「정보시스템연구」 제30권 제2호
한국정보시스템학회
2021년 6월, pp. 57~78

http://dx.doi.org/10.5859/KAIS.2021.30.2.57

혼합 임베딩을 통한 전문 용어 의미 학습 방안

김병태* · 김남규**

〈목 차〉

Ⅰ. 서 론

‘이미 도래한 미래’라는 인공지능 시대의 핵심 과제 중 하나는 자연어 처리(natural language processing)이다. 자연어 처리 (NLP)란 인간의 언어인 자연어와 인공지능으로 대표되는 컴퓨터의 상호 작용, 또는 컴퓨터에 의한 인간 상호의 의사소통을 목적으로 하며, 개별 언어의 발화 집단이 역사적으로 사용해 왔던 말과 글을 컴퓨터에 의한 처리가 가능하도록 연산의 대상으로 변화시키고, 그러한 결과물을 다시 인간이 인식하고 활용할 수 있도록 만드는 작업을 의미한다(Jurafsky and Martin, 2009). 그 처리 대상을 기준으로 할 때 음성인식(speech recognition)과 텍스트 처리(text processing)로 구분될 수 있으나, 이들 모두는 결국 정보 검색, 정보 추출, 감성 분석, 질의 응답, 기계 번역 등의 구체적 과제(specific task) 해결을 목표로 하는 일련의 과정이다. 이러한 자연어 처리 연구 영역은 일상 생활 문제의 해결에 국한되지 않고, 정치, 법률, 행정, 의료, 보건 등 다양한 전문분야와 결합되어 확대되고 있다.

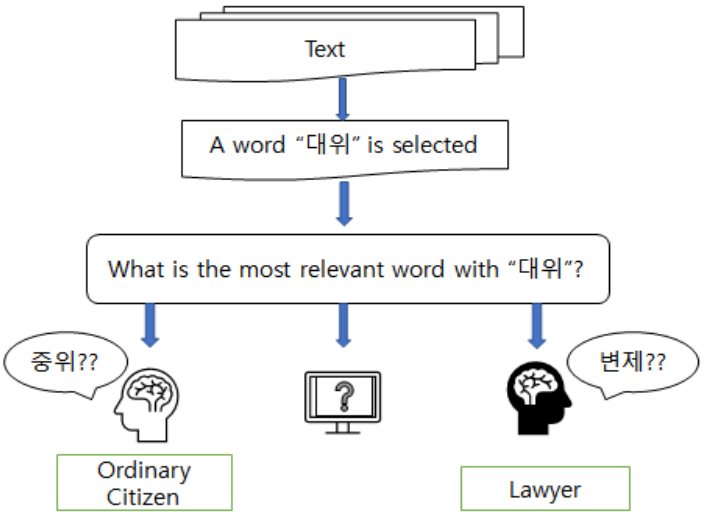
본 연구는 이중 법률 분야와 관련한 자연어 처리 과제를 다룬다. 2019년 경우 한해에 법원에 접수된 사건 건수가 17,720,328 건에 달하

* 국민대학교 비즈니스 IT 전문대학원 석사과정 augustino@kookmin.ac.kr
** 국민대학교 비즈니스 IT 전문대학원 교수 (교신저자) ngkim@kookmin.ac.kr

고, 부수 사건을 제외한 본안 사건만도 1,461,440 건에 달한다(대법원 사법연감(통계), 2019). 이들 모든 사건에서 법원에 의하여 작성된 조서나 재판서 등은 물론 당사자가 작성한 서류 및 제3자 작성의 서류와 당사자가 제출한 증거서류 등 일체의 문서에 대해 일정한 보존의무가 부과된다(대법원 재판사무규칙 18조 이하). 이에 더해 사법 절차가 준용되는 각종 재결, 행정심판 사건 등을 고려할 때, 변호사, 변리사 등 법률 전문직과 법학대학원 등 교육기관이 사적으로 생산하는 법률 문헌 등을 제외하더라도 해마다 엄청난 양의 법률문서가 공적인 영역에서 생산되고 있음을 알 수 있다. 나아가 현재 추진되고 있는 차세대재판사무 시스템 및 등기시스템의 구축에서도 자연어 처리를 통한 정보처리의 고도화가 시도되고 있으므로(대법원 보도자료 “전국법원장 회의”, 2020. 12.), 이를 뒷받침하기 위한 법률 분야에서의 자연어 처리 연구 역시 시급히 다루어질 필요가 있다.

자연어 처리를 법률 분야에 접목하는 과정에서 필연적으로 아래와 같은 문제가 상정된다. 우리말은 크게 일상어로 전문어로 나뉘어지는데(강현화, 2016), 전문어는 “학술 기타 전문 분야에서 특별한 의미로 쓰이는 말”로 정의되며, 일상어는 “평상시에 늘 쓰는 말”의 뜻을 가진다. 가령 일상어에만 익숙해서 일반적인 한국어의 어휘와 문법에 관해 매우 뛰어난 사람이 갑자기 법률가의 언어를 이해하고 구사해야 할 상황이 주어진다고 가정하자. 이 경우 당연히 과거 학습의 결과물로서 그가 가진 어휘 이해와 구사력이 법률가에 상응하게 발휘될 것이라 믿기 어려울 것이다. 기계에 의한 학습 과정에서도 이와 유사하게 학습의 대상이 그 결과를 규정하는 문제가 발생한다. 이러한 문제는 일상어만을 대상으로 학습된 자연어 처리 모델이 전문 분야의 과제 해결에 활용될 때 큰 부작용을 야기하게 된다.

이러한 문제의 원인은 크게 두 가지로 요약



<그림 1> 일반어와 전문어에서 상의한 의미로 사용되는 용어 예

된다. 첫째, out-of-vocabulary(OOV)의 문제이다(Garneau et al., 2018). 이는 특정 전문 분야의 용어가 아예 학습 과정에서 제외되어, 해당 전문 용어들이 자연어 모델에 전혀 포함되지 않는 문제를 의미한다. 둘째는 전문 용어의 전문적 의미가 모델에 반영되지 않는 문제이다(Pilehvar et al., 2017). 이는 일상어에서도 사용되는 어휘이지만, 일상어로 사용될 때와 전문 분야에서 사용될 때 상이한 의미를 갖는 용어의 특성이 충분히 학습되지 않음을 뜻한다. 전자의 문제는 일상어 학습 데이터의 크기가 충분히 크다면 상당 부분 극복될 수 있는 문제라고 볼 수도 있으나, 후자의 문제는 해당 전문 분야의 데이터가 학습 자료로 제공되지 않는다면 극복하기 어렵다. 이러한 문제의 부작용은 “대위”라는 용어의 다중 의미로 인해 야기되는 <그림 1>의 현상을 통해 설명할 수 있다.

<그림 1>은 본 연구 과정에서 실제로 수행한 실험 결과의 일부를 활용한 예로, 다음과 같이 해석된다. 일상어만으로 학습된 모델의 경우 “대위”와 가장 관련이 깊은 단어는 “중위”로 나타나며, 법률 전문어만으로 학습된 모델의 경우 “대위”와 가장 관련이 깊은 단어는 “변제”로 나타난다. 이러한 결과는 법률실무에서 흔히 사용되는 “대위변제”와 같은 용례가 반영된 것이다(현암사 법률용어사전, 2019). 이러한 예는 일상어만으로 학습된 모델은 “대위”라는 단어가 법률 분야에서 사용되었을 때 “변제”와 갖는 관련성을 포착하는데 실패할 가능성이 크다는 점을 나타내며, 더 나아가 일반어만을 학습한 모델은 특정 전문 분야의 특성을 충분히 반영하지 못하기 때문에 그 상태 그대로는 해당 전문 분야의 자연어 모델로 사용될 수 없음을 시사한다. 이와 반대로 처음부터 전문 분야의 문서만을 학습하여 해당 분야의 자연어 모델을 구축하려는 시도는, 방대한 데이터 확보의 어려움으로 인해 학습이 충분히 수행되기 어렵다는 한계를 갖는다. 이는 일반인의 일상어를 통한 질의에 대해 법률 전문 문서에 기반한 응답을 제공하는 질의-응답 시스템의 구축 및 검색어 확장, 그리고 법률 문서를 간명한 일반어로 요약하는 등의 응용 시스템 구현을 위해 법률 문서의 특수성과 일반 문서의 일반성이 함께 어우러진 언어 모델이 반드시 필요함을 나타낸다.

