JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 7, pp. 1419-1425, Nov. 2017

연관법령 검색을 위한 워드 임베딩 기반 Law2Vec 모형 연구

김 나 리 · 김 형 중
고려대학교 빅데이터응용및보안학과

A Study on the Law2Vec Model for Searching Related Law

Nari Kim · Hyoung Joong Kim
Department of Big Data Application and Security, Korea University

[요 약]

법률 지식 검색의 궁극적 목적은 법령과 판례를 근거로 최적의 법례정보 획득이라고 할 수 있다. 최근, 대규모 자료에서 효율적으로 검색하여야 하는 목적을 달성하기 위하여 텍스트 마이닝 연구가 활발히 이루어지고 있다. 대표적인 방법으로 Neural Net 기반 학습방법인 워드 임베딩 알고리즘을 들 수 있다. 본 논문에서는 한국 법령정보를 워드임베딩에 적용하여 연관정보 검색방법을 연구하였다. 우선 판례의 참조법령을 순서대로 추출하여 모형의 입력정보로 활용하였다. 추출한 참조법령들은 중심법령을 기준으로 주변 법령을 학습하고 임베딩하는 Law2Vec 모형을 작성하였다. 이 모형으로 법령에 대하여 학습을 수행하고 법령 간의 관계를 추론하였다. 본 연구의 모형을 평가하기 위하여 연관법령으로 도출된 결과가 키워드와 밀접한 관련이 있는지 정밀도와 재현율을 계산하여 검증하였다. 실험결과, 본 연구의 제안방식이 기존의 키워드 검색방법보다 연관된 법령을 추론하는 데 유용함을 알 수 있었다.

[Abstract]

The ultimate goal of legal knowledge search is to obtain optimal legal information based on laws and precedent. Text mining research is actively being undertaken to meet the needs of efficient retrieval from large scale data. A typical method is to use a word embedding algorithm based on Neural Net. This paper demonstrates how to search relevant information, applying Korean law information to word embedding. First, we extracts reference laws from precedents in order and takes reference laws as input of Law2Vec. The model learns a law by predicting its surrounding context law. The algorithm then moves over each law in the corpus and repeats the training step. After the training finished, we could infer the relationship between the laws via the embedding method. The search performance was evaluated based on precision and the recall rate which are computed from how closely the results are associated to the search terms. The test result proved that what this paper proposes is much more useful compared to existing systems utilizing only keyword search when it comes to extracting related laws.

색인어 : 텍스트 마이닝, 법률 정보, 머신 러닝, 워드임베딩, Word2Vec, 키워드 검색

Key word : Text Mining, Legal Tech, Machine Learning, Word Embedding, Word2Vec, Keyword

http://dx.doi.org/10.9728/dcs.2017.18.7.1419

Received 19 October 2017; **Revised** 09 November 2017
Accepted 25 November 2017

***Corresponding Author; Hyoung Joong Kim**

Tel: +82-02-3290-4895, 010-6251-6343
E-mail: khj-@korea.ac.kr

Ⅰ. 서 론

.

법률정보 검색은 개인이나 기업, 그리고 국가기관의 판사, 검사, 수사관에 이르기까지 법률적 판단과 해석에 필요한 근거 규정을 찾는데 활용하는 수단으로, 경우에 따라서는 많은 시간과 노력이 투입되는 일이다. 국가법령정보센터에 수록된 현재 유효한 법령[1]이 약 5천 개에 이르고, 법리를 해석한 판례의 수가 증가하고 복잡해짐에 따라 효율적인 정보 검색에 대한 요구는 크게 늘고 있다. 그러나 현재까지 법률정보검색은 법률지식을 토대로 정교한 키워드를 알아야만 가능하다는 한계가 있었다.

그런데 빅데이터 분석 방법인 텍스트 마이닝을 활용하여 법률을 검색하면 정확한 단어와 용어를 모르더라도 주요 개념과 테마를 캡처하여 숨겨진 의미와 관계를 알아낼 수 있다. 이 방법은 법률 실무가에게 검색 문제의 어려움을 해결하는 대안이 될 수 있다. 방대한 단어로 구성된 법률 문서를 분석하고 의미를 파악할 수 있는 알고리즘을 통하여, 시간과 비용의 효율성을 높이고 기존의 검색보다 정확한 결과를 제공할 수 있다.

