ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 49, No. 10, pp. 848-858, 2022. 10
https://doi.org/10.5626/JOK.2022.49.10.848

키워드와 문장 임베딩을 활용한 조항별 분류모델 기반 계약서 적격성 검증

(Contract Eligibility Verification Enhanced by Keyword and Contextual Embeddings)

이 상 아 † 김 석 기 †† 김 은 진 †††
(Sangah Lee) (Seokgi Kim) (Eunjin Kim)

강 민 지 ††† 신 효 필 ††††
(Minji Kang) (Hyopil Shin)

요 약 최근에는 계약서를 포함한 법률 문서들을 대량으로, 빠르고 정확하게 처리하기 위하여 인공지능을 활용한 자동화된 분석 방법이 요구된다. 계약서는 그 안에 필수적인 조항들이 모두 포함되었는지, 어느 한 쪽에 불리한 조항은 없는지 등을 확인하여 적격성을 검증할 수 있다. 이때 계약서를 이루는 조항들은 계약서의 종류와 관계없이 매우 정형적이고 반복적인 경우가 많다. 본 연구에서는 이러한 성격을 이용하여 계약서 내 조항별 분류 모델을 구축하였으며, 계약서의 관습적인 요구사항에 기반하여 구성한 키워드 임베딩을 구축하고 이를 BERT 임베딩과 결합하여 사용한다. 이때 BERT 모델은 한국어 사전학습모델을 법률 도메인 문서를 이용하여 미세 조정한 것이다. 각 조항의 분류 결과는 정확도 90.57과 90.64, F1 점수 93.27과 93.26으로 우수한 수준이며, 이렇게 계약서를 이루는 각 조항이 어떤 필수조항에 해당되는지의 예측 결과를 통해 계약서의 적격성을 검증할 수 있다.

키워드: 법률 문서, 계약서, 계약서 적격성, BERT, 키워드 임베딩

Abstract Contracts need to be reviewed to be verified if they include all the essential clauses for them to be valid. Such clauses are highly formal and repetitive regardless of the kinds of contracts, and automated legal technologies are required for legal text comprehension. In this paper, we have constructed a simple item-by-item classification model for clauses in contracts to estimate contract eligibility by addressing formal and repetitive properties of contract clauses. We have used keyword embeddings based on conventional requirements of contracts and concatenate them to sentence embeddings of clauses, extracted from a BERT model fine-tuned with legal documents. The contract eligibility can be verified by the predicted labels. Based on our methods, we report reasonable performances with the accuracy of 90.57 and 90.64, and an F1-score of 93.27 and 93.26, using additional keyword embeddings with BERT embeddings.

Keywords: legal text, contract, contract eligibility, BERT, keyword embeddings

· 본 연구는 2019년 중소기업 기술개발사업인 창업성장-기술개발사업의 지원을 받은 'AI를 활용한 웹 기반의 Legal RPA 솔루션 개발 연구'(주관기관 법률)의 위탁연구로 수행되었다.

† 비 회 원 : 서울대학교 언어학과 교수
sanalee@snu.ac.kr
†† 비 회 원 : 서울대학교 데이터사이언스대학원 석사졸업
blaqdraq77@snu.ac.kr
††† 비 회 원 : 서울대학교 언어학과 학생
jyej3154@snu.ac.kr
mnjkng@snu.ac.kr
†††† 정 회 원 : 서울대학교 언어학과 교수(Seoul Nat'l Univ.)
hpshin@snu.ac.kr
(Corresponding author임)

논문접수 : 2022년 5월 27일
(Received 27 May 2022)
논문수정 : 2022년 7월 18일
(Revised 18 July 2022)
심사완료 : 2020년 7월 26일
(Accepted 26 July 2022)

정보과학회논문지 제49권 제10호(2022. 10)

1. 서 론

대규모 데이터에 대한 고도화된 처리 기술이 활발하게 연구되면서, 대규모의 계약서 및 법률 문서의 분석에 인공지능을 활용하는 리걸 테크 역시 조명되고 있다. 계약서의 경우 각각의 조항들이 양측 계약자 중 어느 한쪽에 치우치지 않도록 공정하게 작성되었는지를 판단해야 한다. 계약서에 필요한 항목이 모두 포함되어 있는지, 어느 한쪽에 불리한 조항 또는 필요 없는 내용이 들어 있는지를 확인하는 것인데, 이 작업은 보통 변호사들이 직접 계약서를 확인하며 진행한다. 그러나 계약서를 이루는 조항들은 그 특성상 여러 계약서에서 공통적으로 사용되고, 그 형태가 매우 고정적이며 반복적으로 나타나는 경향이 있다. 예를 들어 계약서에는 일반적으로 용어 정의, 임금, 배상금, 비밀 유지 등의 항목들이 포함되어 있다. 이들 항목은 계약서가 디자이너, 번역가, 웹 개발자 등 다양한 직종과 관련하여 작성되어도 대부분 공통적으로 포함된다.

따라서 변호사들이 이렇게 유사한 항목들을 포함한 계약서들을 일일이 읽고 확인하는 것은 반복적이고 지루한 작업이 될 수 있다. 이는 최근에 법률 문서에 대한 이해 및 처리를 자동화하는 연구인 리걸 테크(legal tech)가 조명되는 이유의 하나일 수 있다. 최근의 연구들은 계약서의 적격성 검증, 계약서 자동 작성, 계약서 템플릿 생성, 문서의 수정 내역 추적, 법률 문서 검색 또는 질의응답 시스템 등 다양한 방향으로 진행된다[1)·2)]. 한국어 법률 문서 역시 조항, 판례[3)], 특허[4)] 등의 검색 엔진, 법률 문서 요약 시스템[5)], 질의응답 시스템 등 다양한 연구와 서비스를 포함한 접근이 이루어지고 있다.

본 연구에서는 계약서의 적격성을 검증하기 위해 각 계약서에 필수 조항들이 포함되어 있는지를 판단하는 간단한 모델 임베딩 기반 모델을 구축하였다. 이는 계약서 내 조항들을 개별적으로 분류하여 대규모의 데이터에 대해서도 부담 없이 동작하도록 한 가벼운 모델이다. 길고 복잡하게 보이는 계약서도 각 조항의 형태가 어느 정도 정해져 있고 반복적인 경우가 많으므로, 이러한 계약서 조항들의 성격을 이용하여 간편하게 동작하는 도메인 한정적인 모델을 구축하였다.