본 연구는 위와 같은 문제의식에서 출발하여, 법률 분야의 소량의 문서가 다량의 법률 문서와 함께 학습되었을 때, 법률 분야의 고유한 언어적 의미가 학습에 어느 정도 반영되는지를 정량적으로 평가할 수 있는 평가 척도를 제시하고자 한다. 먼저 법률 분야 문서를 일반 문서와 함께 학습시켰을 때, 법률 문서에서의 어휘들의 관계가 일반 문서와 함께 학습된 경우에도 그대로 유지되는지, 혹은 변화 또는 왜곡되는지 살피고자 한다. 이는 전문 용어들의 맥락과 의미가 일상어와 함께 학습되는 과정에서 어떤 영향을 받는지 추적하는 작업이다. 다음으로 본 연구에서는 전문 문서 수와 일반 문서 수의 상대적 비율의 변화가 전문 분야 어휘 특성의 반영에 어떤 영향을 끼치는지 살펴본다.

본 연구의 이후 구성은 다음과 같다. 우선 다음 장인 2장에서는 본 연구와 밀접한 관련이 있는 선행 연구의 성과를 요약한다. 다음으로 3장에서는 일반 문서와 전문 문서의 혼합 임베딩(mixed embedding) 방법론을 제안하고, 혼합 임베딩이 용어의 전문적 의미 학습에 미치는 영향을 측정하는 방안을 제시한다. 본 연구에서

제안하는 혼합 임베딩의 효과성 측정 방안을 실제 사례에 적용한 실험 결과는 4장에서 소개하고, 마지막 장인 5장에서는 본 연구의 기여와 한계를 요약한다.

Ⅱ. 선행 연구

2.1 코퍼스(corpus)

코퍼스는 상당한 분량의 구조화된 문서의 집합으로(Leech, 1992), 자연어 처리의 대상이 된다. 우리말로 '말뭉치'로 번역될 수 있는 이 코퍼스는 일련의 규칙을 전제로 의식적으로 수집된 문서들로 구성되며(김한샘, 2019), 그 크기에 있어서도 수집 대상 영역을 전체적으로 나타낼 수 있을 정도의 특성을 가진다. 영어의 경우 Google Books Ngram Corpus(Google, 2020), Oxford English Corpus(Oxford Univ Press, 2014) 및 Brown 대학에서 수집한 Brown Corpus(Brown Univ, 1979) 등이 유명하며, 우리말의 경우, 국립 국어원에 의해 수집된 세종 말뭉치(국립국어원, 2007, 현재 '모두의 말뭉치'에 통합됨) 및 본 연구에서도 활용된 한국어 Wiki 백과(위키미디어재단, 2002) 등이 있다. 코퍼스는 그 문서들의 주된 사용자 집단과 사용영역을 기준으로 일반 코퍼스(general corpus)와 전문 코퍼스(specialized corpus)로 나뉘며(오선영, 2004), 앞에서 소개한 코퍼스의 예들은 모두 해당 문서 사용 집단을 전체 언어 사용자로 하므로 일반 코퍼스에 속한다. 반면 전문 코퍼스는 보다 세분화된 전문 도메인(domain)에서 수집된 것으로서, 해외의 경우 법률 코퍼스로서 British Legal Report Corpus (British Courts, 2010)와 의료 분야의 경우 Medical Corpus: English Corpus From Web(Kilgarriff et al., 2010) 등이 유명하다. 이를 표로 정리하면 다음 <표 1>과 같다. 본 연구에서는 대법원 판례 문서를 수집하여 전문 코퍼스를 신규 구축하고 이를 실험에 활용하였다.

2.2 법률 분야 자연어 연구

국내 연구의 경우 법률 분야에 특화된 코퍼스는 충분하지 않으며 관련 연구도 대부분 동향 연구에 국한되어 있다. 국내의 연구 중 법률 도메인의 자연어 처리를 다룬 최근 연구로는

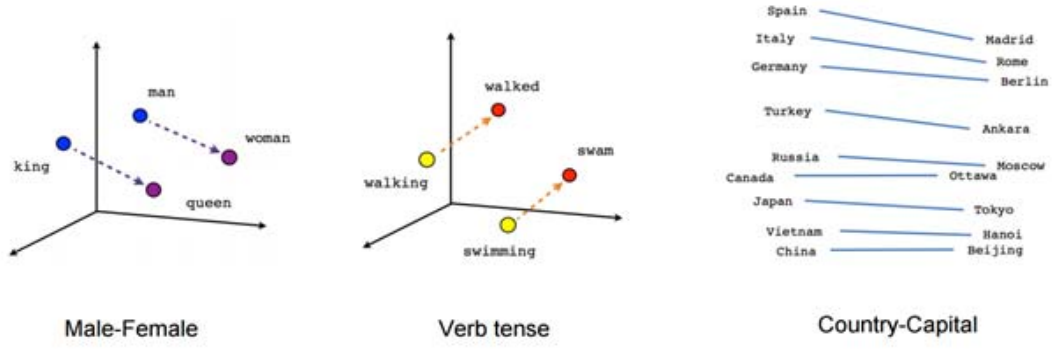
<표 1> 일반 코퍼스와 전문 코퍼스의 예

분야	명칭	언어	수집자
일반	Brown Corpus	영어	Brown Univ.
	Google Books Ngram	영어	Google
	세종말뭉치	한국어	국립국어원
	한국어 WIKI 백과	한국어	위키미디어재단
법률	British Legal Report Corpus	영어	British Courts
의료	Medical English Corpus From Web	영어	Kilgarriff et al.

한국어-영어의 법률용어 번역 및 검색을 위해 로컬 이중 언어 임베딩을 시도한 연구(최순영 등, 2018)와 Word2Vec 기반 모형을 법령 검색에 활용한 연구(김나리 등, 2017) 등을 드물게 찾을 수 있다. 결국 해당 분야의 국내 연구는 아직 충분히 성숙되었다고 보기 어려운 상황이다. 이에 반해 해외의 경우 법률 분야에서 다양한 코퍼스가 만들어졌음은 물론 IBM AI Lawyer Watson으로 대표되는, 활발한 법률 자연어 연구 성과가 실제 실무에의 적용 단계로까지 이어지고 있음을 알 수 있다(Ashley, 2017). 해외 연구의 경우, 법률 분야의 자연어 처리의 응용 과제로서 판결 예측(Judgement Prediction), 질의-응답 시스템(Question Answering), 유사 사례 매칭(Similar Case Matching) 및 텍스트 요약(Text Summarization) 등이 주로 논의되고 있다(Zhong et al., 2020). 이중 판결 예측의 경우 딥러닝 모델로 피의자의 수감 기간을 예측한 Huajie Chen 등의 연구(Chen et al., 2019)가 최근 발표되었으며, 질의-응답 시스템의 경우 미국 사법시험(Bar Exam)의 문제라는 질의에 대한 응답을 시도한 연구(Kim, M., and Goedel, R., 2019)가 주목받고 있다. 또한 유사 사례 분류의 경우 온톨로지 기반으로 법률문서의 의미적 유사성에 기반한 매칭을 시도한 Cardellino 연구(Cardellino et al., 2017)가 이루어졌으며, 법률 텍스트의 요약에 관해서도 기존의 텍스트 요약의 다양한 알고리즘이 법률 문헌을 대상으로 시도되었을 때의 성능 문제를 논의한 연구(Bhattacharya et al., 2019)가 수행되었다. 이처럼 법률 분야 자연어 처리 영역의 해외와 국내 연구의 차이는 부분적으로는 우리 법률 시장의 상대적 협소성 및 이로 인한 연구자의 관심 부족 등에서 그 원인을 찾을 수 있다.

2.3 단어 임베딩 (word embedding)

Bengio et al.,(2003)이 단어 임베딩을 “단어의 분산 표현”이라는 개념으로 규정한 이래, 단어 임베딩은 자연어 처리의 핵심 연구 과제 중 하나로 꼽히고 있다. 이 연구에 따르면 임베딩이란 코퍼스 내의 개별 단어들을 컴퓨터에 의한 처리가 가능하도록 미리 정의된 벡터 공간에서 특정한 값을 가지도록 대응(mapping)시키는 작업을 의미한다. 임베딩을 통해 코퍼스 내의 개별 단어(token)는 N-차원의 벡터 공간 내에서 특정한 위치 값을 갖게 되어, 이후 구체적



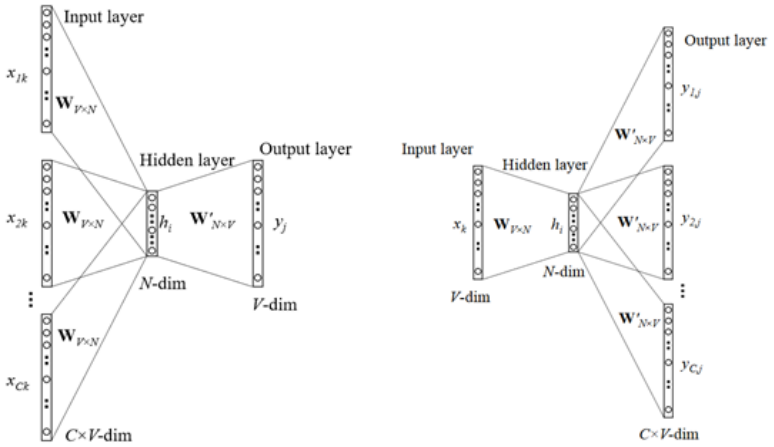
<그림 2> 벡터 공간에서의 단어 표현 예(www.tensorflow.org)

인 하위 작업(downstream tasks)에 활용될 수 있다. 임베딩 학습을 통해 도출된 단어 벡터는 <그림 2>에서와 같이 해당 코퍼스의 구체적인 문법적, 의미적 특성을 반영해야 한다. 이렇게 추출된 임베딩 결과물은 구체적인 추가 하위작업(최병설, 김남규, 2020; 윤상훈, 김근형, 2021 등)에 투입되게 된다.