최근에는 복잡하고 전문적인 법률데이터에 대해서도 인공지능을 활용한 텍스트 마이닝의 연구가 활발하게 이루어지고 있다. 북미와 유럽권 등에서는 법률과 ICT가 융합된 정보서비스인 리걸테크(Legaltech) 산업이 활성화 되고 있는 추세이다. Westlaw와 같은 세계적인 규모의 회사는 원하는 문서를 찾고 의미 있는 정보를 추출하여 법조인의 업무를 지원하는 서비스를 제공하고 있다. 반면 국내에서는 최근 들어 정부가 연구개발 투자를 추진하고 있으나 리걸테크 스타트업이 등장하는 산업 초기 단계에 진입한 상황이다[2].

본 논문에서 법률정보 검색의 효율성을 제고하기 위하여 법령, 판례를 분석해 연관법령정보를 제공하는 방법을 연구하였다. 이 연구의 효과는 정확한 내용을 모르더라도 검색어와 연관된 법령정보를 '쉽고, 정확하게' 검색하는 것이다. 연구 순서는 판례에서 참조법령을 추출하여 법령들을 임베딩하는 Law2Vec 모형을 구축하여 학습을 통해 법령 간의 관계를 추론하는 방법이다. 모형의 결과로 도출된 연관단어와 전문가가 추출한 관련 핵심어와 유사도를 비교하여 연관 법령의 정확도 및 성능을 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구로서 국내 법령 정보 검색 서비스 현황과 텍스트 마이닝 기법을 살펴보고, 3장에서 학습 기반의 법령 데이터를 구축하기 위한 설계 및 검증 결과를 분석한다. 마지막으로 4장에서 본 연구의 결론 및 향후 연구 계획을 제시한다.

Ⅱ. 선행 연구

본 장은 국내 법령정보 검색 연구와 텍스트마이닝 기법을 분석하고, 시사점을 도출하였다.

2-1 국내 법령 정보 서비스 현황조사

법령정보를 빠르고 정확한 검색하는 일은 법령을 해석하고 적용하거나 입법을 검토하는 업무, 그리고 공무원의 행정소관 업무에서도 중요한 과정이다. 그러나 법령정보의 수가 늘어나고 종류도 다양해짐에 따라 법령 상호간의 관계를 파악하는 일이 점점 어려워지고 생활관계나 경제구조도 복잡해져 어느 법령의 한 두 조문만 가지고는 문제를 해결할 수 없는 경우가 많다[2]. 따라서 신속한 해결 필요성 증대 등으로 인하여 연관법령정보의 정확하고 빠른 검색의 중요성은 커지고 있다.

우리나라의 법령정보 제공 현황을 분석하면 법제처가 2009년 국가법령정보센터를 구축한 후 법령정보 검색의 편의를 개선하기 위하여 법령내용과 관련된 상·하위법 비교, 신·구조문 비교, 판례 및 생활법령정보 등 다양한 검색방법을 제공하고 있다. 또한 법령정보 제공방식, 범위 등을 개선하기 위하여 계속해서 노력하고 있다[3].

국내 법령정보를 텍스트 분석 기술을 활용하여 제공하는 방법에 대한 연구는 텍스트 분석 기술 분야에서 초기단계에 있다. 국내 법령을 온톨로지로 변환하는 방법에 관한 연구[4,5,6]가 주를 이루며, 그 밖에 생활용어를 법률용어에 대응하기 위한 연구사례[7]와 토픽 모델링을 활용한 판례분류 연구사례[8]가 있다. 그러나 온톨로지를 기반으로 하는 연구는 도로교통 관련 법규, 철도 분야의 주요 규정에 구축하여 의미 있는 성과를 도출하였지만 수동으로 구축하여야 하는 문제가 있다[4,5]. 자동 구축 방법에 대한 최근의 연구 또한 법령의 문장에 온톨로지 패턴을 명확하게 적용할 수 없을 경우에 변환하기 어렵다는 한계가 있다[6]. 생활용어의 법률용어 대응에 대한 연구는 포털 사이트의 태그정보를 데이터 클러스터링 기법으로 가장 높은 확률의 대응관계를 찾아 자동적인 탐색이 가능하도록 하였다. 하지만 이 연구는 생활용어와 법률용어를 탐색하기 위하여 법률용어 시소로스를 조회하여 수행하여야 하므로[7] 사전에 온톨로지로 구축된 법률용어 시소로스가 필요하다. 토픽 모델링의 판례 분류 방법은 하나의 토픽으로만 분류가 가능하다[8]는 한계가 있어서 방대한 법령 정보에 적용하기 어렵다.