본 연구에서 대상으로 한 데이터는 한국어로 작성된 용역 계약서들로 구성되어 있다. 이 데이터를 참고하여 법률 전문가들이 20가지의 필수적인 조항들의 목록을 사전 정의하였고, 이는 본 연구의 조항 분류 모델의 정답 레이블(label)로 사용되었다. 각 조항의 분류 결과에 기반하여 주어진 계약서가 20가지 사전 정의된 필수 조항들 중 일부만을 포함한다면, 본 연구에서 최종 구축한 시스템이 누락된 조항들의 목록을 출력하고, 사용자는 이에 기반하여 해당 계약서의 적격성을 검증해 볼 수 있을 것이다.

모델에서 분류할 조항들은 형태소 분석[6)]된 뒤 법률 문서를 이용해 미세 조정(fine-tuning)한 BERT 모델[1]을 통해 임베딩으로 변환된다. 이에 각 필수 조항과 관련하여 얻은 유의미한 키워드(보증서, 요율, 지체상금, 특허 등)의 사용 분포에 따른 20차원의 임베딩을 결합하고, 곧 사전 정의된 20가지의 필수 조항 중 하나 또는 그 이상의 레이블을 갖도록 분류된다. 이에 따라 본 연구에서는 정확도(accuracy) 90.57과 90.64, 그리고 F1 점수 93.27과 93.26을 얻을 수 있었다.

2. 기존 연구

2.1 법률 분야에서의 인공지능 활용

인공지능을 활용하여 계약서를 검토하는 과정을 자동화하면, 인력은 점검 과정에서 2차적으로 투입되기 때문에 비용이 상당히 감소되는 효과가 있다. 그러나 대량의 양질의 데이터의 부족, 알고리즘의 한계와 불투명성 때문에[2], 이러한 잠재력에도 불구하고 법률 문서에 자연어 처리를 적용한 연구는 부족한 실정이고, 영어가 아닌 언어에서는 특히 더 그렇다.

Soh 외[3]의 판결문 분류 비교 연구에서는 전통적인 토픽 모델링, 단어 임베딩 모델, 특히 파인튜닝한 사전 학습 모델이 부족한 데이터에서도 충분한 성능을 보여주었지만, Manor와 Li[4]의 계약서 자동 요약 실험에서는 사전 학습된 언어와 법률 문서에서 사용되는 언어의 사용역 차이로 인해 성능에 한계가 있다는 것이 드러났다.

Simonson 외[5]는 계약서를 비롯한 법률 문서에 쓰이는 언어가 어휘와 통사구조 면에서 상당히 반복적이라는 것을 보여주었고, 이는 언어 모델을 개발할 때 통사구조의 중요성이 약화된다는 것을 시사한다. 법률 문서에 쓰이는 언어가 가독성이나 효율성보다는 어휘적, 통사적으로 통일성을 주어 의미를 명확하게 하는 것에 초점을 두고 있기 때문이다. 법률 언어에서 관용적으로 사용되는 고정 언어 표현이 이러한 언어 특성의 예시가 된다[6]. 이러한 특징은 범언어적으로 나타나는 현상으로, 법률 분야에 특화된 사전 학습 모델의 필요성이 대두되었고[7-9], 프랑스어와 루마니아어로 작성된 법률 문서를 학습한 BERT 모델이 등장하였다[10,11].

1) https://www.checkthiscontract.com/
2) https://donotpay.com/
3) https://legalengine.co.kr/
4) https://legaltech.co.kr/
5) https://aihub.or.kr/
6) https://bitbucket.org/eunjeon/mecab-ko

한국어 법률 문서는 언어 모델 개발 이전에 언어 사용역 연구조차 부족한 실정이다[12]. 양윤영 외[13]는 한국어 법률 문서에 높은 빈도로 등장하는 합성명사의 벡터를 조정하여 언어 모델의 성능을 향상시킨 바가 있다. 이는 모델의 강건성을 위한 비지도학습 모델의 필요성을 시사한다.

2.2 법률 문서에서의 분류 태스크

법률 문서에서의 분류 태스크 연구는 크게 규칙기반 방법과 머신러닝 알고리즘을 이용한 방법을 기반으로 한 연구로 나뉜다. Curtotti와 McCreath[14]는 the Australian Contract Corpus에서 256개의 계약서를 모아 행 단위로 처리하고 각 행을 32개의 class로 레이블을 붙여주었다. 32개의 class는 해당 행이 구조적으로 중요한 요소이거나 메타데이터임을 나타낸다. 위 연구에서는 나이브 베이즈(Naïve Bayes)나 SVM과 같은 모든 머신러닝 알고리즘에 직접 정한 규칙(hand-coded rules)을 함께 사용하였을 때 성능이 개선되었음을 확인하였다. 또한 규칙이 완벽하지 않더라도 머신러닝 알고리즘과 함께 쓰였을 때 효과적임을 보였다.

Waltl 외[15]는 독일법에서 법의 기본 요소를 나타내는 법 규범을 의미에 기반하여 9개의 카테고리로 나누었다. 실험 결과 UIMA framework을 사용한 규칙 기반 모델의 F1 score는 78이었고, 머신러닝 기반 모델 중 성능이 가장 좋았던 모델은 linear kernel과 결합한 SVM으로 F1 score는 83이었다. 또한 Padhy 외[16]는 건설 분야의 계약서에서 필수조항을 찾는 연구를 진행하였다. 계약서에서 공통된 용어들을 찾아 Python 라이브러리인 Spacy를 이용하여 규칙 기반 모델을 개발하였고 실험 결과 모델의 정확도는 80.43%였다.

대규모 사전 학습 모델은 법률 문서 분류에도 사용되었다. Soh 외[3]의 장문 판결문 분류 연구에서는 파인튜닝을 거치지 않은 사전 학습 모델이 다른 분야보다도 법률 분야에서 특히 좋은 성능을 보이지 않는다는 것을 보여주었다. 또한 Tziafas 외[17]는 코로나바이러스감염증-19 대응방식을 분류하기 위한 다국어 문장단위 법률 문서 코퍼스를 개발하였고, 파인튜닝을 거친 XLM-ROBERTa 모델이 F1 score가 59.8로 가장 높았다. 이 연구는 법률 분야는 그 특수성 때문에 법률 문서를 통한 파인튜닝이 필수적이라는 것을 보여주었다.

인공지능을 법률 문서 분류에 도입하는 것의 단점인 불투명성[18]을 개선하기 위한 노력도 있었다. Chalkidis 외[6]는 EU 입법 문서 데이터셋 EURLEX57K를 개발하여 극단적인 다중 레이블 텍스트 분류(Extreme Multi-Label Text Classification)를 진행하였고, BIGRU 모델이 가장 좋은 성능을 보여줌과 동시에 이에 대한 설명으로 어텐션 히트맵(attention heat-map)을 사용하였다. Ferro 외[19]는 법률 문서 코퍼스에서 결함이 있는 조항에 이슈(Issue)와 요소(Factor)를 도입하여 주석을 닮으로써 어떤 결함이 발생하였는지 밝히려는 시도를 했다. 예측 모델의 초기 임베딩 결과를 이슈와 요소로서 보여주는 방식이었다.

