운수회사에게 입금할 의무가 있음에도 불구하고 미

2004다27105
(Score: 0.7362) [1] 택시기사에 대한 급여에 관하여 형식상 운송수입금 전액관리제를 시행하고 있었을 뿐 실

88다카34100
(Score: 0.7335) 가. 개인사업을 경영하는 사람의 일실수익을 산정함에 있어서는 총수입금으로부터 그 사업에

2000다30240
(Score: 0.7151) 지입차주가 자기 명의로 사업자등록을 하고 사업소득세를 납부하면서 기사를 고용하여 지입차

89누4888
(Score: 0.7140) 갑이 화물자동차 1대를 운수회사에 지입한 후 차주 겸 운전사로서, 위 자동차를 운전하면서

Fig. 14. S-BERT embedding for statutory data sets and queries

또 다른 예로, 질의자가 '제가 골프장에서 캐디로 근무하고 있는데요 근로자라고 할 수 있나요?' 라는 질문을 했을 때, '골프장 캐디가 근로기준법상 근로자에 해당하는지 여부'라는 판결문이 검색되었으며, 코사인 유사도 값은 Score: 0.8198이 나왔다. '근로자'를 '노동자', '근로'를 '일'이라는 유의어로 변경하여 질의했을 때도 동일한 법률 상담 질문이 검색되었으며, 코사인 유사도 값은 Score: 0.8183으로서 매우 큰 값이 나온 것을 확인할 수 있다.

이렇게 검색된 결과에서 유사도가 가장 높은 법률 상담 제목과 질문, 답변을 추출한 다음 참조 판결번호를 스크래핑하였다. 참조 판결번호는 대법원 판례 데이터에서 검색하여 판례에 대한 '판시사항', '판결요지' 등 판결과 관련한 다양한 정보를 사용자에게 제공해줄 수 있다.

Fig. 15. Supreme Court case search results for queries

IV. Experiment result

Table 1을 보면 각각의 임베딩 모델이 어떤 방식으로 문장 간의 유사도를 판단하는지 추측할 수 있다. 먼저, 키워드 기반으로 유사도를 계산하는 TF-IDF 모델 방식은 특정 단어에 대해 가중치를 높게 줘서, 유사한 문장을 판단하는 것으로 보인다. Doc2Vec 모델은 단어 및 문서에 대해 모두 임베딩을 수행함으로써, 문맥에 대한 정보를 함께 반영할 수 있었고, 정확도가 높게 나온 것을 볼 수 있다. Doc2Vec 모델링시 각 판결문을 한 개의 id 값을 갖는 단어처럼 취급한 다음 학습을 수행하게 된다. 그에 따라 판결문서 id는 일반적인 단어들처럼 임베딩 벡터 공간에 하나의 위치 좌표값을 갖게 되는 방식으로 학습이 진행된다. 유사도 결과는 다소 낮지만, TF-IDF보다는 우수한 정확도를 확인할 수 있었다.

S-BERT 모델로 임베딩 된 법률 상담 질의문과 대법원 판결문 벡터 간의 코사인 유사도를 이용한 검색의 경우에는 다른 알고리즘에 비해 높은 정확도를 도출할 수 있다. 평균적으로 S-BERT를 이용한 검색의 정확도는 TF-IDF에 비해 6.62%, Doc2Vec에 비해 2.62% 향상되었다. 이것은 질의문과 판결문에 대한 문맥 정보가 잘 반영된 결과라고 할 수 있다.

Table 1. Accuracy and cosine similarity by model

model	n	accuracy	Average Similarity
TF-IDF	10	52.9	0.6193
	20	55.6	0.5825
	30	58.3	0.5268
	40	60.2	0.4874
	50	62.1	0.4492
Doc2Vec	10	56.2	0.7352
	20	59.4	0.6897
	30	61.3	0.6556
	40	64.2	0.6478
	50	68.0	0.6187
S-BERT	10	58.1	0.8053
	20	61.9	0.7482
	30	63.9	0.7154
	40	67.1	0.6924
	50	71.2	0.6735

Table 1에서 n은 각각의 모델에 의해 법률 질의문과 가장 유사도가 높은 대법원 판결문 검색 개수가 된다. 예를 들어 n=10은 법률 질의문이 입력되었을 때, 가장 유사도가 높은 대법원 판결문 10개가 추출된 것이라고 볼 수 있다. accuracy는 가장 유사도가 높은 대법원 판결문 각각에 대한 판결번호와 법률 질의문에 관한 법률 전문가가 작성한 답변에 포함된 참조 대법원판결 번호가 서로 일치하는지를 확인하여 계산한 정확도이다. Average Similarity는 각 모델 별 검색 결과에 대한 코사인 유사도의 평균값이다. 상대적으로 S-BERT의 코사인 유사도 평균값이 높게 형성되었으며, n이 증가함에 따라 코사인 유사도 평균값이 낮아진 걸 확인할 수 있다. 이것은 검색 개수가 늘어남에 따라 코사인 유사도 값이 낮아지므로, 유사도 평균값 낮아지는 것이라고 볼 수 있다(Fig 16 참조).