법무실무가들은 법령의 연관정보가 함께 검색되는 것을 필요로 한다[9]. 그러나 적용 가능한 연구사례가 드물고, 대법원 종합법률정보에서 제공하는 기능을 살펴보면 연관정보는 키워드에 그치고 있다(그림 1,2 참조).

그림 1. 키워드 기반의 연관검색어
Fig. 1. Related query based on keyword

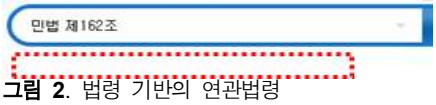

그림 2. 법령 기반의 연관법령
Fig. 2. Related query based on article

표 1. 연관검색 기능현황
Table 1. Related query search systems

Information provider	Related keyword	Related article
Supreme court of Korea	O	X
Korea Ministry of Government Legislation	O	X
LaWnB	O	X

따라서 연관법령을 검색할 수 있는정보 검색 기법에 대한 연구가 요구된다(표 1 참조). 본 연구에서는 판례를 학습하여 법령의 관계를 도출하고 연관법령을 추론할 수 있는 방법을 연구하였다.

2-2 Word Representation

텍스트 분석 기법 중에서 자연어 처리는 사람의 언어를 기계가 분석하기 위한 연구분야로, 벡터표현(vector representation) 방법에 대한 연구가 진행 중이다[10].

단어를 '벡터화'하는 방법으로 초기 연구모델에 'one-hot encoding' 방식이 있다. 길이 n짜리의 벡터를 만들고 그 단어가 해당하는 자리에 1을 넣고 나머지 자리에 0을 넣는 방식이다. 그러나 '단어가 본질적으로 다른 단어와 어떤 차이점을 가지는지 이해할 수가 없다' 는 한계가 있었다.

이러한 한계를 극복하기 위해 연구자들은 단어 자체가 가지는 의미를 다차원 공간에 '벡터화' 하는 방식을 고안하게 되었다. '비슷한 분포를 가진 단어들은 비슷한 의미를 가진다' 는 언어학의 가정에 입각하여 1990년대부터 여러 모델이 제안되었는데 2000년대에 와서 neural network 의 학습 원리에 기반을 두는 'NNLM(Neural Network Language Models)' [11]이 만들어졌다. 'NNLM'은 n-gram의 언어모델을 신경망을 이용하여 구현한 후 목표단어의 앞뒤 단어를 입력받아 목표단어들과 의미적으로 연관성을 갖도록 학습시킨다. 그러나 NNLM의 학습에는 많은 시간이 필요하며 대부분의 시간이 hidden layer와 output layer 사이의 계산에서 소요되는 문제가 있었다[9]. 2013년에 제안된 Word2Vec[12]은 hidden layer를 감축하고 속도와 정확도를 높일 수 있는 forward/backward propagation 개념을 도입하여 신경망 구성의 단순함에 비해 학습된 단어의 벡터표현에 대한 우수한 성능을 보여준다. Word2Vec은 대표적으로 두 가지 모델이 사용되는데 CBOW(Continuous Bag-of-Words)와 skip-gram 모델이 있다. CBOW는 주변 단어가 주어졌을 때 들어갈 적절한 중심단어를 예상하는 모델이고, skip-gram은 CBOW와 반대로 한 단어를 입력으로 받고 그에 대한 주변단어를 예상하는 모델이다(그림 3 참조).

이러한 word embedding 모델은 '벡터'를 통해 단어의 관계를 추론하거나 의미적인 유사성으로 분류하는 것이 가능해졌다는 점에서 의의를 가진다. 다양한 분야에서 이를 활용한 연구가 진행되고 있는데 국내에서도 문서 분류, 감성분석 등에서 word embedding을 시도하는 연구가 활발히 이루어지고 있다.