임베딩을 수행하는 구체적 알고리즘은 매우 다양하다. 대표적인 임베딩 알고리즘으로 Word2Vec(Mikolov et al., 2013 a), GloVe (Pennington et al.,2014), ELMo(Peters, 2018) 및 비교적 최근의 BERT(Devlin et al., 2018)와 GPT-3(OpenAI, 2020) 등을 들 수 있다. 이중 GloVe의 경우 역시 형태소 분석을 거쳐 토크나이징된 개별 단어를 이용하여 임베딩을 수행하며, 다만 Global한 단어 분포를 이용한다는 점에서만 Word2Vec과 차이점을 갖는다. 양자는 유사한 결과를 기대할 수 있으나 GloVe가 보다 큰 컴퓨팅 리소스를 요구한다는 점에서 차이가 있다. 또한 ELMo와 BERT의 경우 사전학습(Pretrained)된 모델을 이용한다는 점에서 Word2Vec과는 근본적인 출발점의 차이를 보인다. 감성 분석이나 문서 분류와 같은 구체적인 응용과제에 있어 BERT 등이 훨씬 향상된 성능을 보인다는 점은 의심할 여지가 없으나 본 연구에 활용되기에는 한계가 있는데, 그것은 BERT에 내장된 토크나이징 기법인 WordPiece Tokenizing이 서로 다른 Corpus에 적용되었을 때 일관된 토크나이징을 보장하지 않는다는 점 때문이다. 따라서 본 연구에서는 연산 부담을 줄이고 상이한 Corpus간의 일관된 토크나이징을 통한 상호 비교를 수행하기 위해 Word2Vec을 통한 분석을 채택하였으며, 그 개념과 구조에 대해서는 다음 절에서 보다 자세히 소개한다.

2.4 Word2Vec

Word2Vec은 Continuous Bag of Words (CBOW)와 skip-gram의 두 가지 구조로 제안되었다(Mikolov, 2013 b). CBOW는 주변 단어들로부터 특정 단어를 예측하는 과정에서 학습이 이루어지는데, 이때 주변 단어들의 수는 파라미터로서 외부로부터 주어지며, 그들의 순서는 학습에 영향을 미치지 않는다. 반면 skip-gram 구조에서는 특정 단어로부터 주위의 단어를 예측하는 방식으로 학습이 이루어진다. 이때 문장 내에서 보다 가까이 위치한 단어가 멀리 위치한 단어에 비해 높은 관련성이 부여된다 <그림 3>. 이 CBOW와 skip-gram 모델은 함께 또는 개별적으로 학습될 수 있으며, Mikolov et al.(2013 b)에 따르면 skip-gram 모델이 학습 시간이 더 걸리는 대신 희소한 단어의 임베딩에도 강하다는 특성을 갖는다. 또한 Word2Vec은 코퍼스의 크기와 파라미터들에 매우 민감하다는 특징이 보고되고 있으며, 이에 따라 차원(dimension), 학습에 동원되는 주변 단어의 수(windows), 그리고 학습 횟수(iteration)의 최적값을 찾기 위한 연구가 다수 진행되어 왔다 (Adewumi et al., 2020). Word2Vec에 의해 임베딩된 개별 단어들의 의미적 유사성은 기본적으로 코사인 유사도에 의해 측정되며, 본 연구의 유사도 측정 또한 이에 의한다.



<그림 3> Word2Vec의 구조 (Mikolov et al., 2013b)

2.5 코사인 유사도 (cosine similarity)

자연어 처리 연구(NLP)에 있어 단어간 유사도의 판단은 크게 시소러스(thesaurus) 기반의 측정과 통계(stochastics) 기반의 측정으로 나뉘어 진다. 시소러스는 단어들의 관계를 보여주는 단어집의 일종으로서, 유의어, 반의어 등으로 구조화된 방식으로 그 관계를 나타낸다(Mooers, 1950). 전산화된 시소러스로서 가장 대표적인 것으로 WordNet(Princeton Univ, 1985)이 있으며, 시소러스에서 규정된 친소관계로써 단어들의 관계를 평가하게 된다. 많은 장점에도 불구하고 시소러스 기반의 평가는 자의적이고 정량화되기 어렵다는 비판에서 자유롭지 않다(Doerr, 2006). 단어간의 유사도를 정량적으로 평가하는 방법으로서 단어간 유클리디언 거리를 이용하는 방법, 자카드 유사도에 따르는 방법, 코사인 유사도에 따르는 방법 등이 제안되어 왔으며, 이들 중 코사인 유사도가 다양한 응용에서 유사도 기준으로 가장 널리 사용되고 있다(Faruqui et al., 2016; Garain et al., 2019 등). 이에 본 연구에서도 위 선행 연구의 방법론을 좇아 Word2Vec을 통해 임베딩된 단어 벡터들에 대해 코사인 유사도(cosine similarity)를 산출함으로써 어휘의 관련성을 파악한다. 코사인 유사도(Church, 2016)는 벡터 공간에 사영(projected)된 두 단어 벡터 사이의 사잇각(angle between two vectors)의 크기로 정의되며, 아래 수식에서과 같이 각 단어의 벡터 값을 내적(Dot Prodcut)한 후, 각 단어의 크기(norm)을 곱한 값으로 나눠주는 방법에 의해 계산된다. 이때 코사인 유사도는 -1과 1 사이의 값을 가지게 되어 1에 가까울수록 유의어, -1에 가까울수록 반의어로 판단된다(Islam, 2008).

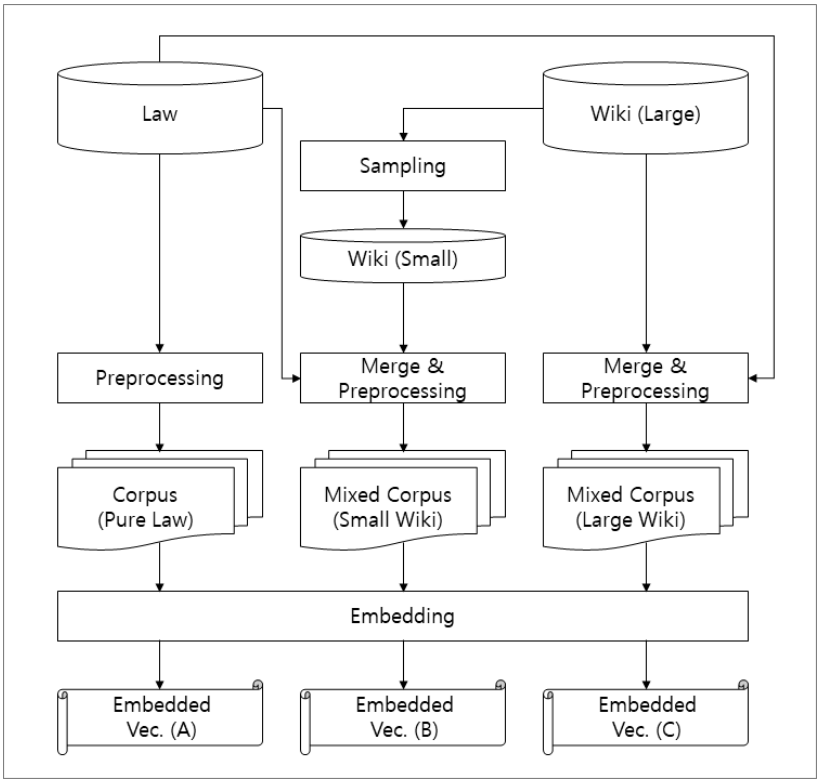
$$cosine(x, y) = \frac{x.y^T}{||x||.||y||}$$

Ⅲ. 제안 방법론 및 실험 설계

3.1 전문 코퍼스 혼합 임베딩 방법론

본 장에서는 전문 코퍼스 문서와 일반 코퍼스 문서에 대한 다양한 비율의 혼합 임베딩을 수행한 뒤, 혼합 임베딩 방법 및 비율에 따라 전문 코퍼스 문서들에서 나타나는 고유한 특성이 임베딩 후에도 유지되는 수준을 측정하는 방안을 제시한다. 또한 전문 문서로서 법률 분야의 대법원 판례 문서, 그리고 일반 문서로서 한국어 위키백과를 선정하여 혼합 임베딩 및 결과 평가를 수행하는 과정을 소개한다. 대법원 판례 문서와 위키백과의 혼합 임베딩은 <그림 4>와 같이 세 종류의 조합으로 실시한다.