한편 한국어 계약서 분류 태스크에 대한 연구는 적은 편에 속한다. 우선 전명준[20]에서는 sentence-wise attention BERT가 계약서 문서에서 조항을 분류하는데 좋은 성능을 보이는 것으로 나타났다. 위 연구에서는 근로계약서와 사채인수계약서를 모아 문서를 문장으로 분리하였고 각 문장은 법률 전문가에 의해 필수 조항인지 아닌지 태깅되었다. 그리고 KorBERT(ETRI)와 LSTM을 결합하였는데, LSTM에서 sentence-wise attention layer를 활용하여 context vector를 생성하였다. 다중 레이블(multi-label) 분류인 근로계약서 분류에서는 F1 score가 88.18이었고, 다중 클래스(multi-class) 분류인 사채인수계약서 분류에서는 97.86이었다. 하지만 계약서의 도메인이 제한적이고 모델이 도메인에 따라서 따로 미세 조정되었다는 점에서 한계가 있다.

이치훈 외[21]에서는 기계독해 모델을 사용하여 계약서에서 위험요소를 감지하는 연구를 수행하였다. 우선 법률 문서들과 법률용어 사전에서 모은 정보를 기반으로 계약서의 위험 요소를 분석하였다. 이를 통해 계약서에서 누락될 시 피해를 야기하는 필수 조항을 규정하였고, 질의응답(Question-Answering) 데이터셋을 구축하기 위해 변호사들이 조항이 필수 조항인지 아닌지 정하기 위한 체크리스트를 만들었다. 총 1315개의 데이터셋을 구축하였고 ENLIPLE Large-v1 모델을 적대적 학습으로 파인튜닝하였다. 실험 결과 가우시안 노이즈를 적용한 모델이 가장 성능이 높았으며 F1 score는 87.93이었다. 하지만 이치훈 외 4인(2021)[21]에서 사용한 데이터셋의 양이 매우 적기 때문에 사전학습된 모델에 파인튜닝 시 모델에 과적합(over-fitting) 문제를 야기할 수 있다는 문제가 있다.

따라서 본 연구의 주요한 목적은 간단한 구조를 가진 계약서 조항 단위 분류 모델을 개발하는 것으로, 과적합 가능성을 줄이기 위해 충분한 양의 데이터셋을 구축한다. 또한 기존 한국어 분류 모델을 개선하기 위해서 다양한 산업 분야에서 사용된 계약서를 새로운 입력단위로 가공하여, 계약서에 사용된 다양한 문장 통사구조에 강건한 모델을 제시하고, 문장의 키워드 기반 태깅을 통해 분류 기준에 관한 설명도 제공한다.

2.3 키워드를 활용한 문서 분류

문서에서 자주 등장하는 단어를 기반으로 문서를 분류하는 기법은 비교적 단순한 머신러닝 기법으로도 유효한 성능을 보여왔다[22-25]. 특히 의생명과학 분야[26]와 블

로그 글 분류[27] 등 도메인이 한정되어 있을 때 상당히 효과적이었다. 그러나 문서의 도메인을 분류할 때 IDF(역문서빈도)로 빈출 단어를 사용하는 경우[28]에는 거의 성능 향상을 보이지 않았다. 이는 기존 연구들에서 지적하였듯[22, 29], 키워드를 추출할 때 맥락이 반영된 임베딩이 중요하다는 사실을 보여준다. 이러한 점을 개선하기 위해 TextCNN[24]에서는 각 키워드에 분류 레이블별로 등장 확률을 임베딩하는 multi-weighting 기법을 적용하여 성능 향상을 보였다. 그러나 BERT 등 사전학습된 트랜스포머 모델과 도메인에 적합한 키워드를 결합하여 연구한 사례는 찾아볼 수 없다.

3. 데이터

실험에 사용한 데이터셋은 총 2355개의 계약서로, 기업용 법무 관리 솔루션을 제공하는 한국의 기업 법틀[7]로부터 제공받았다. 계약서에서 이름, 회사 정보, 금액의 양 등과 같은 개인정보나 회사의 정보를 담은 데이터는 익명화하였다. 데이터셋은 광고, 디지털산업, 공동구매 등 다양한 산업의 계약서로 구성되어 있다. 법률 전문가가 크게 필수조항과 옵션조항으로 나누어 총 20개의 조항 구분을 정의하였고 각 조항에 직접 레이블을 태깅하였다. 라벨링이 된 조항은 계약서에서 꼭 포함되어야 하는 조항이고, 그 중 필수조항은 계약서의 적격성을 판단하는 데 더 필수적인 조항임을 의미한다. 또한 하나의 조항이 여러 레이블(multi-label)을 가지는 경우도 있다. 조항 구분에 대한 정의는 표 1에 제시되어 있다.

표 1 법률 전문가들이 정의한 조항 구분의 내용
Table 1 Content of Classes Defined by Legal Experts

Class	Content
Required1	Contract period
Required2	Contract deadline
Required3	Contract amount
Required4	Compensation for delay
Required5	Intellectual property right
Required6	Non-disclosure
Required7	Warranty against defects/Detect repair
Required8	Advance/Intermediate payment/Balance
Required9	Inspection for the result/product
Required10	Penalty for the principal/service provider
Required11	Contract termination/cancellation/ automatic renewal
Required12	Compensation for damage
Required13	Transfer of ownership
Optional1	Performance Guarantee
Optional2	Definition of term
Optional3	Changes to contract
Optional5	Direct payment request
Optional6	Rejection of products
Optional7	Subcontracting
Optional8	Jurisdiction
None	None of the above

7) https://buptle.com/

원본 계약서에서 각 행은 장, 조, 항, 호, 목으로 구분되어 있고 일련의 입력 전처리 과정을 통해 항 단위의 입력 데이터를 새로 구축하였다. 입력 전처리 과정을 거친 후 조항 구분의 분포는 표 2에서 확인할 수 있다. 실험을 위해 우선 500개의 계약서를 먼저 평가 데이터셋으로 분리해 두었고, 나머지 계약서를 항 단위로 학습 데이터와 검증 데이터셋으로 나누었다. 이후 남은 조항들을 표 3에서와 같이 각 조항 구분의 비율을 동등하게 학습 데이터셋과 검증 데이터셋으로 나누기 위해 scikit-learn의 iterative-stratification 패키지를 사용하였다.