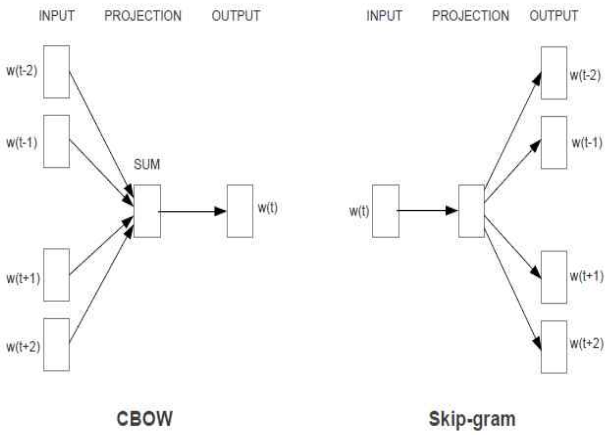

그림 3. Word2Vec 아키텍처
Fig. 3. Word2Vec architecturet

그러나 영어를 근간으로 연구된 Word2Vec 모델을 국내법령에 응용하기 위해서는 법률 정보의 특성을 고려할 필요가 있다. 한글의 경우, 특성상 생략과 도치가 빈번하고 각종 기능어 및 접사가 많이 쓰이며 용언의 활용이 다양하여 분석의 어려움이 있다[13]. 특히 법률 정보[14]는 전문용어로 용어가 간결하지 않고 다의적인 표현, 복합명사 형태가 빈번하게 사용되므로 형태소분석[15]이 복잡하여 연구사례가 많지 않다.

본 연구에서는 이러한 문제점을 극복하기 위하여 국내 법령정보의 특성을 고려하여 법령정보를 분석할 수 있는 방안에 대한 연구를 진행하였다. 국내 법령을 유의미한 벡터로 계산하고 이를 학습하여 법령 간 연관관계를 추론하였다.

Ⅲ. 연관 법령-추출 방법론 - Law2Vec

3-1 연구 모형

본 논문에서는 Word2Vec을 참고로 법령 및 판례를 대상으로 학습을 통하여 법령들을 벡터화하고 법령 간 관련성을 추론하는 법령 기반 Law2Vec 모형을 제안한다. Word2Vec이 문장에 등장하는 중심단어와 주변단어의 분포정보를 분석하여 중심단어의 의미를 유추하는 것처럼, Law2Vec은 판례에 참조된 법령의 분포를 분석하여 특정 법령과 연관된 법령정보를 추론한다. 기존의 Word2Vec을 적용할 때 발생하는 한글 형태소 분석 및 자연어 처리(NLP)의 어려움을 개선하기 위하여 본 연구에서는 구분자 콤마(,)를 기준으로 법령을 파싱(parsing)하고 연관관계를 분석하였다. 따라서 Law2Vec 모형은 법령 본문을 input하는 것이 아니라 법령 조항을 input하고 학습함으로써, 복잡한 한글 전처리의 어려움을 해결하고 결과적으로 연관법령 검색 성능을 높이고자 하였다.

Law2Vec 모형을 연구하기 위해 다음의 세 가지의 모듈로 구성하여 진행한다(그림 4 참조).

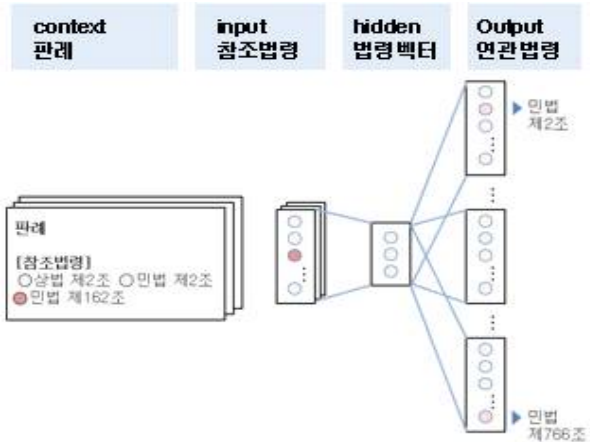

그림 4. Law2Vec 절차도
Fig. 4. Law2Vec workflow

첫째, 입력모듈은 제시어로 판례를 검색하여 corpus를 만들고 판례의 참조법령 정보를 추출하여 context를 구성한다. 판례는 사건에 대하여 법리를 근거로 합리적 해석을 하는 내용으로서 판시사항, 판결요지, 참조법령, 판결내용으로 정형화된 틀을 갖추고 있다. 판례에서 참조한 법령은 해당사건 및 해당법리의 주제가 담긴 가장 기초적인 정보라고 할 수 있다. 쟁점이 비슷한 판례들에서 하나의 참조법령이 반복해서 등장하거나 여러 개의 참조법령이 동시에 등장하는데, 이러한 참조법령의 특성은 같은 맥락에서 등장하는 단어들이 같은 의미(semantic) 를 공유한다는 distributional hypothesis [16]와 유사하다.

