Journal of the Korea Convergence Society
Vol. 9. No. 10, pp. 45-53, 2018

ISSN 2233-4890
https://doi.org/10.15207/JKCS.2018.9.10.045

한국어-영어 법률 말뭉치의 로컬 이중 언어 임베딩

최순영[1], Andrew Stuart Matteson[1], 임희석[2*]
[1]고려대학교 컴퓨터학과 석사과정
[2]고려대학교 컴퓨터학과 교수

Utilizing Local Bilingual Embeddings on Korean-English Law Data

Soon-Young Choi[1], Andrew Stuart Matteson[1], Heui-Seok Lim[2*]
[1]Student, Dept. of Computer Science and Engineering, Korea University
[2]Professor, Dept. of Computer Science and Engineering, Korea University

요 약 최근 이중 언어 임베딩(bilingual word embedding) 관련 연구들이 각광을 받고 있다. 그러나 한국어와 특정 언어로 구성된 병렬(parallel-aligned) 말뭉치로 이중 언어 워드 임베딩을 하는 연구는 질이 높은 많은 양의 말뭉치를 구하기 어려우므로 활발히 이루어지지 않고 있다. 특히, 특정 영역에 사용할 수 있는 로컬 이중 언어 워드 임베딩(local bilingual word embedding)의 경우는 상대적으로 더 희소하다. 또한 이중 언어 워드 임베딩을 하는 경우 번역 쌍이 단어의 개수에서 일대일 대응을 이루지 못하는 경우가 많다. 본 논문에서는 로컬 워드 임베딩을 위해 한국어-영어로 구성된 한국 법률 단락 868,163개를 크롤링(crawling)하여 임베딩을 하였고 3가지 연결 전략을 제안하였다. 본 전략은 앞서 언급한 불규칙적 대응 문제를 해결하고 단락 정렬 말뭉치에서 번역 쌍의 질을 향상시켰으며 베이스라인인 글로벌 워드 임베딩(global bilingual word embedding)과 비교하였을 때 2배의 성능을 확인하였다.

주제어 : 이중 언어 워드 임베딩, 자연어처리, 영역 특수적, 법률 영역, 단어집, 반지도 학습, 단락 정렬, 단어 유사도, skip-gram, 로컬 임베딩

Abstract Recently, studies about bilingual word embedding have been gaining much attention. However, bilingual word embedding with Korean is not actively pursued due to the difficulty in obtaining a sizable, high quality corpus. Local embeddings that can be applied to specific domains are relatively rare. Additionally, multi-word vocabulary is problematic due to the lack of one-to-one word-level correspondence in translation pairs. In this paper, we crawl 868,163 paragraphs from a Korean-English law corpus and propose three mapping strategies for word embedding. These strategies address the aforementioned issues including multi-word translation and improve translation pair quality on paragraph-aligned data. We demonstrate a twofold increase in translation pair quality compared to the global bilingual word embedding baseline.

Key Words : Bilingual word embedding, natural language processing, domain-specific, law domain, dictionary seed, semi-supervised training, paragraph-aligned, word similarity, skip-gram, local embedding

1. 서론

CBOW(Continuous Bag of Words)모델과 네거티브 샘플링(Negative Sampling)을 사용하는 skip-gram (SGNS)[1] 모델의 등장으로 단어 표현(Word Representation) 모델이 단순하고 빠르게 학습될 수 있게 되었으며 자연 언어 처리 분야에서는 다층 신경망(Multi-layered Neural Network)을 사용하여 워드 임베딩(Word Embedding)을

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017M3C4A7068189).
*Corresponding Author : Heui-Seok Lim (limhseok@korea.ac.kr)
Received August 7, 2018 Revised September 27, 2018
Accepted October 20, 2018 Published October 28, 2018

학습시키는 것이 표준이 되었다.[2-3]

대부분의 워드 임베딩 연구는 단일 언어에 대해서 문법적으로 혹은 의미적으로 유사한 단어를 임베딩 공간(Embedding space)에서 가깝게 표현하도록 하는 단어의 표현을 학습하는 데 초점을 맞추었었다[4]. 최근 들어서는 다중 언어 워드 임베딩(Multilingual Word Embedding)에 관한 관심도 많아졌다[5]. 그리고 다중 언어 워드 임베딩은 기계번역(Machine Translation), 품사 태깅(POS tagging), 감정분석(Sentiment Analysis) 등 자연어 처리 분야에서 폭 넓게 사용될 수 있다. [6]에서는 형태소 단위 자질을 표현하기 위해 워드 임베딩을 사용하였고, [7]에서는 아마존 패션 상품 리뷰 데이터를 이용하여 각 단어들의 의미론적 특성을 반영하기 위해 워드 임베딩을 사용하였다.