표 2 전체 조항의 개수 분포
Table 2 The Distribution of Total Lines

Train	Eval	Test	Total
53,787	17,929	19,938	91,654

4. 방법론

4.1. 데이터 입력 전처리

본 연구에서는 계약서를 모델에 적합한 형태로 넣기 위해 새로운 입력 단위를 설정하여 실험용 계약서 데이터를 구축하였다. 원본 계약서는 장, 조, 항, 호, 목 등의 순서로 계층을 이루는 조항들로 구성된다. 법틀에서 제공한 데이터는 이들 조항을 문장 경계에 따라 나누고, 각 문장에 대해 표 1에 기술된 조항 클래스에 따른 레이블을 주석한 것이다. 표 4는 이러한 원본 계약서의 일부를 나타낸다.

첫 번째로 모든 계약서에서 장 단위는 삭제하였다. 대부분 장 단위는 내용이 비어있거나 조 단위와 내용이 같기 때문에 장 단위는 본고에서 수행하는 분류 태스크에 유의미한 내용을 담고 있지 않다고 판단하였다. 두 번째로 항 단위 이하의 호, 목 등의 세부 단위들은 [SEP] 토큰으로 구분하여 항과 결합하였다. 이 과정에서 기존의 레이블들도 결합되며 하나의 항이 여러 개의 레이블을 갖게 될 수도 있다.

또한 원래부터 하나의 항이 여러 레이블을 가지는 경우에는 동일한 항을 여러 번 반복하여 레이블을 주지 않고, 하나의 항에 여러 개의 레이블을 붙여주도록 처리하였다. 항을 처리한 이후 상응하는 조를 [SEP] 토큰으

표 3 데이터셋에서 각 조항 구분의 개수. 학습, 검증, 평가 데이터셋은 6:2:2로 나뉘었다.

Table 3 Ratio of Each Class in the Dataset. Train, Eval, Test datasets were split to be close to 6:2:2

Class	Train	Eval	Test
Required1	1,427	476	522
Required2	292	97	100
Required3	1,260	420	470
Required4	1,076	359	408
Required5	2,269	756	836
Required6	3,917	1,306	1,358
Required7	1,357	452	477
Required8	1,156	385	451
Required9	1,469	490	495
Required10	1,258	419	438
Required11	3,849	1,283	1,414
Required12	3,334	1,111	1,193
Required13	464	154	149
Optional1	729	243	316
Optional2	704	235	251
Optional3	2,243	748	860
Optional5	111	37	40
Optional6	86	29	29
Optional7	171	57	66
Optional8	1,112	371	387
None	27,570	9,175	10,389
Total	55,854	18,603	20,649

로 결합하였다. 이를 통해 각 입력 문장이 출처 계약서와 내용적으로 일관성을 유지하도록 하였다. 따라서 모델이 하나의 입력 문장마다 충분한 정보를 가질 수 있게 하였고 모델에 넣기 위해 레이블을 원핫 벡터(one-hot vector)로 변환 시 덜 희소한 벡터가 되도록 하였다. 데이터 입력 전처리 과정은 그림 1에서 확인할 수 있다. 표 5는 이러한 처리를 통해 구성한 모델 입력 형태의 예시이다.

표 4 원본 계약서 예시

Table 4 Examples of Raw clauses

Clause	Content	Class
Article10	지적재산권	
Article10 Paragraph1-1	본 계약상 “을”이 제작하여 “갑”에게 제공한 일체의 산출물은 제3자의 지식재산권을 침해하지 않아야 한다.	Required5
Article10 Paragraph1-2	만약 이로 인하여 “갑”에게 손해가 발생한 경우 “을”은 이를 배상하여야 한다.	Required5
Article10 Paragraph1-2	만약 이로 인하여 “갑”에게 손해가 발생한 경우 “을”은 이를 배상하여야 한다.	Required12
Article10 Paragraph2	“을”이 “갑”에게 제공한 일체의 산출물에 대한 권리(소유권 및 지식재산권 등)는 “갑”의 수수료 지급과 동시에 “갑”에게 귀속한다.	Required5

표 5 10조 1항 조항구분의 원핫 벡터는 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0]이다.

Table 5 Examples of Processed clauses. One-hot label vector for Article10 Paragraph1 is [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0]

Clause	Content	Class
Article10 Paragraph1	지적재산권 [SEP] 본 계약상 “을”이 제작하여 “갑”에게 제공한 일체의 산출물은 제3자의 지식재산권을 침해하지 않아야 한다. [SEP] 만약 이로 인하여 “갑”에게 손해가 발생한 경우 “을”은 이를 배상하여야 한다.	Required5, Required12
Article10 Paragraph2	지적재산권 [SEP] “을”이 “갑”에게 제공한 일체의 산출물에 대한 권리(소유권 및 지식재산권 등)는 “갑”의 수수료 지급과 동시에 “갑”에게 귀속한다.	Required5

4.2. 모델

주어진 입력 문장이 해당되는 조항 클래스를 분류하기 위해서 모델을 제작하고자 하였다. 본 연구는 먼저 데이터 도메인의 특성을 반영하여 특정한 키워드의 목

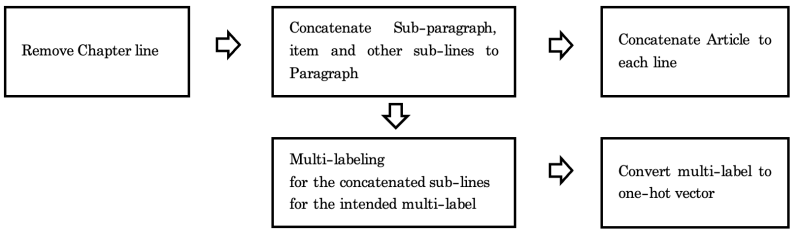

그림 1 입력 데이터 처리 과정

Fig. 1 Data Input Processing Procedure

록을 정의하여 사용하였다. 법률 도메인의 데이터에 포함되는 문장들은 정형적이고 반복적인 표현을 사용하는 경우가 많으며, 특히 본 연구에서는 데이터의 범위가 용역 계약서로 한정되었으므로 이러한 특성이 더욱 분명하게 나타나리라고 기대할 수 있다.