둘째, 학습모듈은 context의 각 법령들을 hidden layer와 output layer에서 벡터화한다. 추출된 참조법령은 주변법령의 분포정보와 분석하여 hidden layer에서 고유한 벡터로 산출하고 법령 간 관계를 도출하는 학습을 진행한다. 수집된 corpus를 학습시키기 위한 네트워크 모델은 skip-gram 모델을 사용하여 중심법령 v_c를 기준으로 주변법령 u_o를 예측하는 모형을 구축한다. 계산식은 아래와 같으며, o는 주변법령(surrounding law), c는 중심법령(context law)이다. 따라서 $P(o|c)$는 중심법령(c)가 주어졌을 때 주변법령(o)이 등장할 조건부 확률을 의미한다 [17]. skip-gram 모델에 업데이트하여 학습을 진행하여 중심법령이 주어졌을 때 주변법령을 예측한다.

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \quad (1)$$

마지막으로 출력모듈은 벡터의 유사도를 계산하여 검색하는 법령에 대응하는 연관법령을 출력한다. 생성된 Law2Vec 모델에 참조법령리스트를 넣어 학습시키고, 학습이 완료되면 각 법령의 단어벡터의 코사인 유사도가 가장 높은 법령 정보를 제시할 수 있다. 이러한 모형의 학습결과를 검증하기 위하여 제시어에 따른 관련 법조문을 출력하고 이결과를 휴리스틱에 의하여 추출된 핵심어의 포함여부를 기준으로 성능을 평가한다.

3-2 실험

본 연구에 제안된 모형을 실험하기 위하여 '대법원 종합법률정보'에서 공개하는 대법원 민사소송 판결을 수집하였다.

표 2. 소멸시효 말뭉치
Table 2. Corpus of extinctive prescription

Keyword	Category		Doc	Corpus
	Adjudicating court	Litigation Case		
Extinctive Prescription	Supreme court of Korea	Civil case	911	Yes
		Criminal case	11	No
		Family case	6	No
		Tax case	93	No
		Administrative litigation case	58	No

표 3. 손해배상 말뭉치
Table 3. Corpus of indemnification for damage

Keyword	Category		Doc	Corpus
	Adjudicating court	Litigation Case		
Indemnification for damage	Supreme court of Korea	Civil case	6,988	Yes
		Criminal case	141	No
		Family litigation	35	No
		Tax case	67	No
		Administrative litigation case	137	No

Corpus는 모든 텍스트 데이터를 대상으로 분석을 수행하는 것이 이상적이나, 시간과 자원의 한계를 고려하여 특정 주제를 선정하고 이에 대한 사례 연구를 진행하였다. 특정주제로써 '소멸시효','손해배상'으로 검색한 판례를 수집하였다. 검색결과는 Python으로 자동화하여 수집하였다. 수집한 대상은 표 2,3과 같다.

약 7,000개의 판례가 수집됐고 판례의 참조법령 부분을 추출하여 학습에 사용하였다. 먼저 참조법령을 구분자(,)를 기준으로 파싱(parsing)하였다. 다음으로 법령의 단위가 다양해서 의미단위가 분산되는 것을 최소화하기 위하여 해당 참조법령이 법의 조, 항, 호, 목 등의 하위단위로 명시되어있더라도 조 단위로 통일되도록 추가적인 전처리 절차를 진행하여 텍스트 분석이 용이한 데이터구조로 변환하였다.

훈련 데이터를 학습하기 위하여 Google에서 제공하는 Word2Vec 패키지[18]를 활용하여 중심법령과 주변 법령을 skip-gram 방식으로 학습하였다. 등장 횟수가 20 이하인 단어는 학습 샘플링에서 제외하였고, 300차원짜리 벡터스페이스에 임베딩하였다. 특정 법령의 유사한 법령 결과를 도출한 결과는 아래 내용과 같다.