본 논문에서 실험한 다양한 모델별 검색은 오답으로 처리되었지만, 정답과 유사한 대법원 판결이 여러 개 있을거라 예상된다. 이것은 검색 결과의 정확도가 더 낮아지게 하는 요인이 된다고 판단된다. 이와 같은 상황에서 S-BERT를 이용한 법률 검색 결과에 대한 정확도가 다른 알고리즘에 비해 높다는 점을 확인할 수 있다.

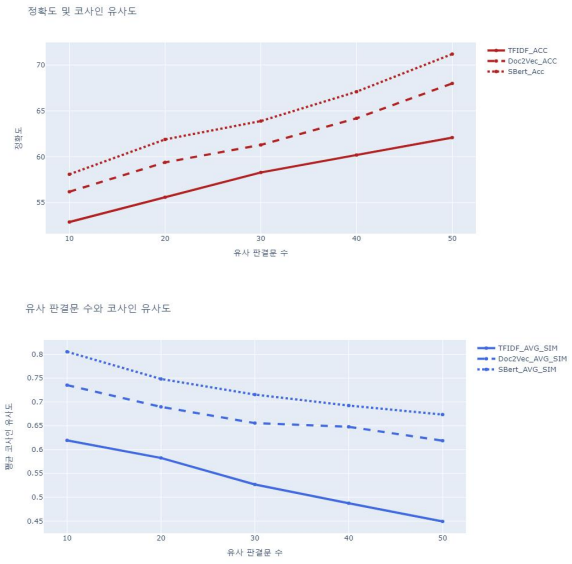


Fig. 16. Supreme Court case search results for queries

V. Conclusions

본 논문에서는 법률 용어를 모르는 일반인들이 사용하는 일상 용어로도 정확하고 풍부한 판례 및 판결 정보 등을 쉽게 분류하고 받을 수 있도록 S-BERT를 사용하여 일반인의 질의에 대해 가장 유사한 대법원 판결문을 검색하는 방법을 제안하였다.

법률 분야와 같이 전문적인 분야에서도, 전처리 과정을 거쳐 학습에 사용할 수 있는 양질의 데이터셋을 구축하고, S-BERT 모델을 이용하여 실제 법률 서비스에 활용할 수 있을 만큼의 충분한 성능을 이끌어 낼 수 있었다.

향후에는 현재 연구에서 검증되지 않은 일반인의 다양한 법률 질의문 사례에 대해서도 유의미한 결과를 도출하기 위한 연구를 할 예정이다. 온톨로지를 S-BERT 모델과 결합하여 온톨로지 기반으로 단어들 사이에 정의된 관계 및 규칙을 추론한 결과를 S-BERT 모델에 반영하여, 문장 사이의 관계를 좀 더 명확하게 함으로서 판결문의 검색 정확도를 향상시키기 위한 연구를 계속 진행할 것이다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (2021R1A2C2008414)

REFERENCES

[1] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, Maosong Sun, "How does NLP benefit legal system:A summary of legal artificial intelligence", arXiv:2004.12158, pp.5218-5230, Apr. 2020. DOI:10.18653/v1/2020.acl-main.466

[2] Kang myeon gu, [Court Column]'Advice for Solitary Litigation', http://www.kwnews.co.kr/page/view/2022031500000000016

[3] T Mikolov, K Chen, G Corrado, J Dean, "Efficient estimation of word representations in vector space", arXiv:1301.3781, Jan. 2013. DOI:10.3126/jiee.v3i1.34327

[4] Chang, W. C., Yu, F. X., Chang, Y. W., Yang, Y., & Kumar, S., "Pre-training tasks for embedding-based large-scale retrieval.", In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, Apr. 2020. DOI:10.48550/arXiv.2002.03932

[5] Jeffrey Pennington, Richard Socher, Christopher D. Mannin, "GloVe: Global Vectors for Word Representation", In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, Doha, Qatar, Oct. 2014. DOI:10.3115/v1/d14-1162