위의 세 종류의 임베딩은 모두 유사한 절차에 의해 이루어지지만 세부 구성에는 서로 차이가 존재한다. 우선 순수 전문 임베딩인 A 임베딩은 전문 코퍼스인 대법원 판례 문서만을 대상으로 수행되며, 전처리 과정에서 불용어와 기타 문장 부호 등을 제거하고 명사만을 추출한 후 임베딩이 이루어진다. 다음으로 혼합 임베딩인 C 임베딩의 경우, 법률 문서와 일반 문서 전체를 대상으로 결합(merge)한 후, 이에 대한 전처리와 명사 추출을 거쳐 임베딩을 수행한다. 이 과정에서 추후 전문성 평가에 활용하기 위한 중간 산출물, 즉 위키 문서에는 존재하지 않고 오로지 법률 문서에서만 나타나는 어휘를 추출하여 확보하게 된다. 일반적인 경우



<그림 4> 혼합 임베딩의 세 가지 조합

전문 문서에 비해 일반 문서의 양이 훨씬 많을 것으로 예상하며, 따라서 혼합 임베딩의 경우 전문 문서의 내용이 임베딩 결과에 미치는 영향은 매우 미미하게 나타난다. 따라서 임베딩 과정에서 전문 코퍼스의 내용을 충분히 반영하기 위해서는 전문 문서와 일반 문서의 비율에 대한 조정이 필요하며, 이를 반영한 절차가 부분 혼합 임베딩, 즉 B 임베딩이다. 부분 혼합 임베딩의 경우 전문 문서는 전체를 사용하되 일반 문서의 일부만을 사용하는 방식으로 전문 문서와 일반 문서의 혼합비를 조절하게 된다. 부분 혼합 임베딩에서 일반 문서의 표본 추출(sampling) 비율에 따라 임베딩 결과가 다르게 나타날 것으로 예상하며, 그 비율이 100%인 경우는 혼합 임베딩(C 임베딩), 비율이 0%인 경우는 순수 전문 임베딩(A 임베딩)과 동일한 임베딩이 이루어진다.

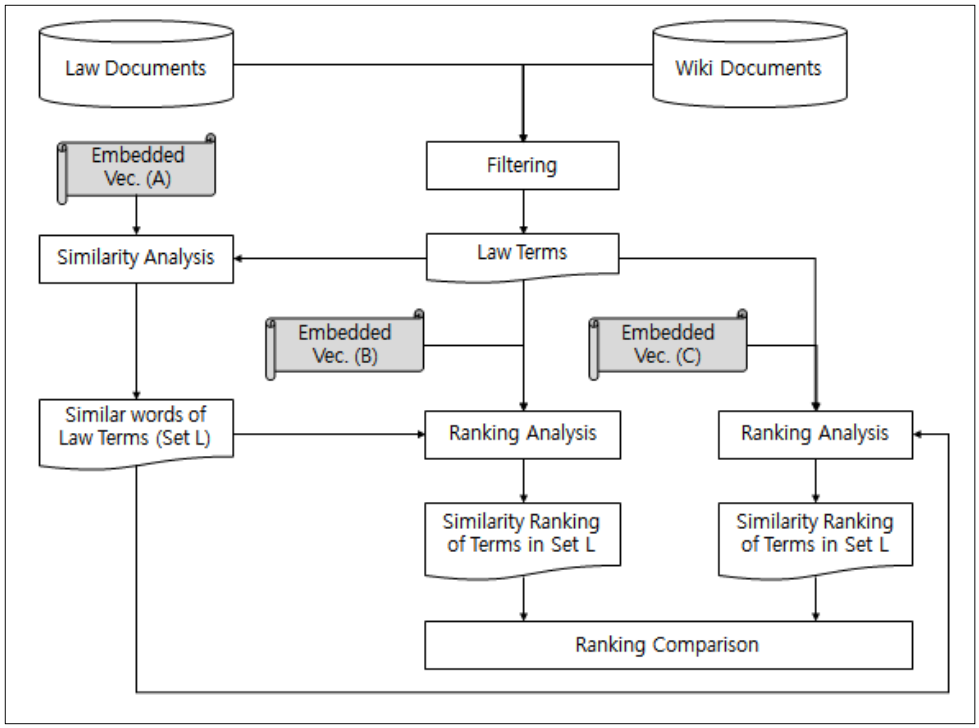
위의 연구 설계에 기초하여 본 연구가 밝히고자 하는 바는 아래와 같이 정리할 수 있다. 우선 본 연구는 상이한 도메인 코퍼스에서 어휘적 특성 및 문맥적 특성이 상이하게 나타날 것이라는 가정에서부터 출발한다. 또한 그 가정이 허용될 수 있다면, 혼합 임베딩의 경우 개별 도메인의 특성은 어떤 방법으로 측정될 수 있는가의 연구 문제에 대해 그 절대적 수준은 측정하기 어렵고 다만 상대적인 비교를 통해 간접적으로 평가할 수 있다는 관점을 견지한다. 즉 위의 세 가지 임베딩 조합 중 전문 문서만 사용한 순수 전문 임베딩이 해당 전문 분야의 특수성을 가장 정확하게 반영할 수 있음은 자명하므로, 이와의 상대적인 비교를 통해 전문 분야 문서만의 임베딩에서 개별 단어가 맺고 있던 최유사 단어의 쌍이 혼합 임베딩의 과정에서 달라지는 정도를 추적하고자 한다. 그 결과 혼합 임베딩 후 최유사 단어의 쌍이 달라지는 현상이 발견된다면 이는 전문 문서에서 사용된 전문 단어의 특성이 변화한 것으로 해석될 수 있으며, 따라서 이를 통해 전문성의 보전/와해 정도를 확인할 수 있을 것이다.

본 장의 이후 부분에서는 위의 세 가지 임베딩 조합 중 전문 문서만 사용한 순수 전문 임베딩이 해당 전문 분야의 특수성을 가장 정확하게 반영하고 있는 것이라고 가정하고, 각기 다른 비율의 혼합 임베딩을 통해 전문 분야의 특수성, 즉 전문성이 임베딩 결과에 어느 정도로 보존되는가를 평가하기 위한 척도 및 측정 방안을 제안한다.

3.2 전문어 의미 존속성 평가 방안

본 절에서는 전문 문서의 특성이 혼합 임베딩 후에 어느 정도 유지되는가를 평가하기 위한 방법으로, 전문어와 일반어 간 유의어 분석과 전문어 간 유의어 분석의 두 가지 방안을 제시한다. 두 가지 방안에 대한 자세한 내용은 본 절의 부절에서 각각 다루며, 두 방법에 공통적으로 적용되는 전체 과정은 <그림 5>와 같다.

우선 <그림 5>에서 음영으로 구분되어 나타난 Embedded Vec.(A) ~ (C)는 <그림 4>에서 제시된 세 가지 임베딩 방법을 각각 적용한 최종 산출물이다. 또한 그림에서 “Filtering”을 통해 도출한 “Law Terms”는 법률 문서에는 나타나지만 일반 문서에는 나타나지 않는 용어, 즉 법률 분야의 전문 용어를 뜻한다. 이러한 전문 용어들의 유의어가 각 임베딩 방법에 따라 얼마나 상이하게 나타나는지를 분석하는 것이 본



<그림 5> 전문 분야의 특수성 유지 수준 평가 개요

실험의 핵심이며, 구체적으로는 순수 전문 임베딩의 결과물에서 확인된 법률 문서 고유 단어의 유의어 관계가 혼합 임베딩과 부분 혼합 임베딩을 통해서도 얼마나 변질되지 않고 유지되는지를 분석하고자 한다. 이때 두 가지 품질 평가 방안 모두 유의어 분석의 기준 용어로 “Law Terms”, 즉 법률 분야의 전문어를 사용하는 점은 동일하며, 3.2.1절에서는 각 전문어와 가장 유사한 일반어(전문어 포함)의 순위를 비교하는 분석을, 3.2.2절에서는 각 전문어와 가장 유사한 다른 전문어의 순위를 비교하는 분석을 수행한다는 점에서 차이가 있다.