데이터 탐색을 통해 각 조항별로 자주 나타나는 단어들이 존재하는 것을 확인하였다. 예를 들어 계약 금액과 관련된 조항들에는 '수당', '운임', '기술료'와 같은 단어가, 지적재산권에 관한 조항들에는 '라이센스', '로얄', '창작권' 등의 단어가 특징적으로 사용된다. 데이터 특성에 기반해 수동 추출한 이들 단어를 키워드로 하여 본 연구의 모델이 각 조항을 분류하는 데에 추가적인 정보가 되도록 하였다. 따라서 모든 조항 클래스 중 특정 조항에서만 자주 등장하는 단어를 기반으로 키워드 목록을 작성하였다. 키워드 목록 중에는 하나의 조항에서만 자주 사용되는 키워드도 있고, 3~4개 이상의 조항에 포함되는 키워드도 존재한다. 조항을 분류하는 것이 목적이기 때문에 대부분의 조항 클래스에 자주 등장하는 단어는 키워드 목록에 포함하지 않았다. 작성한 키워드 목록을 이용하여 각 키워드에 대한 벡터를 구성하였다. 각 조항 클래스에 존재하는 각 키워드의 비율을 계산하여 조항의 수만큼의 크기를 가지는 벡터로 만들었다. 예를 들어 '보증'이라는 키워드에 대한 벡터는 [0.1891, 0, 0.0055, .., 0]과 같이 표현된다. 여기에서 '0.1891'은 이 키워드가 첫 번째 조항 클래스로 분류된 문장의 18.91%에 존재한다는 것을 의미한다. 또한 모델은 주어진 입력 문장에 대해 하나의 클래스로만 분류하는 것이 아니라 각 클래스마다 분류가 될지의 여부를 결정하는 다중 레이블 분류(multi-label classification)이기 때문에 비율은 전체 데이터 안에서의 비율이 아닌 각 클래스 내에서의 비율로 계산하였다. 이러한 방법을 통해 모델은 주어진 입력 문장에 키워드가 존재할 때 각 조항으로 분류될 확률을 반영할 수 있다. 이렇게 생성된 키워드 벡터는 상수 벡터로서, 추후 구성되는 조항 임베딩과 결합되어 활용된다.

그림 2는 모델 구조를 통해 각 조항에 대응하는 문장 임베딩과 키워드 벡터가 구성되는 과정을 보여 준다. 그림의 왼쪽 부분에서 볼 수 있듯, 모델에 조항 문장이 입력되면 토큰별로 KR-BERT-MEDIUM을 통과하여 나온 마지막 은닉 상태에서 [CLS] 토큰과 [SEP] 토큰에 해당하는 부분을 max pooling하여 문장 임베딩 벡터를 얻는다. 그림의 오른쪽은 입력 문장에 존재하는 모든 키워드들의 벡터를 mean pooling하여 최종 키워드 벡터를 얻는 과정을 보인다. 이때 그림에서 *keyword_n*은 미리 정의된 키워드 목록 중 해당 문장에 포함된 키워드의 개수를 나타내는 표현이다. 이렇게 얻어진 두 벡터를 서로 연결하여 linear layer에 통과될 최종 벡터를 구성한다.

이렇게 키워드 벡터와 문장 임베딩을 연결한 최종 벡터를 입력받은 선형 분류기를 통해 조항 분류를 수행하였다. 이때 조항 클래스 판단 기준에 threshold를 적용하여 다중 레이블 분류를 수행할 수 있도록 하였다. threshold로는 0.0부터 0.9까지 0.1 단위로 바꾸어 가며 수행한 실험 중 가장 결과가 높았을 때의 값을 채택하였다. 모델이 예측한 조항 클래스별 logit이 threshold 값 이상인 경우 해당 클래스에 대한 예측 여부가 참이 된다.

4.3 계약서 적격성 검증

모델 학습의 최종적인 목적은 계약의 적격성을 결정

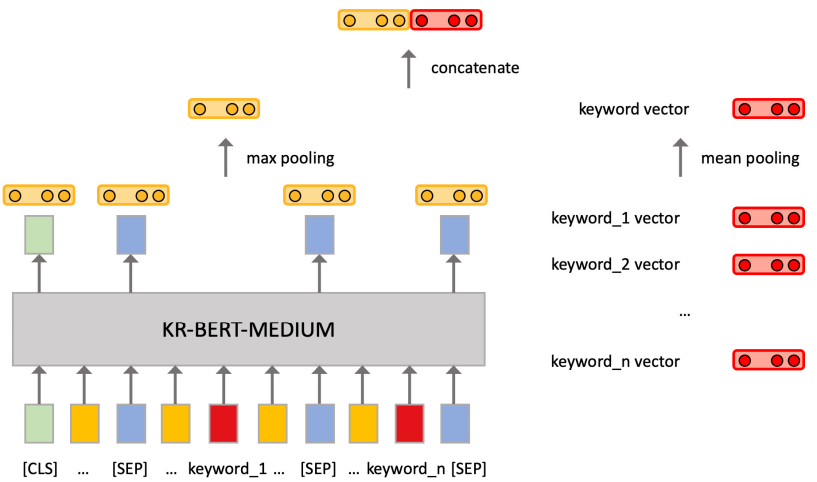

그림 2 임베딩 벡터 구성 방법

Fig. 2 The Structure to Obtain the Embedding Vector

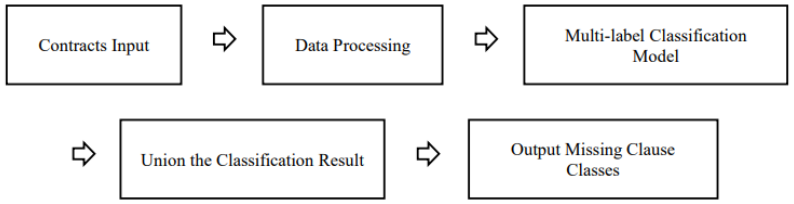

그림 3 계약서 적격성 검증 과정
Fig. 3 Contract eligibility verification process

하는 것이다. 학습된 모델을 사용하여 계약서에 반드시 포함되어야 하는 필수조항이 포함되어 있는지의 여부를 판단할 수 있다. 그림 3은 계약서 적격성 결정 과정을 보여준다. 먼저 입력으로 계약서를 넣으면 데이터 처리를 통해 학습된 모델을 통과할 수 있는 형태의 문장으로 변환한다. 변환된 각 문장은 다중 레이블 분류 모델을 거쳐 어떤 조항으로 분류되는지 확인된다. 모든 문장에 대한 분류가 끝나면 계약서에 있는 모든 문장의 분류 결과를 집계하여 계약서에 존재하는 모든 조항 클래스를 확인한다. 마지막으로 얻은 결과를 이용하여 계약서에 반드시 포함되어야 하는 필수 조항 중 계약서에 존재하지 않는 것을 출력한다. 이때 경우에 따라 필수조항 또는 특정 옵션 조항들을 모두 포함할 경우 적합한 계약서라는 기준을 마련한다면, 위와 같은 과정을 통해 학습된 모델을 활용하여 계약서의 적격성을 검증할 수 있다.