표 4. 민법 제 168조의 연관법령 결과 비교
Table 4. Comparison of the output - Article 168 o the Civil Act

Option	Target article	Number of results	Search results	Legal text
Keyword Search (Existing methods)	Article 168 of the Civil Act (Causes Interrupting Extinctive Prescription)	3	Article 168 of the Civil Act	(Causes Interrupting Extinctive Prescription) Extinctive prescription shall be interrupted in any of the following cases: 1. Demand; 2. Attachment, provisional attachment or provisional disposition; 3. Acknowledgment.
			Article 54 of the Framework Act on National Taxes	(Extinctive Prescription of National Tax Refund) (…) (2) The provisions of the Civil Act shall be applied to the extinctive prescription under paragraph (1), except as otherwise provided for in this Act or tax-related Acts. (…) the claim under subparagraph 1 of Article 168 of the Civil Act shall be deemed (…)
			Article 111 of the Industrial Accident Compensation Insurance Act	(Relationship with other Acts)(1) With respect to an interruption of prescription, the filing of a request for examination or reexamination as prescribed in Articles 103 and 106 shall be deemed a judicial claim as prescribed in Article 168 of the Civil Act.
Law2Vec (Proposed methods)	Article 168 of the Civil Act (Causes Interrupting Extinctive Prescription)	3	Article 169 of the Civil Act	(Effect of Interruption of Prescription) The interruption of prescription shall be effective only between the parties themselves and their successors in title.
			Article 17 of the Bills Act	(Applicable regulations of draft bill) For the promissory note, the provisions of the following draft shall apply mutatis mutandis to the nature of the promissory note.(…)8.Extinctive Prescription (…)
			Article 430 of the Civil Act	(Appendant Nature of Surety Obligation) If the burden of a surety is greater than that of the principal obligor either as to its subject or its terms, it shall be reduced to the extent of the principal obligation.

표 5. 민법 제 755조의 연관법령 결과 비교
Table 5. Comparison of the output - Article 755 of the Civil Act

Option	Target article	Number of results	Search results	Legal text
Keyword Search (Existing methods)	Article 755 of the Civil Act (Supervisor's Liability)	1	Article 755 of the Civil Act	(Supervisor's Liability) (1) If a person who has caused any damage to another is exempt from liabilities under Article 753 or 754, the person who is under a legal duty to supervise such person shall be liable to make compensation for the damage: Provided, That the same shall not apply, if the person so supervising has not been negligent in performing his/her duty of supervision. (…)
Law2Vec (Proposed methods)	Article 755 of the Civil Act (Supervisor's Liability)	3	Article 753 of the Civil Act	(Minor's Competency)In the event that a minor has caused damage to another, if he was not in possession of sufficient intelligence to understand his responsibility for the act, he shall not be liable for damages resulting therefrom.
			Article 709 of the Civil Act	(Presumption of Representative Power of Managers) The partners who manage the partnership affairs shall be presumed to have a representative power of the management of the partnership affairs.
			Article 82 of the Labor Standard Act	(Bereaved family's compensation) (1) If an employee dies in work, the employer shall compensate the survivor for the average wage of 1,000 days after the employee's death, without delay. (2) The extent of the survivor's remuneration, the rank of survivor's compensation, and the rank of survivor's compensation in the case of death of a person determined to receive compensation shall be determined by Presidential Decree.

법령을 검색하기 위하여 기존의 키워드로 법령을 검색한 결과와 본 논문에서 제안하는 Law2Vec으로 법령을 검색한 결과를 각각 비교하여 살펴보면, Law2Vec 모형이 단순 키워드로 검색하는 결과에 비해 의미상 관련이 높은 연관법령까지 도출하는 것을 확인할 수 있다. Law2Vec을 통한 법령의 검색 결과는 1건에서 n건으로 output의 크기를 설정할 수 있는데, 본 논문에서는 Law2Vec 검색결과를 3건으로 설정하여 그 법령의 내용을 표 4, 5와 같이 비교하였다.

Law2Vec을 통해 분석된 결과로써 첫째, '민법 제168조(소멸시효 중단사유)'의 연관법령은 '민법 제169조(시효중단의 효력)', '어음법 제77조(환어음의 규정의 준용)', '민법 제430조(목적, 형태상의 부종성)'이 도출되었고, 둘째 '민법 제755조(감독자의 책임)'의 연관법령은 '민법 제753조(미성년자의 책임능력)', '민법제709조(업무집행자의 대리권추정)', '근로기준법 제82조(유족보상)'이 도출되었다.