3.2.1 전문어와 일반어 간 유의어 분석

본 부절에서는 전문 문서에서만 존재하는 단어들(Law Terms)과 가장 유사한 것으로 확인된 단어들이 일반 문서와의 혼합 임베딩 후에도 얼마나 유사성이 높게 유지되는지를 측정함으로써 혼합 임베딩의 전문성의 존속 정도를 측정하는 방안을 제시한다. 즉, <그림 5>에서와 같이 전문 문서에만 존재하는 단어들을 대상으로 최유사 단어를 식별한 후 각각 확인한 다음, 다양한 비율의 혼합 임베딩 후에 그 유사성 관계가 유지되는 비율을 확인한다. 만약 순수 전문 임베딩에서 나타나는 특수성이 그대로 보존된다면, 혼합 임베딩 후에도 순수 전문 임베딩에서 발견되는 개별 단어들의 최유사 단어쌍은 그대로 보존될 것임을 예상할 수 있고, 그 변화의 비율 역시 0에 가깝게 나타날 것이다. 반면 그 특수성이 혼합 임베딩 과정에서 영향

을 받는다면 단어들 간의 유사도는 변화를 보일 것이며, 그 결과 혼합 임베딩 후의 최유사 단어쌍은 순수 전문 임베딩의 경우와는 다르게 나타나고 변화의 횟수와 비율 역시 증가할 것이다.

또한 전문 문서와 일반 문서의 혼합비를 조절하였을 때 각 임베딩 결과에서 발견되는 최유사 단어쌍이 의미있는 변화를 보이지 않는다면, 문서들의 혼합비와 특수성 보존의 문제는 관련성이 크지 않다고 판단할 수 있을 것이다. 이와 반대로 전문 문서에 대한 일반 문서의 혼합 비율이 높아짐에 따라 순수 전문 임베딩과 혼합 임베딩에서 도출되는 최유사 단어쌍의 차이가 크게 나타난다면, 전문 문서의 특수성과 일반 문서의 크기는 일종의 길항관계(tradeoff)로서 전문성의 보존을 위해서는 혼합비의 적절한 조정이 필요하다고 판단할 수 있다.

3.2.2 전문어 간 유의어 분석

본 부절에서는 앞에서 소개한 전문성 존속 정도의 측정 방안을 보완하기 위한 방안으로, 전문 문서에만 존재하는 단어들 사이의 유사도 변화 양상 측정 방안을 제시한다. 만약 이들 전문어 상호간의 유의어 순위 및 유사도의 크기가 순수 전문 임베딩과 혼합 임베딩에서 크게 다르게 나타난다면, 이는 혼합 임베딩 과정에서 벡터 공간의 변화로 인해 전문어의 성질 역시 영향을 받은 것으로 추론할 수 있다. 또한 앞에서와 마찬가지로 혼합 임베딩에서 전문 문서와 일반 문서의 혼합비를 조절하였을 때 전문어 상호간의 유의어 순위 및 유사도의 크기에 큰 변화가 없다면, 혼합 임베딩은 순수 전문 임베딩으로 형성되는 전문어의 벡터 공간에 큰 왜곡을 주지 않는 것으로 판단할 수 있다. 이 경우 역시 3.2.1에서 소개한 전문어와 일반어 간 유의어 분석의 경우와 마찬가지로, 보다 직관적인 이해를 위해 유지 비율과 횟수에 의한 평가를 시도하였다.

3.3 실험 설계

본 연구의 실험을 위해, 전문 문서로서 일반에게 공개된 대한민국법원 종합법률정보 중 대법원 판례 문서를 이용하였다(법원종합법률정보, 2019). 법률 도메인의 대표적 문서로 평가될 수 있는 대법원 판례는 크게 사건번호, 요지, 주문 및 이유로 구성되며 본 연구에서는 이중 이유만을 발췌하여 사용하였다. 웹 크롤링(web crawling)을 통해 수집된 위 문서는 대법원 선고 일자 기준 2010년 1월 1일부터 2019년 9월 12일까지의 것으로 총 4,253 건, 문장 단위로는 139,232 문장에 달하며, 이는 일반 문서 기준 약 9,000 쪽에 해당한다. 3심제의 사법제도를 채택하는 우리나라의 경우 대법원의 판례는 사실상 하급법원을 구속하는 관계로(법원조직법 제8조), 개별 판례 역시 하급심 및 기타 법률 실무에 대한 지도적 성격을 가지므로 대법원 판례는 법률 문서의 대표로서 손색이 없다고 할 수 있다.

또한 본 연구는 일반 문서로서 한국어 위키백과를 대상으로 하였다. 한국어 위키백과(이하 Wiki)는 전 세계 누구나 편집할 수 있는, 다중 언어 웹 기반 자유 콘텐츠 백과사전 프로젝트의 한국어판이며, 정치, 경제, 사회, 문화, 연예, 취미 등 거의 전 분야를 대상으로 하는 한국어 일반 코퍼스의 대표적인 문서 집합 중 하나

이다(이기창, 2019). 2020년 3월 기준 한국어 위키백과에는 535,000건의 문서가 등재되어 있으며, 평균적으로 매일 122개의 문서가 새로 등재되고 있다(한국어위키백과, 2021). 본 연구에서는 일반 문서의 다양성을 확보하기 위해, 이 중 300,000개의 문서를 분야를 가리지 않고 무작위로 수집하여 실험에 활용하였다.

개별적 임베딩 절차들은 일반적인 자연어 처리의 절차에 따른다. 즉, 각각의 코퍼스에 대한 자료수집, 전처리, 토크나이징(tokenizing), 형태소 분석(parts of sentence), 불용어(stop words) 제거 등의 절차를 진행한 후 각 임베딩을 진행하였다. 혼합 임베딩의 경우, 문서의 순서가 학습에 영향을 줄 가능성을 방지하기 위해, 판례 문서와 Wiki 문서들을 문서 단위로 무작위 혼합하여 임베딩을 수행하였다. 본 연구의 경우 포괄적인 언어 모델을 생성하는 것이 아니므로, 분석의 목적상 판례 문서와 Wiki 문서 공히 명사(NN)만을 대상으로 하였다. 순수 전문 임베딩은 오롯이 판례 문서만을 사용하여 수행되었으며, 혼합 임베딩에는 판례 문서와 Wiki 문서가 사용되었다. 혼합비의 변동에 따른 전문어 의미 존속성의 변화를 추적하기 위해 우선 300,000건의 Wiki 문서 전체와 4,253건의 판례 문서로 혼합 임베딩을 수행하였으며, 다음으로 위 300,000건의 Wiki 문서 중 무작위로 추출한 40,000건의 문서와 4,253건의 판례 문서로 부분 혼합 임베딩을 수행하였다.

구체적인 실험 환경은 아래와 같다. 먼저 전반적인 프로그래밍은 파이썬 3.7 환경에서 수행하였으며, 토크나이징과 형태소 분석은 오픈소스 한국어 형태소 분석기의 일종인 Mecab (Mecab Project, 2016)을 활용하였다. 또한 개별 임베딩은 오픈소스 파이썬 라이브러리인 Gensim을 사용하였으며, 개별적 유사도의 확인은 Word2Vec 모듈 내 most_similar 함수 코드를 일부 수정하여 활용하였다. Word2Vec의 경우 하이퍼파라미터(hyper-parameter)의 값에 상당히 민감하여, 그 최적값을 찾으려는 연구도 다수 수행된 바 있다(Adewumi et al., 2020). 통상적으로 학습되는 주변 단어의 수(windows)로서 5 내지 10이, 차원의 수(dimensions)는 문서의 크기와 단어의 숫자에 따라 50 내지 300이 제시되고 있으며, 반복 학습횟수(iteration) 역시 50 내지 200회가 일반적으로 사용된다. 또한 최소 학습 단어(min_count)의 경우 Word2Vec의 디폴트(default)는 5로 설정되어 있으며, 일반적으로 데이터 셋의 크기와 분석 목적에 따라 출현 빈도 5 내지 30 이하의 단어는 제외할 것이 권장된다(Adewumi et al., 2020). 본 연구에서는 위 선행 연구의 결과들을 참고하여, 모든 임베딩에서 학습 주변 단어의 수는 10, 차원은 200, 반복 학습횟수는 200, 최소 학습 단어 수는 21, 그리고 학습률(learning rate)은 0.01로 설정하여 실험을 수행하였다.

Ⅳ. 실험 결과

4.1 각 임베딩별 유의어 분석

일련의 전처리 과정을 거친 후 각 임베딩의 입력으로 사용한 문서들의 수는 아래의 <표 2>와 같다. 순수 전문 임베딩의 경우 판례 문서 4,253건이 사용되었으며, 부분 혼합 임베딩의 경우 판례 문서 4,253건과 Wiki 문서 40,000건

<표 2> 실험에 활용된 문서와 단어의 크기

	문서 수	단어 수	단어 수 (중복 제거)
순수 전문 임베딩	4,253	4,952,217	7,129
부분 혼합 임베딩	44,253	10,888,335	44,278
혼합 임베딩	304,253	84,191,006	350,907

<표 3> 판례 문서 고유 단어(Law Terms)의 예 (출현 문서 빈도 순 상위 9개)

순위	단어	빈도	순위	단어	빈도	순위	단어	빈도
1	불허가	122	4	불능미수	68	7	수인한도	59
2	지료	95	5	소송요건	67	8	등기필증	57
3	보관증	68	6	임료	63	9	불고불리	54

<표 4> 각 임베딩 결과로 나타나는 “불허가”의 상위 유사어 및 코사인 유사도 값

유의어 순위	임베딩 방법		
	순수 전문 임베딩	부분 혼합 임베딩	혼합 임베딩
1_{st}	'허가', 0.4200	'처분', 0.5389	'신청', 0.5994
2_{nd}	'취허', 0.3973	'허가', 0.5360	'행정청', 0.5818
3_{rd}	'토석', 0.3831	'토석', 0.4734	'처분', 0.5491
4_{th}	'이민', 0.3765	'신청', 0.4586	'행정소송', 0.5443
5_{th}	'처분', 0.3750	'취허', 0.4548	'허가', 0.5418

이 함께 사용되었다. 또한 혼합 임베딩의 경우 판례 문서 4,253건과 Wiki 문서 300,000건이 사용되었다.