다중 언어 임베딩 중 하나인 이중 언어 워드 임베딩에서는 서로 다른 언어 말뭉치(Corpus)를 이용하여 모델을 학습시키게 된다. 학습에 사용되는 말뭉치는 정렬 기준에 따라서 나뉘는데, 이 중 어떤 정렬 기준을 사용한 말뭉치를 밑바탕 어휘집(seed lexicon)으로 사용하는지에 따라 서로 다른 언어 간의 단어 번역 쌍 연결(mapping) 강도가 결정되어 임베딩의 질에 영향을 미치게 된다[8]. 단어 번역 쌍의 연결은 단어 정렬, 문장 정렬, 단락 정렬, 순으로 강도가 낮아진다. 하지만 단어 기준 정렬 코퍼스와 문장 기준 정렬 말뭉치의 경우, 많은 양의 질 좋은 데이터를 얻는 데 한계가 있고, 언어 번역 쌍에 따라서도 얻기가 어려워진다.

한국어와 특정 언어를 사용한 이중 언어 워드 임베딩의 경우 특히, 병렬 데이터로 이중 언어 워드 임베딩을 학습하는 데에 있어서는 병렬 데이터가 양에서도 한정적이고, 질이 좋은 데이터는 더욱 희소하다. 위와 같은 이유로 한국어를 사용한 이중 언어 워드 임베딩 연구는 활발히 이루어지지 못하고 있다.

또한 그 중에서도 영역 특수적인 목적으로 자연어 처리 관련 분야에서 사용될 수 있는 로컬 이중 언어 워드 임베딩(local bilingual word embedding)의 경우 연구는 더 희소하다. [9]에서는 로컬 워드 임베딩(local word embedding)이 정보 검색(information retrieval)에서 문서를 검색하는 데 쿼리 확장(query expansion)을 더 효과적으로 할 수 있음을 보였고, 따라서 임베딩을 하는 데에도 로컬 워드 임베딩을 고려하는 것이 중요하다는 것을 제안하였다.

본 연구에서는 단락 기준으로 정렬(paragraph-aligned)된 한국 법률 데이터의 한국어-영어 말뭉치(corpus)와 작은 단어집(dictionary seed)을 사용하여 한국어-영어 이중 언어 워드 임베딩(bilingual word embedding)을 반지도(semi -supervised)로 학습시키는 방법을 제안한다. 해당 문서는 문장 기준으로 정렬된 말뭉치가 아닌 단락 기준으로 정렬된 말뭉치이므로 첫째, 유사도(similarity)가 높은 서로 다른 언어의 두 단어를 효율적으로 연결(mapping)하기 위해서 둘째, 한 단어로 이루어진 한국어 어휘가 여러 개의 단어로 이루어진 영어 어휘와 연결되는 문제를 해결하기 위해서 3가지의 연결 전략을 제안한다. 그리고 이 3가지 연결 전략을 검증하기 위해서 학습에 사용된 한국어-영어로 표현된 법률 말뭉치를 바탕으로 150쌍의 단일 단어로 이루어진 검증 사전과 50쌍의 다중 단어(multi-word)로 이루어진 검증 사전을 구축하고, 이 사전을 기준으로 전략 간의 정확도를 비교한다. 또한 위키피디아 글로벌 이중 언어 워드 임베딩을 베이스라인으로 사용하여 본 논문의 전략과 정확도 비교한다.