기존의 키워드 검색은 '민법 제168조'라는 '키워드'가 일치하는 경우에 한하여 검색되었다. 그러나 Law2Vec은 '민법 제168'조로 검색하면 '민법 제168조'의 벡터정보를 통해, '소멸시효의 중단'과 의미상 유사성이 있는 것으로 볼 수 있는 '시효중단의 효력'을 다룬 법조문이나 '감독자의 책임'과 관련이 있는 '미성년자의 책임능력', '업무집행자의 대리권 추정'에 대한 규정이 검색된다는 점에서 기존 키워드 검색과 차이가 있다.

3-3 성능평가

본 모형의 성능을 검증하기 위하여, Law2Vec 모형에서 도출된 연관법령 정보들이 전문가가 도출한 핵심어와 얼마나 관련이 있는지 비교하였다. 법무부 보고서[19]를 참고하여 핵심어를 gold standard로 마련하였고(표 7 참조), 연관법령 상위 30개의 결과가 gold standard keyword를 얼마나 포함하는지 정밀도와 재현율을 평가하였다. 정밀도와 재현율은 표 6을 참고한 계산식으로 분석하였다. system output의 positive와 negative는 Law2Vec에서 most_similar 함수의 positive 인수와 negative 인수로 나온 결과값에서 gold standard가 포함되면 'true positive'로 판단하여 계산하였다.

$$\text{Precision} = \frac{tp}{tp + fp} \quad (2)$$

정밀도= 검색된 적합 법령 수/(검색된 적합 법령 수 + 검색된 부적합 법령 수) * 100

$$\text{Recall} = \frac{tp}{tp + fn} \quad (3)$$

재현율= 검색된 적합 법령 수 /(검색된 적합건수 + 검색되지 않은 적합건수)* 100

표 6. 정밀도, 재현율 계산식
Table 6. Confusion matrix

Confusion matrix		Gold standard	
		True	False
System Output	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

표 7. 소멸시효 핵심어
Table 7. Gold standard keyword

Keyword in Ministry of Justice 2007' Research
extinctive prescription, claim, marriage. extinctive prescription, property of inheritance, short-term extinctive prescription, extension of period, request, termination of a right, seizure,period of prescription, starting point of reckoning,temporary disposition, illegal act, approval, indemnification for damage, incompetent (…)

표 8. 구간별 정밀도, 재현율 결과
Table 8. Precision and recall by rank

Targeted - Civil Act	Rank	Precision	Recall
Article 766 (Prescription in respect of Right to Claim for Damages)	10	60%	55%
	20	45%	50%
	30	57%	63%
Article 166 (Starting Point of Computing Extinctive Prescription)	10	70%	70%
	20	63%	60%
	30	64%	60%
Article 168 (Causes Interrupting Extinctive Prescription)	10	50%	63%
	20	55%	65%
	30	50%	68%
Average	20	57%	62%

연관법령 결과를 상위10개, 20개, 30개 구간별로 나누어 정밀도와 재현율을 비교하였다. Rank는 연관법령 결과의 순위를 의미하고, 결과는 표 8과 같다. 상위 10개의 법령들은 정밀도가 재현율보다 높은 것으로 측정되고, 평균적으로 정밀도는 57%이고, 재현율은 62%로 재현율이 더 높게 측정되었다. 상위에 rank된 연관법령이 핵심어를 포함하고 있는 유사한 법령일 확률이 높았다. 연관법령을 많이 추출할수록 positive관계로 도출된 법령은 negative관계로 도출된 법령보다 핵심어가 포함된 법령이 많이 제시되었다. 이러한 결과는 Law2Vec의 성능이 Precision뿐만 아니라 Recall 측면에서도 결과가 균형을 이루는 것을 의미한다. 따라서 Law2Vec모형을 통해 도출된 연관법령이 검색하는 법령과 의미상 관련 있는 결과를 제공하는 것을 평가할 수 있었다.

Ⅳ. 결 론

본 연구에서는 판례의 법령을 벡터로 하여 법령벡터를 학습하고 임베딩한 결과가 법령 간의 관계 및 연관된 법령을 추론하는데 유용함을 알 수 있었다. 기존에는 법령을 검색하는 경우, 연관법령을 찾으려면 검색어를 여러 번 수정하거나 판례 내용의 법령정보를 참고하기 위하여 많은 판례를 열람해야하는 번거로움이 있었다. 본 연구에서 법령 간 유사도 검색으로 연관법령을 쉽게 검색할 수 있는 새로운 방법을 제시하고 이론적 근거를 마련함으로써 법령 활용의 효과를 높일 수 있을 것으로 여겨진다.