앞서 언급한 바와 같이 각각의 문서에 21회 이상 출현하는 단어들의 수는 중복 제거 후 각 임베딩별로 7,129개, 44,278개, 그리고 350,907개로 나타났다. 이중 판례 문서에는 사용되지만 Wiki 문서에는 나타나지 않는 판례 문서 고유의 단어의 수는 440개로 파악되었다. 이러한 판례 고유 단어는 대부분 일상 생활에서는 흔히 사용되지 않는 전문 법률용어로, <표 3>은 440개의 단어 중 개별 문서 중 출현 빈도가 가장 높은 상위 9개의 단어를 보인다.

이후 순수 전문 임베딩, 부분 혼합 임베딩 및 혼합 임베딩을 개별적으로 실시한 결과 3개의 독립된 임베딩 결과를 얻게 되었다. 이 때, 3개의 임베딩 결과 각각에 대해 위의 440개 Law Terms의 유의어를 추출하였다. 동일한 단어라도 각 임베딩의 결과로부터 도출된 유의어는 상이하게 나타날 수 있으며, 일례로 Law Terms 중 가장 출현 빈도가 높은 “불허가”의 유의어는 <표 4>와 같이 나타난다.

4.2 전문어와 일반어 간 유의어 비교

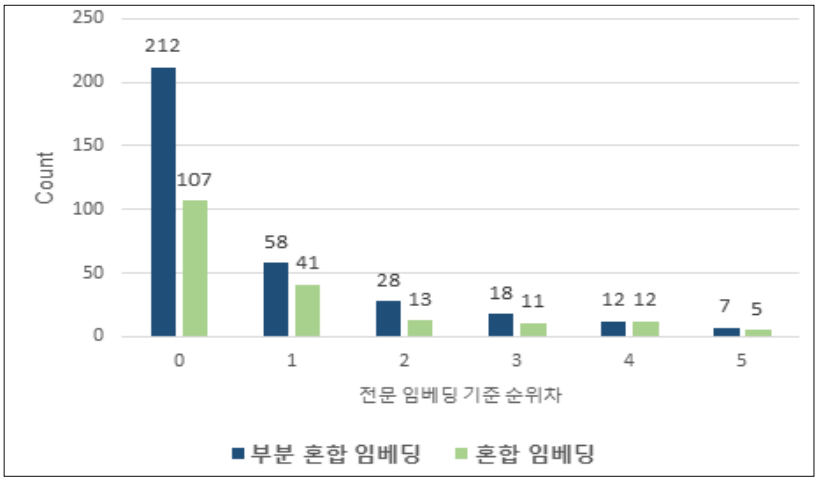
앞의 절에서 살펴본 바와 같이, 440개 전문 용어의 유의어는 각 임베딩에 따라 서로 다르게 나타나게 된다. 이 때 전문 용어가 해당 분야에서 갖는 고유의 의미는 해당 분야의 문서들만을 대상으로 한 임베딩인 순수 전문 임베딩에서 가장 정확하게 파악된 것으로 간주될 수 있다. 따라서 본 실험에서는 순수 전문 임베딩에서 나타난 전문 용어의 최유사 단어가 부분 혼합 임베딩 및 혼합 임베딩을 통한 결과에서 어느 정도 순위의 유의어로 파악되는지 측정함으로써 각 임베딩 방법의 전문어 의미 존속성을 평가하고자 한다. 예를 들어 <표 4>에서 순수 전문 임베딩을 통한 “불허가”의 최유사 단어는 “허가”이며, “허가”는 부분 혼합 임베딩에서는 2위를, 혼합 임베딩에서는 5위를 차지하였다. 즉 최유사 단어 기준으로 판단하면 “불허가”의 전문 임베딩에서의 의미는 혼합 임베딩에서보다 부분 혼합 임베딩에서 더욱 잘 유지된 것으로 해석할 수 있다.

이와 동일한 분석을 440개 단어 전체에 대해 수행한 결과가 <표 5>에 요약되어 있다. <표 5>에서 ‘최유사 단어 존속 수’는 440개 단어 중 순수 전문 임베딩에서의 최유사 단어가 부분 혼합 임베딩 또는 혼합 임베딩 후에도 여전히 최유사 단어로 남아 있는 단어의 수를 의미한다. 또한 ‘1,000위 이상 변동 수’는 440개 단어 중 순수 전문 임베딩에서의 최유사 단어가 부분 혼합 임베딩 또는 혼합 임베딩 후에 유사도 순위 1,000위 내에 포함되지 못한 단어의 수를 의미한다. 따라서 ‘최유사 단어 존속 비율’이 높고 ‘1,000위 이상 변동 비율’이 낮을수록 바람직한 임베딩, 즉 전문 용어가 해당 분야에서 갖는 고유의 의미를 더욱 잘 유지한 임베딩이라고 볼 수 있다.

<표 5>에서 부분 혼합 임베딩의 ‘최유사 단어 존속 수’는 212로, 전체 단어 440개 중 약 48%에 해당하는 단어들의 최유사 단어가 부분 혼합 임베딩을 통해서도 최유사 단어로 유지되었음을 알 수 있다. 이는 혼합 임베딩에서의 약 24%에 비해 약 24%p 높은 결과이다. 이와 반대로 종전 최유사 단어가 혼합 임베딩 후 유사도 1,000위 밖으로 순위가 대폭 하락하는 경우는 부분 혼합 임베딩의 경우 약 1.6%에 불과하였으나, 혼합 임베딩의 경우에는 10%에 해당하는 것으로 나타났다. 여기서 440개 단어 쌍의 개수는 모수(Population)이지만, 표본 추출로 파악하더라도 대부분의 통계학 서적(Gravetter and Wallnau, 2017)에서 원용되는 $np > 10$ 과 $n(1-p) > 10$인 이항 확률 분포의 정규 분포로의 근사요건(여기서 n은 표본의 수, p는 최유사 단어 존속 비율)을 동시 충족하므로, 이들 간의 비교는 충분히 의미 있는 것으로 평가할 수 있을 것이다(부분 혼합 임베딩의 경우 각 212와 251.7 및 혼합 임베딩의 경우 각 106.9 및 333.1). 각 임베딩에서 나타난 혼합 비율의 증가에 따른 유의어 순위의 변화 정도를 도표로 나타내면 아래의 <그림 6>과 같다.

<표 5> 최유사 단어의 변동 비율

	최유사 단어 존속 수	최유사 단어 존속 비율	1,000위 이상 변동 수	1,000위 이상 변동 비율
부분 혼합 임베딩	212	0.482	7	0.016
혼합 임베딩	107	0.243	44	0.1



<그림 6> 전문어/일반어 간 유의어 순위 변화

위 실험에 따른 결과는 아래와 같이 정리할 수 있다. 최유사 단어의 보존 정도로써 평가되는 판례 문서의 특수성이 위키 문서와의 혼합 임베딩의 과정에서 약해지고 있으며, 그 정도는 위키 문서의 크기 증가와 관련성이 있음이 확인된다. 구체적으로 위키 문서의 비율을 증가시켰을 때 순수 전문 임베딩에서 나타났던 최유사 단어들의 관계가 부분 혼합 임베딩과 혼합 임베딩 이후에도 그대로 유지되는 비율이 감소하는 것은, 순수 전문 임베딩에서 나타나는 전문용어의 의미적 고유성의 약화로 이해될 수 있다. 한편 각 혼합 임베딩에 활용된 위키 문서의 크기가 혼합 임베딩이 부분 혼합 임베딩에 비해 문서수 기준 10배, 고유 단어 기준 대략 6.2배 크다는 점을 감안할 때, 고유성의 약화 정도는 일반 문서의 크기 증가 정도에 반드시 비례하지는 않는다는 점도 확인할 수 있다. 또한 순수 전문 임베딩에 투입된 전문 문서의 양에 비해 부분 혼합 임베딩 및 혼합 임베딩의 각 임베딩에서의 일반 문서가, 문서 수 기준 각 약 10배 및 70배, 고유 단어 기준 각 6.2배 및 49.2배의 크기임을 감안할 때, 압도적인 전문 문서와 일반 문서의 크기 차이에도 불구하고 전문 문서의 의미적 고유성이 혼합 임베딩 후에도 일정 부분 존속되는 것으로 이해할 수 있다.