본 연구의 기여는 첫째, 단락 정렬 말뭉치를 사용한 워드 임베딩과 같은 문장 정렬 말뭉치를 사용한 워드 임베딩 보다 상대적으로 임베딩의 질이 떨어질 수 있는 말뭉치에 대해 번역 쌍 연결의 질을 향상시키기 위해 다양한 매핑전략을 적용하였다. 둘째, 한국어-영어 번역 쌍을 연결시키는 과정에서 발생하는 하나의 원시 어휘(source language vocabulary)에 여러 개의 단어로 이루어진 목적 어휘(target language vocabulary)가 연결되는 문제를 해결하기 위해 3가지의 전략을 제안하고 각 전략들의 결과를 비교분석하였다. 셋째, 법률 데이터를 사용하여 로컬 워드 임베딩을 학습시킨 결과, 영역 특수적인 자연어 처리 관련 분야에서 위키피디아 데이터를 사용하여 글로벌 워드 임베딩을 학습시킨 이중 언어 임베딩보다 2배의 정확도를 얻음을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 소개하며, 3장에서는 본 연구에서 제안하는 3가지 연결 전략에 대해서 설명하며, 4장에서는 3가지 연결 전략을 실험에 적용하고, 5장에서는 결론 및 향후 연구로 구성된다.

2. 관련 연구

2.1 Word Embedding

Word2vec은 Google의 Tomas Mikolov와 그 팀원들이 제안한 워드 임베딩 기술(technique)이며[10], 아주 큰 데이터 셋으로부터 워드를 연속된 벡터 표현(continuous vector representation)으로 계산해내는 모델이다[11].

Word2Vec은 일반적으로 skip-gram(SGNS) 또는 continuous bag of words(CBOW), 두 가지 방법 중 하나를 통해 워드를 연속된 벡터표현으로 나타낸다. 두 가지 방법 중 하나인 SGNS는 입력으로 단어가 주어졌을 때 특정한 윈도우(window) 크기 w내의 문맥을 예측하는 모델이며, 또 다른 방법인 CBOW는 반대로 문맥이 입력으로 주어지면 그 문맥을 사용하여 단어를 예측하는 모델이다.

GloVe는 2014년에 미국 스탠포드 대학에서 개발된 워드 임베딩 방법론이며, LSA와 Word2Vec의 단점을 극복하고자 고안된 모델이다. GloVe는 단어 벡터 사이의 유사도 측정을 용이하게 하면서 말뭉치 전체의 통계 정보를 비교적 더 반영시킬 수 있는 방법이다. GloVe의 목적(objective)은 단어 벡터들의 내적이 전체 말뭉치의 단어의 동시등장확률(the word's probability of co-occurrence)에 로그를 취한 값이 되도록 학습하는 것이다. 그러나 단어의 개수만큼 큰 동시등장행렬을 만들고 그에 따른 matrix factorization을 수행해야하기 때문에 계산 복잡성이 커진다.

따라서 본 논문에서는 Word Embedding 기법중 Word2Vec을 사용하였으며, SGNS 모델이 상대적으로 소량의 데이터로 보다 안정적으로 작동하고 희소한 단어를 보다 효과적으로 나타낼 수 있기 때문에[1], 법률 영역 로컬 이중 언어 워드 임베딩을 학습시키는 데 SGNS 모델을 적용한다. 따라서 본 논문에 적용된 skip-gram 모델은 인덱스 i 와 윈도우 크기 c 그리고, 중심 단어 w_i 가 주어졌을 때, 영역 특수적인 법률 문맥 단어인 w_j, $(i-c \leq i \leq i+c, j \neq i)$를 예측한다.

2.2 교차 언어 임베딩(Cross-lingual embedding)

이중 언어 워드 임베딩의 경우에는 네 가지의 기본 접근법이 있다[12]. 첫째, monolingual mapping은 먼저 원시 언어(source language)와 목적 언어(target language)의 워드 임베딩을 각각 학습시킨 후, 두 단어 표현 간의 선형 매핑을 학습하는 접근법이다. 이러한 접근법에는 projection 방법에 따라 linear projection 적용한 [13]와 CCA를 적용한 [14]가 있다. 둘째, pseudo cross-lingual은 서로 다른 언어의 문맥을 혼합한 말뭉치를 사용하여 워드 임베딩을 학습시키는 접근법이다. 본 접근법에는 문맥을 혼합시키는 방법에 따라서 [4],[15]가 있다. 셋째, cross-lingual training은 교차 언어 목적(cross-lingual objective) 함수 최적화에 집중하는 접근법이다. 본 접근법은 교차 언어 어휘집과 병렬 말뭉치보다 문장 정렬 말뭉치에 의존하는 방법이며 [16]와 같은 연구가 있다. 넷째, joint optimization은 이중 언어 제약뿐만 아니라 단일 언어(mono-lingual)와 이중 언어 간의 목적 함수를 공동 최적화(joint optimization)하는 접근법이며 [17]과 같은 연구가 있다.