5. 실 험

5.1 학습

조항 분류 모델을 학습하기 위해 세 가지 형태로 실험을 수행했다. 실험은 다음과 같은 형태로 진행하였다.

- **KR-BERT-MEDIUM:** 한국어로 사전 학습된 BERT 모델 중 하나인 KR-BERT-MEDIUM을 사용하여 각 문장에 해당하는 조항을 분류하는 모델을 학습시켰다.
- **KR-BERT-MEDIUM(CLS) + 키워드 벡터:** 4.2에서 설명하였듯이 키워드 벡터와 BERT를 통과하여 얻은 마지막 은닉 상태를 연결한 임베딩 벡터를 linear layer에 통과시킴으로써 조항 분류하도록 모델을 학습시켰다. 차이점은 그림 2의 [SEP] 토큰에 해당하는 최종 은닉 상태를 사용하여 max pooling하는 것과 다르게 [CLS] 토큰에 해당하는 최종 은닉 상태만 사용하고 키워드 벡터와 연결한다는 점이다.
- **KR-BERT-MEDIUM(CLS+SEP) + 키워드 벡터:** 4.2에서 설명된 실험 방법으로, 입력 문장마다 [SEP] 토큰의 개수가 다르기 때문에 [CLS] 토큰과 [SEP] 토큰에 해당하는 최종 은닉 상태를 max pooling하여 벡터를 얻었다.

3가지 형태의 실험 모두에서, 조항 분류 모델을 학습시키기 위해 linear layer를 transformer layer 뒤에 붙여서 사용하였다. 다중 레이블 분류(multi-label classification) 문제에서 자주 사용되는 binary cross entropy loss를 손실 함수로 사용하였다. 활성화 함수로는 sigmoid 함수를 사용하였다.

다음과 같은 하이퍼파라미터들을 사용하여 KR-BERT-MEDIUM 모델을 미세 조정(fine-tuning)하였다.

- 학습 에폭: 5
- 배치 크기: 12
- 옵티마이저: AdamW
- 학습률: 2e-5
- 최대 토큰 길이: 256

5.2 성능 평가

세 가지 형태로 조항 분류 모델을 학습하고 테스트 데이터셋을 통해 성능을 평가하였다. 표 6은 5.1에서 언급한 세 가지 모델의 실험 결과와, 비교를 위해 수행한 베이스라인 실험 결과, 그리고 ablation test를 포함한다.

베이스라인 실험을 위해서는 BERT 임베딩 대신 각 조항을 구성하는 단어의 인덱스로만 구성된 기본 임베딩을 사용하고, 여기에 키워드 벡터를 더하는 경우를 추가로 상정하여 레이어 2개로 이루어진 LSTM 모델을 적용하였다. 또한 본 연구의 각 상세 방법론의 성능을 평가하기 위해 LSTM과 BERT 임베딩 모델, 그리고

표 6 조항 분류 태스크 성능
Table 6 Classification Performance of Our Models

Model	Accuracy	F1-score
keyword vector only	9.77	12.94
LSTM	83.12	81.75
LSTM + keyword vector	84.38	82.21
KR-BERT-MEDIUM	88.76	86.93
KR-BERT-MEDIUM (CLS) + keyword vector	90.57	**93.27**
KR-BERT-MEDIUM (CLS+SEP) + keyword vector	**90.64**	93.26

키워드 벡터를 각각 단독 인풋으로 입력한 ablation test를 수행하였다. 이때 모든 실험에서 하이퍼파라미터는 위 5.1에 기술한 것과 같으나, 키워드 벡터만을 단독으로 입력한 경우 학습 에폭은 20, 학습률은 5e-5로 설정하였다.

성능 평가 결과 한국어로 사전 학습된 BERT 모델만 사용한 경우 정확도가 약 88.76%이고 F1-점수가 약 86.93인 것으로 나타났다. 반면 확률을 이용하여 구성한 키워드 벡터를 추가적으로 사용한 경우 정확도가 90% 수준으로 증가하고 F1-점수도 약 93으로 증가하였다. 이는 LSTM 모델로 구성한 베이스라인 실험 결과에 비해 매우 높은 수준이며, 베이스라인 실험 안에서도 키워드 벡터를 사용한 경우가 그렇지 않은 경우보다 우수한 성능을 보인다. 따라서 본 연구에서 구성한 법률 키워드 벡터가 조항 분류에서 중요한 역할을 할 수 있다는 것을 알 수 있다. CLS 토큰과 CLS+SEP 토큰을 사용하는 모델을 비교한 경우 성능에는 큰 차이가 없었다.

결과에서 알 수 있듯이 사전 학습된 BERT 모델에 키워드 임베딩 벡터를 추가하는 간단한 방법을 통해 조항 분류 성능을 개선시킬 수 있었다. 키워드 임베딩 벡터는 각 법률 키워드가 조항 클래스와 연관될 확률을 반영하기 때문에 조항을 분류하는 데 도움을 주게 된다. 이는 조항 클래스 별로 자주 등장하는 키워드가 존재하기 때문에 가능하다. 다만 키워드 벡터만을 조항 클래스 분류기의 인풋으로 사용한 경우에는 매우 저조한 성능을 보여, 도메인 내 특징적인 키워드의 사용이 문장의 의미를 반영한 임베딩과 결합되었을 때 비로소 효과를 나타냄을 알 수 있다.

표 7은 위 모델 중 가장 높은 F1-점수를 기록한 모델인 KR-BERT-MEDIUM(CLS) + 키워드 벡터가 실제 데이터 내 조항을 분류한 결과의 일부를 나타낸 것이다.