모델의 확장성을 위해서는 input layer 측면에서는 유사한 법령정보 집합의 corpus가 필요하고, output layer측면에서는 성능평가를 위하여 gold standard로 활용될 검증자료가 필요하다. 휴리스틱에 의한 핵심어로 연관법령의 결과와 대조하는 평가방식은 의미 상 일치하는 법령을 일률적으로 평가하기 어렵다. 이러한 부분을 이 분야의 전문가 집단의 폭넓은 검토와 협업이 이루어진다면, 의미 있는 법률 정보 지식 제공 방식을 연구하는데 보다 효과적일 것으로 판단된다.

참고문헌

[1] Statute Status Report , [Internet] available at http://www.moleg.go.kr/lawinfo/status/statusReport

[2] H. J. Jeon, "Legal Tech Industry Status and Implications," Hyundai Research Institute, vol. 16-31. no. 669. pp. 1-11. Dec 2016.

[3] M. H, Koh, "A Study on Advancement Provision of Legal Information," Korea Ministry of Government Legislation, no. 11-1170000-000460-01, pp. 1-121. Sep 2012.

[4] I. H. Chang, "Developing and Evaluating an Ontology-based Legal Retrieval System," Journal of the Korean

Society for Library and Information Science, vol. 45, no. 2, pp. 345- 366, Mar 2011.

[5] M. J. Won, “A Development of Ontology-Based Law Retrieval System: Focused on Railroad R&D Projects,” Journal of Society for e-Business Studies, vol. 20, no. 4, pp. 209-225, Nov 2015.

[6] J. H. Kim, “A Study on Legal Ontology Construction,” Journal of the Korea Society of Computer and Information, vol. 19, no. 11, pp. 105-113, Nov 2014.

[7] J.H. Kim, “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.

[8] J. S. Shim, “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.

[9] J. H. Kim, “Exploring the Lawyers’ Legal Information Seeking Behaviors for the Law Practice,” Journal of the Korean Society for Information Management, vol. 32, no. 4, pp. 55-76, Dec 2015.

[10] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent Trends in Deep Learning Based Natural Language Processing,” arXiv preprint arXiv:1708.02709, 2017.

[11] Y. Bengio, R. Ducharme, P. Vincent et al., “A neural probabilistic language model,” Journal of Machine Learning Research, vol. 3, pp. 1137–1155, 2003.

[12] 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.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781v3, 2013.

[14] R. Andrii, “Semiotic Analysis of Korean Legal Terms,” Journal of Korean Culture, vol. 10, pp. 26-30, Feb 2008.

[15] C. Park, K. Kim, and D. Seong, “Automatic IPC Classification of Patent Documents Using the Term Clustering,” Journal of Korean Institute of Information Technology, vol. 12, no. 9, pp.127-139, Sep 2014.

[16] Z. S. Harris, "Distributional Structure," Word, vol. 10, no. 2-3, pp. 146–162. 1954.

[17] Word2Vec Research, [Internet] available at https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/03/11/embedding/

[18] Word2Vec Tutorial, [Internet] available at https://rare-technologies.com/deep-learning-with-Word2vec-and-gensim/

[19] K. Y. Lee, “Jurisprudence for the Advancement of the Statute of Limitations in Korean Civil Law,“ Ministry of Justice, Republic of Korea, Research Report, Dec 2007.

김나리(Nari Kim)

2012년 : 동국대학교 법학과 학사
2016년~현재 : 고려대학교 정보보호대학원 빅데이터응용 및 보안학과 (석사과정)
2013년～현 재: 딜로이트 안진회계법인

※관심분야 : 빅데이터 분석, 텍스트 마이닝, 머신러닝, e-Discovery, 디지털 포렌식 등

김형중(Hyoung Joong Kim)

1978년 : 서울대학교 전기공학과 학사
1986년 : 서울대학교 제어계측공학과(공학석사)
1989년 : 서울대학교 제어계측공학과(공학박사)
1989년～2006년: 강원대학교 교수
2006년～현 재: 고려대학교 정보보호대학원 교수

※관심분야 : 컴퓨터보안, 패턴인식, 가역정보은닉, 머신러닝, 빅데이터분석 등