앞서 언급한 네 가지 접근법 중 첫 번째 접근법인 전체 문장을 하나의 문장표현으로 인코딩하는 CCA 모델과 세 번째 접근법인 교차 언어 학습 방법 모델은 위키피디아와 같은 많은 양의 데이터가 필요하다는 단점을 가지고 있다. 또한 네 번째 접근법인 Joint optimization은 단일 언어와 이중 언어를 공동으로 목적 함수로 최적화시키기 때문에 이중 언어 임베딩의 질이 저하될 수 있다. 따라서 본 연구에서는 이중 언어 임베딩의 질을 최대화시킬 수 있고, 말뭉치의 크기가 작고 희박성(sparsity)이 높은 경우에도 잘 학습될 수 있는 pseudo cross-lingual 접근법을 사용하였다.

3. 이중 언어 워드 임베딩 모델

Fig. 1은 본 논문에서 제안하는 이중 언어 워드 임베딩 모델의 전체 구조를 나타낸다. 본 모델은 데이터 전처리, 제안된 모델의 학습, 실험 세 단계로 이루어져있다. 데이터 전처리 단계에서는 한국어-영어 법률 데이터에 대해서 어근화를 하고, 학습 단계에서는 전처리 후의 단락 정렬 법률 데이터에 대해서 제안된 알고리즘을 사용하여 이중 언어 문장을 생성한다. 실험 단계에서는 학습된 이중 언어 워드 임베딩을 바탕으로 번역 쌍의 성능을 측정한다.

본 연구에서는 이중 언어 워드 임베딩을 학습시키기 위한 영역 특수적(domain-specific)인 말뭉치 C를 한국어와 영어로 이루어진 단락 기준 정렬 한국 법률 말뭉치로 가정한다. 그리고 원시 언어(source language)를 한국어로 하고 목적 언어(target language)를 영어로 설정한 후, 한국어 단락을 입력으로 하여 [4]와 같은 방식으로 독립확률(independent likelihood) a를 0.5로 설정하고 a에