[4] Lee, K., Chang, M. W., & Toutanova, K., "Latent retrieval for weakly supervised open domain question answering.", arXiv preprint arXiv:1906.00300., pp. 6086-6096, Florence, Italy, Jul. 2019. DOI:10.18653/v1/p19-1612

[5] Reimers, N., & Gurevych, I., "Sentence-bert: Sentence embeddings using siamese bert-networks.", arXiv preprint arXiv:1908.10084., pp. 3982-3992, Hong Kong, China, Nov. 2019. DOI:doi.org/10.18653/v1/d19-1410

[6] Edwina L. Rissland et al., "AI and Law: A fruitful synergy", Artificial Intelligence, Vol. 150, No. 1, pp. 1-15, Nov. 2003. DOI:https://doi.org/10.1016/s0004-3702(03)00122-x

[7] Bruce G. Buchanan and Thomas E. Headrick. "Some Speculation About Artificial Intelligence and Legal Reasoning", Stanford Law Review, Vol. 23, No. 1, pp.40-62, Nov. 1970. DOI:https://doi.org/10.2307/1227753

[8] L. Thorne McCarty, "Reflections on 'TAXMAN': An Experiment in Artificial Intelligence and Legal Reasoning", Harvard Law Review, Vol. 90., No. 5, pp. 837-893, Mar. 1977. DOI: https://doi.org/10.2307/1340132

[9] J, S, Sim, H, J, Kim, "A Searching Method for Legal Case Using LDA Topic Modeling", Journal of The Institute of Electronics and Information Engineers, Vol. 54, NO. 9, pp. 67-75, Sep. 2017., DOI:10.5573/ieie.2017.54.9.67

[10] Ji Hyun Kim,.Jong-Seo Lee,.Myungjin Lee,Wooju KimJune, Seok Hong, "Term Mapping Methodology between Everyday Words and Legal Terms for Law Information Search System", Journal of Intelligence and Information Systems Vol 18, No. 3, pp.137-152., Sep. 2012. DOI: 10.13088/jiis.2012.18.3.137

[11] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent Trends in Deep Learning Based Natural Language Processing,” arXiv preprint arXiv:1708.02709, Aug. 2018. DOI: 10.1109/mci.2018.2840738

[12] Y. Bengio, R. Ducharme, P. Vincent et al., “A neural probabilistic language model,” Journal of Machine Learning Research, Vol. 3, pp. 1137-1155, Feb. 2003. DOI: 10.1162/153244303322533223

[13] H. Y. Lee, and J. S. Lee, “Functional Expansion of Morphological Analyzer Based on Longest Phrase Matching For Efficient Korean Parsing“, Journal of Digital Contents Society, Vol. 17, No. 3, pp. 203-210, Jun. 2016. DOI: 10.9728/dcs.2016.17.3.203

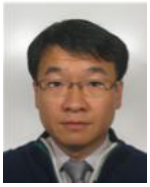
[14] Quoc Le and Tomas Mikolov. ”Distributed representations of sentences and documents.“, In International Conference on Machine Learning, pp. 1188-1196. May. 2014. DOI: 10.48550/arXiv.1405.4053

[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. “Bert:Pre-training of deep bidirectional transformers for language understanding.”, NIPS, pp. 4171-4186, May. 2019., DOI:10.48550/arXiv.1810.04805

[16] Nils Reimers and Iryna Gurevych, “Sentence-BERT Sentence Embeddings using Siamese BERT-Networks”, arXiv:1908.10084v1[cs.CL] 27, pp. 3982-3992 ,Nov. ,2019. DOI:10.48550/arXiv.1908.10084

[17] Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, Hyungjoon Soh, "Kornli and korsts: New benchmark datasets for korean natural language understanding", Findings of the Association for Computational Linguistics: EMNLP 2020, pp.422-430, Nov. 2020, DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.39

Authors



Gil-sik Park received the M.S. degree, Ph.D. candidate in Computer Engineering from Chung-ang, Dongguk University Korea, in 2008, 2013, respectively. He is interested in Data Mining, NLP, Machine Learning, and AI.

Jun-tae Kim received the B.S. degree in Control and Measurement Engineering from Seoul National University, Korea, in 1986. He received the M.S., Ph.D. degrees in Computer Engineering from University of Southern California, USA, in 1990, 1993, respectively. He is interested in Data Mining, NLP, Machine Learning, and AI.