4.3 전문어 간 유의어 비교

본 절에서는 전문 문서에만 나타나는 Law Terms만을 대상으로, 이들 단어 사이의 유사도의 순위가 혼합비의 변동에 의해 어떤 영향을 받는지 측정한 결과를 정리한다. 유사도는 코사인 유사도 기준을 적용하였으며, 각 단어에 대해 자신과의 유사도는 1로 나타난다. 전체 440개의 Law Terms에 대해 자기 자신을 제외한 단어 간 나타날 수 있는 유사도의 순위는 1위에서 439위 사이에 분포한다. 아래 <표 6> ~ <표 8>은 각 임베딩을 통해 '불허가', '지료', '보관증', 그리고 '불능미수'의 4개의 전문어의 상호 코사인 유사도가 상이하게 나타남을 보인다.

<표 6> 순수 전문 임베딩(전문 100%)

	불허가	지료	보관증	불능미수
불허가	1	0.0822	0.0461	0.0099
지료	0.0822	1	0.1118	0.0969
보관증	0.0461	0.1118	1	0.0655
불능미수	0.0099	0.0969	0.0655	1

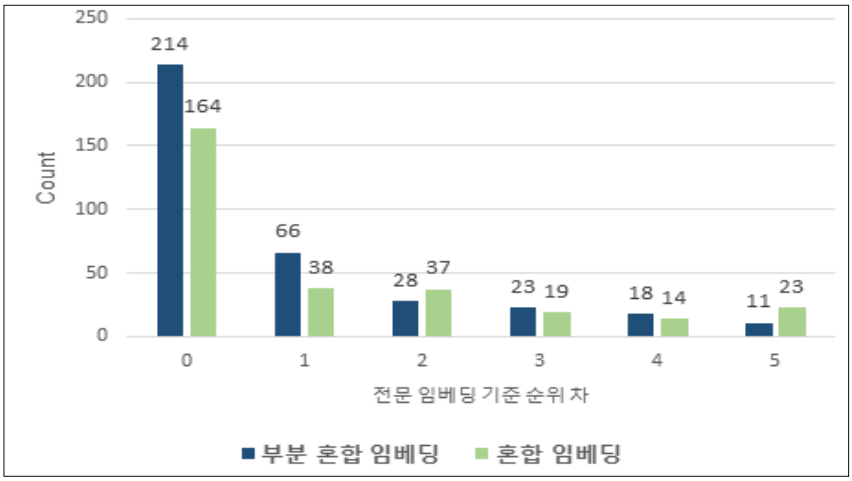
<표 7> 부분 혼합 임베딩 (전문 10% + 일반 90%)

	불허가	지료	보관증	불능미수
불허가	1	0.2965	0.1564	0.1222
지료	0.2965	1	0.1990	0.2099
보관증	0.1564	0.1990	1	0.1515
불능미수	0.1222	0.2099	0.1515	1

<표 8> 혼합 임베딩 (전문 1.4% + 일반 98.6%)

	불허가	지료	보관증	불능미수
불허가	1	0.2294	0.1957	0.1783
지료	0.2294	1	0.3464	0.2087
보관증	0.1957	0.3464	1	0.1573
불능미수	0.1783	0.2087	0.1573	1

위의 표는 4개의 단어에 대해서만 유사도 변화를 요약한 것이며, 전체 440개 단어 간 유의어 순위를 분석한 결과는 <그림 7>과 같다. <그림 7>은 Law Terms에 있어 순수 전문 임베딩에서 가장 높은 유사도 순위를 보이던 단어가 부분 혼합 임베딩 및 혼합 임베딩에서 어느 정도의 순위로 나타나는지를 보인다. 순위차가 0인 경우, 즉 혼합 순수 전문 임베딩에서 특정 단어와 가장 높은 유사도를 보였던 단어가 여전히 가장 높은 순위로 나타나는 경우는 부분 혼합 임베딩의 경우 214건(48.7%) 이었으나, 혼합 임베딩의 경우에는 164건(37.4%)으로 그 비율이 감소하였음을 확인할 수 있다. 또 한 최유사 단어의 순위가 한 계단 낮은 순위로 변동한 경우는 각 66건과 38건으로, 부분 혼합 임베딩의 경우 순위가 그대로이거나 한 단계만 감소한 경우는 약 68%에 해당하는 반면 혼합 임베딩의 경우 약 46%에 그치는 것을 알 수 있다.



<그림 7> 전문어 간 최유사 단어의 순위 변화 분석

<표 9> 전문어 일반어 간 유의어 분석 VS 전문어 간 유의어 분석

	전문어 일반어 간 유의어 분석		전문어 간 유의어 분석	
	최유사 단어 존속 수	최유사 단어 존속 비율	최유사 단어 존속 수	최유사 단어 존속 비율
부분 혼합 임베딩	212	0.482	214	0.486
혼합 임베딩	107	0.243	164	0.373

<표 9>는 전문어와 일반어 간 유의어 분석, 그리고 전문어 간 유의어 분석의 결과를 요약하여 나타낸다. 순수 전문 문서 임베딩에서 형성되는 Law Terms 사이의 유사도 관계가 일반 문서의 혼합 비율의 증가에 따라 약해지고 있음을 알 수 있으며, 결국 이는 일반 문서의 크기 증가와 고유성의 완화 정도가 정의 상관성을 가지게 된다는 관찰 결과를 뒷받침하는 것으로 해석할 수 있다.

Ⅴ. 결 론

최근 자연어 처리의 기술 발전은 많은 문제를 새로운 방식으로 해결하고 있다. 그럼에도 불구하고 특정 도메인의 특성이 충분히 반영된 자연어 처리 모델을 형성하여 전문 분야의 문제 해결에 적용하는 작업은 일부를 제외하고는 초기 단계에 머무르고 있다. 이에 본 연구에서는 먼저 법률 분야의 고유한 특성이 반영된 전문 문서 임베딩 및 소량의 전문 문서와 다량의 일반 문서가 함께 학습되는 혼합 임베딩을 시도하였으며, 이때 전문 분야의 고유한 언어적 의미가 학습에 어느 정도 반영되는지를 정량적으로 평가할 수 있는 평가 척도를 제시하였다.

본 연구는 우선 전문 문서로서 대법원 판례 문서, 일반 문서로서 한국어 위키백과를 선정하였다. 이후 전문 문서에서만 관찰되는 고유 단

어들의 최유사 단어쌍과 유사도를 추출한 다음, 그 값들이 혼합 임베딩의 과정에서 어떻게 변화되고 있는지를 관찰하였다. 본 연구에서 소개한 측정 방법에 따라, 전문 문서만의 임베딩에서 파악되었던 개별 단어들간 유사도의 관계가 함께 학습되는 일반 문서의 크기가 증가함에 따라 변화하는 현상을 관찰할 수 있었다. 이는 곧 전문 문서의 특수성이 함께 학습되는 일반 문서의 크기와 정의 상관관계를 맺으며 약화되는 것임을 확인할 수 있었으며 이는 본 연구의 학술적 기여로 인정받을 수 있다. 나아가 본 연구에서 채택한 성능 평가 방안이 개별 전문 분야 문서와 일반 문서 사이의 임베딩에서는 물론, 다양한 전문 분야 문서들이 결합되는 학제(cross-domain)적 연구에서의 임베딩에 있어 개별 도메인에서 사용되는 전문어의 특성 유지 수준을 측정할 수 있는 지표로서 역할을 할 수 있을 것이다. 또한 실무적인 측면에서는 우선 가까운 시일 내에 도입 예정인 자연어 처리 기반의 법률 정보 시스템 등에서 시민과 법률 전문가 집단 간 사용 용어의 간극을 다소나마 메우는데 활용될 수 있을 것으로 기대한다.

본 연구의 한계 및 향후 연구 방향은 다음과 같다. 우선 본 연구는 전문 코퍼스로서 법률 분야 문서만을 대상으로 하였으나, 보다 다양한 도메인 코퍼스의 확보를 통한 추가 실험이 필요하다. 또한 본 실험에서는 혼합 임베딩 전후의 전문성 유지 수준을 측정하기 위해, 코사인 유사도에 의한 단어 간 유사도를 측정한 후 각 단어들의 최우선 유사도의 유지 횟수와 비율을 사용하였다. 하지만 이는 직관적인 설명이 가능하다는 측면에서는 바람직한 방법으로 인정받을 수 있지만, 다양한 수치 정규화 방법 및 유사도 기준에 따라 그 결과가 달라질 수 있다는 한계를 가진다. 따라서 향후 연구에서는 유사도 측정의 방안을 포함한 수치 해석에 대해 더욱 정교한 분석이 이루어질 필요가 있다. 끝으로 본 연구는 일반 문서의 크기를 두 가지로 나누어 전문 문서와의 혼합 임베딩 수행 결과를 평가하였다. 그러나 코퍼스의 크기와 고유성의 약화 정도를 보다 세밀하게 관찰하기 위해서는 전문 문서와 일반 문서의 혼합비를 보다 다양하게 조정하여 관찰할 필요가 있다. 최근 학습 기술의 발전으로 학습 도중 학습 결과를 일부 동결하는 방식은 물론, 투입 배치(batch)의 크기를 증가시키면서 학습의 성과를 모니터링하는 방안들도 제시되고 있다. 향후 연구에서는 이러한 방식을 활용하여 보다 다양한 실험을 통해 더욱 정교한 분석을 수행할 필요가 있다.