또한 표 8은 KR-BERT-MEDIUM(CLS) + 키워드 벡터의 실험 결과를 조항 클래스별 Precision, Recall,

표 8 조항 클래스별 분류 성능
Table 8 Classification Performance per Class

Class	Precision	Recall	F1-score
Required1	97.87	96.93	97.40
Required2	87.38	90.00	88.67
Required3	91.71	91.91	91.82
Required4	95.50	93.63	94.55
Required5	90.83	88.88	89.84
Required6	94.80	96.61	95.70
Required7	91.95	90.99	91.46
Required8	89.21	89.80	89.50
Required9	92.46	94.14	93.29
Required10	79.54	71.00	75.03
Required11	97.41	95.62	96.50
Required12	92.07	95.39	93.70
Required13	72.67	78.52	75.48
Optional1	98.36	94.94	96.62
Optional2	97.21	97.21	97.21
Optional3	95.69	95.47	95.58
Optional5	95.12	97.50	96.30
Optional6	80.65	86.21	83.33
Optional7	90.77	89.40	90.08
Optional8	99.48	99.74	99.61
Average	93.38	93.20	93.27

표 7 KR-BERT-MEDIUM(CLS) + 키워드 벡터의 조항 분류 예시
Table 7 Examples of Classification Performed by the KR-BERT-MEDIUM(CLS) + Keyword Vector Model

Content	Gold Label	Model Prediction
물가변동으로 인한 계약금액의 조정 [SEP] 계약상대자는 제3항의 규정에 의하여 계약금액의 증액을 청구하는 경우에는 계약 금액 조정내역서를 첨부하여야 한다. <신설 '99.11.4>	Optional3	Optional3
개인정보의 활용제한 [SEP] 본 조는 계약기간의 만료 및 계약의 해제·해지가 된 후에도 효력이 있다. 이 조항의 의무를 위반하는 경우 "수행사"는 "고객사"에게 발생한 모든 손해에 대하여 배상할 의무가 있으며, 손해배상과 별도로 이 계약의 즉시 해지사유가 된다.	Required6, Required11, Required12	Required6, Required11, Required12
"프로젝트 결과물"의 검수 [SEP] "을"은 제4조의 기한 내에 "프로젝트 결과물" 및 기타 협의 자료 등을 완성하여 "갑"에게 제출하고 "갑"의 검수를 받아야 한다.	Required9	Required9
소유권, 권리, 의무 [SEP] "을"은 종합기업서비스정보망 구축사업목적에 맞는 정보를 제공하며, "갑"이 정보서비스상에서 정보에 대한 오류가 발생하였을 경우에는 "을"은 신속하게 정정한다.	Required5	Required6
불가항력의 사유 및 업무대가 반환 [SEP] 전항의 불가항력에 해당되는 것으로 판단된 경우 "갑"은 본 계약을 해제/해지할 수 있으며, "을"은 기 지급받은 대가 전부를 "갑"에게 반환하여야 한다	Required10, Required11	Required11
계약기간 [SEP] 본 계약기간은 OOOO년 OO월 OO일부터 OOOO년 OO월 OO일까지로 하며, [SEP] 계약기간 만료 1개월 전까지 상호 합의를 통하여 계약기간을 연장할 수 있으며, [SEP] 별도의 합의가 없는 한 계약기간 만료로 본 계약은 자동 종료된다.	Required1, Required11, Optional3	Required1, Required11, Optional3

F1-점수 값으로 나타낸 것이다. 이 값들은 대부분의 조항 클래스에서 90 이상의 높은 수치를 보이나, 필수조항10, 필수조항13, 옵션조항6에서와 같이 낮은 값을 나타내는 경우도 있다. 옵션조항6처럼 해당 조항 클래스를 포함하는 문장의 개수가 매우 적거나, 필수조항10과 옵션조항6과 같이 해당 조항 클래스와 관련된 키워드가 정의되지 않은 경우 성능에 영향을 미쳤을 수 있다. 또한 필수조항5와 필수조항13과 같이 서로 밀접한 관련이 있으나 별개의 클래스로 구분된 조항의 경우에도 보다 세밀한 수준의 분류까지는 수행되지 못했을 가능성이 있다. 이러한 분류 결과들로부터 유추할 수 있는 약점을 보강하여 본 연구가 제안하는 모델의 성능을 향상시킬 수 있을 것이다.

이렇게 문장 임베딩과 키워드 벡터를 활용한 모델을 사용하여 조항을 분류함으로써 입력으로 들어오는 계약문서의 모든 문장에 대한 분류 결과를 합하면 누락된 조항 클래스를 확인할 수 있다. 이를 통해 누락된 조항 중 '필수' 조항이 포함된 경우 해당 계약이 부적합하다는 결정을 내릴 수 있다. 예를 들어 한 계약서의 각 조항에 대한 분류 결과들을 소거한 뒤 남은 조항 클래스가 '필수조항2, 필수조항4, 필수조항7, 필수조항13, 옵션조항1, 옵션조항2, 옵션조항5, 옵션조항6, 옵션조항7'인 경우, 4종류의 필수조항이 계약서에 포함되어 있지 않으므로 해당 계약서는 부적합하다고 판정하는 것이다.

6. 결 론

본 연구에서는 한국어 용역 계약서 데이터셋을 대상으로 하여 계약서 내 조항 단위의 분류 모델을 제안하고, 이를 이용해 계약서의 적격성을 검증할 수 있도록 하였다. 계약서는 조항 단위로 처리되는데, 각 조항은 전처리를 거쳐 한국어 BERT 임베딩과 사전 정의된 키워드의 임베딩을 결합한 것으로 변환된다. 이때 키워드는 정형적이고 반복적인 계약서 조항들의 특성을 이용하여, 또한 법률 전문가들의 20가지 필수조항 정의를 이용하여 구성된 것이다. 본 연구는 이러한 모델을 통하여 조항 분류 성능으로 유의미한 정확도(accuracy)와 F1 점수를 확인하였다. 각 조항에 대한 이들 분류 결과를 이용하여, 실제 계약서에서 특정 조항 분류들이 모두 존재하는지의 여부를 통해 계약서의 적격성을 검증할 수 있을 것이다.

이러한 방법론을 추후에 다양한 도메인의 계약서 및 법률 문서에도 확장 적용해 보고자 한다. 본 연구에서 사용한 자료인 용역 계약서 외에도 거래 계약서, 매매 계약서, 임대 계약서 등 여러 종류의 문서들도 유사하게 분석할 수 있을 것이다. 또한 본 연구의 한국어 계약서 데이터를 활용하여 문장쌍 인코딩 모델을 이용한 유사 조항 검색 시스템 등, 리걸 테크와 관련된 더욱 다양한 작업들을 추가로 시도해 볼 계획이다.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171-4186, 2019.

[2] M. E. Kauffman and M. N. Soares, "Ai in legal services: New trends in ai-enabled legal services," *Service Oriented Computing and Applications*, Vol. 14, No. 4, pp. 223-226, 2020.

[3] J. Soh, H. K. Lim, and I. E. Chai, "Legal area classification: A comparative study of text classifiers on singapore supreme court judgments," *Proceedings of the Natural Legal Language Processing Workshop 2019*, pp. 67-77, 2019.

[4] L. Manor and J. J. Li, "Plain English Summarization of Contracts," *Processing of the Natural Legal Language Processing Workshop 2019*, pp. 1-11, 2019.

[5] D. Simonson, D. Broderick, and J. Herr, "The extent of repetition in contract language," *Proceedings of the Natural Legal Language Processing Workshop 2019*, pp. 21-30, 2019.

[6] S. Jeong, "Zur analyse von mehr oder weniger festen wortverbindungen in patentschriften im deutschen und koreanischen," *Zeitschrift für deutschsprachige Kultur Literatur*, No. 23, pp. 359-381, 2014.

[7] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Extreme multi-label legal text classification: A case study in EU legislation," *Proceedings of the Natural Legal Language Processing Workshop 2019*, pp. 78-87, 2019.

[8] J. Savelka and K. Ashley, "Discovering explanatory sentences in legal case decisions using pre-trained language models," *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4273-4283, 2021.

[9] N. Limsopatham, "Effectively leveraging BERT for legal document classification," *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 210-216, 2021.

[10] S. Douka, H. Abdine, M. Vazirgiannis, R. E. Hamdani, and D. R. Amariles, "JuriBERT: A masked-language model adaptation for French legal text," *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 95-101, 2021.

[11] M. Masala, R. C. A. Iacob, A. S Uban, M. Cidota, H. Velicu, T. Rebedea, and M. Popescu, "jurBERT: Romanian BERT model for legal judgement prediction," *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 86-94, 2021.

[12] H. R. Kim, "A corpus-driven analysis of key words in korean translation of chinese statutes," *Interpreting and Translation Studies*, Vol. 25, No. 1, pp. 21-49, 2021.
[13] Y. Yang, S. Shin, J. Y. Jang, and H. Jung, "Regulations corpus-based question and answering system," *Proceedings of Korea Computer Congress 2018*, pp. 696-698, 2018.
[14] M. Curtotti and E. McCreath, "Corpus based classification of text in Australian contracts," *Proceedings of the Australasian Language Technology Association Workshop 2010*, pp. 18-26, 2010.
[15] B. Waltl, G. Bonczek, E. Scepankova, and F. Matthes, "Semantic types of legal norms in german laws: classification and analysis using local linear explanations," *Artificial Intelligence and Law*, Vol. 27, No. 1, pp. 1-29, 2019.
[16] J. Padhy, M. Jagannathan, and V. Delhi, "Application of natural language processing to automatically identify exculpatory clauses in construction contracts," *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, Vol. 13, No. 4, 2021.
[17] G. Tziafas, E. Saint-Phalle, W. de Vries, C. Egger, and T. Caselli, "A multilingual approach to identify and classify exceptional measures against COVID-19," *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 46-62, 2021.
[18] S. Jain and B. C. Wallace, "Attention is not Explanation," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 3543-3556, 2019.
[19] L. Ferro, J. Aberdeen, K. Branting, C. Pfeifer, A. Yeh, and A. Chakraborty, "Scalable methods for annotating legal-decision corpora," *Proceedings of the Natural Legal Language Processing Workshop 2019*, pp. 12-20, 2019.
[20] M. J. Jeon, "Research on the classification of contract articles using bert-based sentence-wise attention," Master's thesis, Yonsei University, 2020.
[21] C. H. Lee, J. W. Noh, J. H. Jeong, K. S. Joo, and D. H. Lee, "Risk prediction model of legal contract based on korean machine reading comprehension," *Journal of Information Technology Services*, Vol. 20, No. 1, pp. 131-143, 2021.
[22] F. Fukumoto and Y. Suzuki, "An Empirical Approach to Text Categorization Based on Term Weight Learning," *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pp. 71-79, 1998.
[23] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf.idf," *Proceedings of 4th International Conference on Data Management Technologies and Applications*, pp. 26-37, 2015.
[24] B. Guo, C. Zhang, J. Liu, and X. Ma, "Improving text classification with weighted word embeddings via a multi-channel TextCNN model," *Neurocomputing*, Vol. 363, pp. 366-374, 2019.
[25] T. C. T. Tran, H. X. Huynh, P. Q. Tran, and D. Q. Truong, "Text Classification Based on Keywords with Different Thresholds," *Proceedings of the 2019 4th International Conference on Intelligent Information Technology*, pp. 101-106, 2019.
[26] G. Szarvas, "Hedge Classification in Biomedical Texts with a Weakly Supervised Selection of Keywords," *Proceedings of ACL-08: HLT*, pp. 281-289, 2008.
[27] C. Hashimoto and S. Kurohashi, "Blog Categorization Exploiting Domain Dictionary and Dynamically Estimated Domains of Unknown Words," *Proceedings of ACL-08: HLT, Short Papers*, pp. 69-72, 2008.
[28] J. Bai, H. Shinnou, and K. Komiya, "Domain Adaptation for Sentiment Analysis using Keywords in the Target Domain as the Learning Weight," *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, pp. 37-42, 2018.
[29] Z. Dai and J. Callan, "Context-Aware Term Weighting For First Stage Passage Retrieval," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1533-1536, 2020.

이 상 아

2013년 서울대학교 언어학과, 컴퓨터공학부 졸업(학사). 2016년 서울대학교 언어학과 졸업(석사). 2021년 서울대학교 언어학과 졸업(박사). 2021년 3월~2022년 2월 서울대학교 데이터사이언스대학원 박사후연구원. 2022년 3월~8월 서울대학교 기초교육원 강의교수. 2022년 9월~현재 서울대학교 언어학과 교수. 관심분야는 자연언어처리, 언어모델, 비정형 텍스트 분석, 논증 마이닝

김 석 기

2020년 서울대학교 재료공학부 졸업(학사). 2022년 서울대학교 데이터사이언스대학원 졸업(석사). 관심분야는 한국어 문장 내 관계 추출을 통한 지식 그래프 구축

김 은 진

2021년 서울대학교 독어독문학과 학사 2021년~현재 서울대학교 언어학과 석사 과정. 관심분야는 자연언어처리, 언어모델, 비지도학습, 정보 추출, 프롬프트 튜닝

강 민 지

2021년 서강대학교 영문학부 학사. 2021년~현재 서울대학교 언어학과 석사. 관심분야는 자연어처리, 자연어생성

신 효 필

1988년 서울대학교 언어학과 졸업(학사) 1990년 서울대학교 언어학과 졸업(석사) 1994년 서울대학교 언어학과 졸업(박사) 1997년 12월 University of Missouri, Computer Science 졸업(석사). 1998년 1월~2001년 1월 Computing Research Lab, New Mexico State University. 2001년 1월~2001년 12월 YY Technology in Silicon Valley. 2001년 9월~2003년 2월 서울대학교 전자공학부 BK교수. 2003년~현재 서울대학교 인문대학 언어학과 교수. 2020년 1월~2022년 2월 서울대학교 데이터사이언스 대학원 교무부원장. 관심분야는 사전학습모델 구축과 이를 활용한 자연어처리