참고문헌

강현화, “전문용어 연구의 맹점,” 나라사랑, 제125집, 2016, pp. 191-215.

김나리, 김형중, “연관법령 검색을 위한 워드 임베딩 기반 Law2Vec 모형 연구,” 한국디지털콘텐츠학회 논문지, 제18권, 제7호, 2017, pp. 1,419-1,425.

김한샘, “말뭉치 기반 한국어 연구의 현황과 전망,” 한국어학회 한국어학, 제83권, 2019, pp. 1-33.

오선영, “코퍼스와 영어교육,” 외국어 교육연구 제7집, 2004, pp 1-38.

윤상훈, 김근형, “Word2Vec를 이용한 토픽모델링의 확장 및 분석사례,” 한국정보시

스템학회 정보시스템연구, 제30권, 제1호, 2021, pp. 45-64

이기창, 한국어 임베딩, 도서출판 에이콘, 2019. pp. 80

최병설, 김남규, “감정 딥러닝 필터를 활용한 토픽 모델링 방법론,” 한국정보시스템학회 정보시스템연구, 제28권, 제4호, 2019, pp. 271-291

최순영, Matteson, A. S., 임희석, “한국어-영어 법률 말뭉치의 로컬 이중 언어 임베딩,” 한국융합학회논문지, 제9권, 제10호, 2018, pp. 45-53.

현암사 법전부, 법률용어 사전, 현암사, 2019.

Adewumi, T. P., Liwicki, F., Liwicki, M., “Word2Vec: Optimal Hyper-Parameters and their Impact on NLP Downstream Tasks”, arXiv:2003.11645, Mar 2020.

Ashley, K. D., Artificial Intelligence and Legal Analytics, Cambridge University Press, Cambridge, UK, 2017.

Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Pochhi, “A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments,” In Proceedings of *ECIR*, Springer, pp. 413-428, April 2019.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C., “A Neural Probabilistic Language Model,” *The Journal of Machine Learning Research*, 2003, pp. 1137-1155.

Cardellino, C., Teruel, M., Alonso L, Villata, S., “Legal NERC with Ontologies, Wikipedia and Curriculum Learning,” In Proceedings of *EACL*, pp.254-259, 2017.

Chen, H., Cai, D., Dai, W., Dai, Z., Ding, Y., “Charge-Based Prison Term Prediction with Deep Gating Network,” arXiv: 1908.11521v1, Aug 2019.

Church, K. W., “Word2Vec,” Natural Language Engineering, Cambridge University Press, Cambridge UK, Dec 2016.

Devlin, J., Chang, M., Lee, K., and Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv: 1810.04805, Oct 2018.

Doerr, M., “Semantic Problems of Thesaurus Mapping,” *Journal of Digital Information*, Vol.1 No 8, 2001.

Faruqui, M., Tsvetkov, Y., Rastogi, P., Dyer, C., “Problems With Evaluation of Word Embeddings Using Word Similarity Tasks,” arXiv:1605.02276, May 2016.

Garain, A,. Mahata, S., K., Dutta S., “Normalyzing Numeronyms - A NLP approach,” arXiv:1907.13356, Jul 2019.

Garneau, N., Leboeuf, J., and Lamontagne, L., “Predicting and Interpreting Embeddings for Out-of-vocabulary Words in Downstream Tasks,” arXiv:1903.00724, Mar 2019.

Golitsyna, O. L., Maksimov, N. V., and Fedorova V. A., “On Determining Semantic Similarity Based on

Relationships of a Combined Thesaurus," *Automatic Documentation and Mathematical Linguistics*, Vol 50, pp. 139-153, 2016.

Gravetter, F., and Wallnau, L., Statistics for Behavior Sciences, Cengage Learning, US, 2017.

Islam, A., and Inkpen, D., "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity," ACM Transactions on Knowledge Discovery from Data, Article No. 10, Jul 2008.

Jurafsky, D., and Martin, J., Speech and Language Processing, Pearson Education, Upper Saddle River, New Jersey, 2009.

Kim, M., Goebel, R., "Two-step Cascaded Textual Entailment for Legal Bar Exam Question Answering," *ICAIL* pp. 283-290, 2017.

Leech, G., and Svartvik, J., "Corpora and Theories of Linguistic Performance," Directions in Corpus Linguistics, Corpora and Theories of lLnguistic Performance, Walter de Gruyter, Berlin, 1992.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., "Distributed Representations of Words and Phrases and Their Compositionality," arXiv: 1310.4546, Oct 2013.

Mooers, C., "The Theory of Digital Handling of Non-Numerical Information and its Implications to Machine Economics," in Proceedings of The Meeting of *The Association for Computing Machinery* at Rutgers University, Mar 1950.

Pennington, J., Socher, R., and Manning, C., "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 *Conference on Empirical Methods in Natural Language Processing*, Oct 2014, pp. 1532-1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., "Deep Contextualized Word Representations," arXiv: 1802. 05365, Feb 2018.

Pilehvar, M., Camacho-Collados, J., Navigli, R., and Collier, N., "Towards a Seamless Integration of Word Senses into Downstream NLP Applications," in Proceedings of the 55th Annual Meeting of the *Association for Computational Linguistics*, Volume 1: Long Papers, Vancouver, Canada, 2017, pp. 1857-186.

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

대법원 보도자료 “전국법원장 회의”, 2020. 12. : https://www.scourt.go.kr/portal/news/News ViewAction.work?seqnum=1931

대법원 사법연감(통계) 2019 : https://www.scourt.go.kr/portal/justicesta/JusticestaListAction.work?gubun=10

대한민국 법원 종합법률정보 : https://glaw.scourt.go.kr/

모두의 말뭉치, 국립 국어원, 2020 : https://corpus.korean.go.kr/

한국어 위키백과 : https://ko.wikipedia.org/wiki/위키백과

British Legal Report Corpus : https://www.sketchengine.eu/blarc-british-law-reference-corpus

Brown Corpus : http://icame.uib.no/brown/bcm.html

Google Books Ngram Corpus : http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

GPT-3 (openAI) : https://openai.com/blog/openai-api/

Mecab : https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/

Medical Corpus: English Corpus From Web: https://www.sketchengine.eu/medical-web-corpus/

Oxford English Corpus : https://languages.oup.com/research/

wordNet (Princeton Univ) : https://wordnet.princeton.edu/

김 병 태 (Kim, Byung Tae)

현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이며, 서울서부지방법원에 재직 중이다. 주요 관심 분야는 Text Mining, Data Modeling 등이다.

김 남 규 (Kim, Nam Gyu)

현재 국민대학교 비즈니스IT전문대학원장 및 경영정보학부 교수로 재직 중이다. 서울대학교 컴퓨터 공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사 학위를 취득하였다. 한국지능정보시스템 학회 부회장, 한국정보기술응용학회 부회장, 한국 경영학회 상임이사, 한국경영정보학회 이사, 한국인터넷 정보학회 이사를 역임하였다. 주요 관심 분야는 Text Mining 및 Data Mining, Deep Learning, Data Modeling 등이다.

<Abstract>

A Method for Learning the Specialized Meaning of Terminology through Mixed Word Embedding

Kim, Byung Tae · Kim, Nam Gyu

Purpose

In this study, first, we try to make embedding results that reflect the characteristics of both professional and general documents. In addition, when disparate documents are put together as learning materials for natural language processing, we try to propose a method that can measure the degree of reflection of the characteristics of individual domains in a quantitative way.

Approach

For this study, the Korean Supreme Court Precedent documents and Korean Wikipedia are selected as specialized documents and general documents respectively. After extracting the most similar word pairs and similarities of unique words observed only in the specialized documents, we observed how those values were changed in the process of embedding with general documents.

Findings

According to the measurement methods proposed in this study, it was confirmed that the degree of specificity of specialized documents was relaxed in the process of combining with general documents, and that the degree of dissolution could have a positive correlation with the size of general documents.

Keyword: Corpus, Legal Document, Word Embedding, Word2Vec, Word Similarity, Evaluation for Mixed Embedding

* 이 논문은 2021년 4월 21일 접수, 2021년 5월 11일 1차 심사, 2021년 5월 27일 2차 심사, 2021년 5월 27일 게재 확정되었습니다.