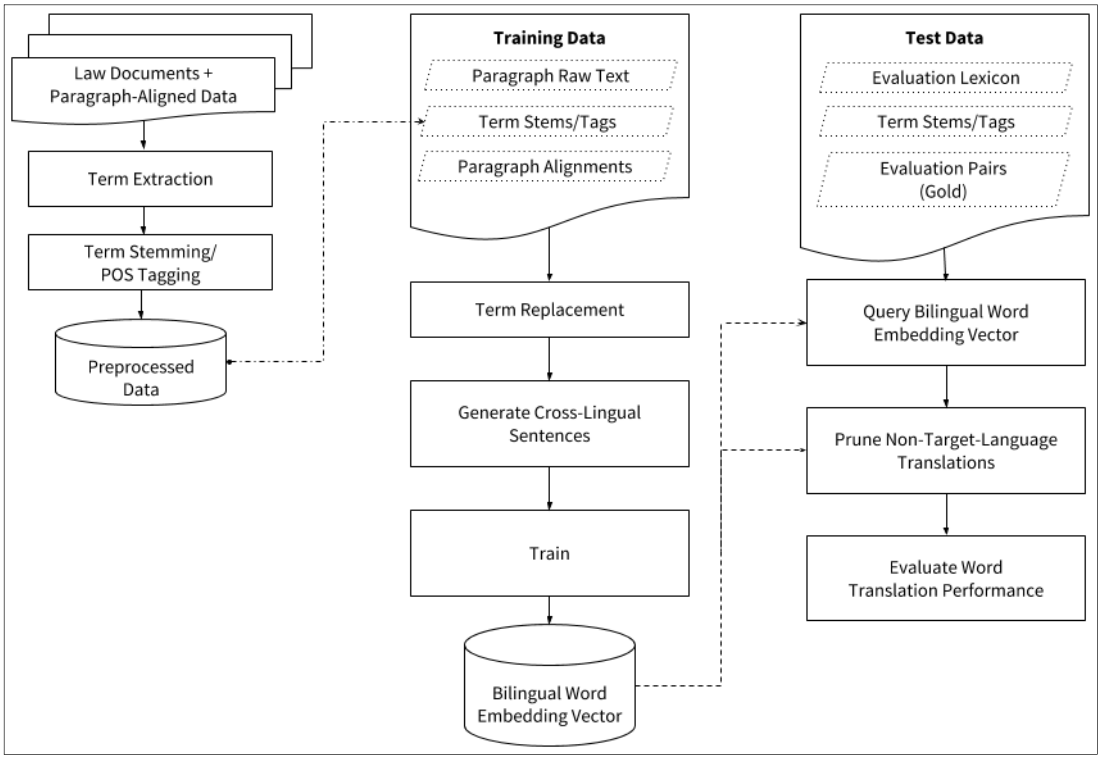


Fig. 1. Overview of local bilingual word embedding system

따라서 한국어 단락에 포함되어 있는 어휘를 어휘집에 존재하는 영어 대역어(equivalent)로 교체한다.

한국어와 영어는 계열이 크게 다르므로 많은 어휘들이 번역될 때 단어의 개수에서 일대일 대응을 이루지 못하는 경우가 많다. 예를 들면 일대다 ('대학원' : 'graduate school', '문화재' : 'cultural heritage')와 같이 불규칙적인 대응관계를 갖는 현상을 쉽게 볼 수 있다 [18]. 따라서 본 논문에서는 이러한 불규칙적 대응 문제를 해결하고, 단락 기준의 말뭉치의 워드 임베딩 질을 향상시키기 위해서 다음과 같은 3가지 연결 전략을 제안한다.

첫 번째로는 무작위 매치(Random Match) : **RM**, 두 번째로는 한 단어 교집합 매치(Single Match Greedy Intersect) : **SMGI**, 세 번째로는 복수 단어 교집합 매치(Multiple Match Greedy Intersect) : **MMGI**(목적 언어로 표현된 법률 말뭉치에 존재하는 단어 중 어휘집에 존재하는 목적 단어로 교차시킨다.)를 제안한다.

[Fig. 2.]는 본 논문에서 샘플 문장에 제안하는 3가지 전략에 따른 대역어 선택하는 과정에 대해 설명하고 있다. 본 논문에서는 단락 기준으로 정렬된 말뭉치에 본 전략을 적용하지만, 설명의 편의를 위해 [Fig. 2.]에서는 문장에 본 전략을 적용한 예를 설명하였다. [Fig. 2.]에는 한국어 문장, 영어로 번역된 문장, 사전이 나타나 있다. 한국어 문장과 영어로 번역된 문장 및 사전은 모두 전략에 적용할 때 어근 형태로 표현하여 사용하게 된다. Fig. 2.에서는 문장에 등장하는 어휘들 중 '국군'에 전략을 적용한 경우에 대해서 설명하고 있으며, 해당 어휘의 경우 번역 쌍이 일대일 대응이 아닌 일대다 대응을 이루는 경우이다.

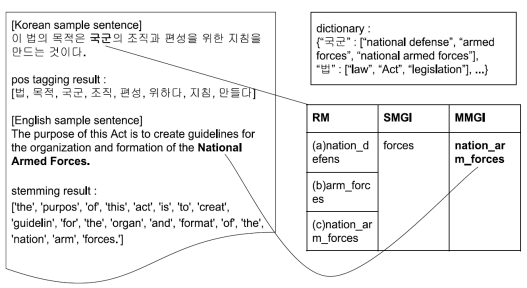


Fig. 2. Effect of each strategy on translation equivalence selection for sample sentence

3.1 Random Match

Random Match 전략에서는 원시 언어의 단락의 어휘를 어휘집에 존재하는 목적 언어의 대역어 중 무작위로 선택하여 교체한다. 해당 전략의 경우에는 학습에 사용되는 법률 말뭉치 중 목적 언어의 단락을 참고하지 않는다. 본 전략의 경우에서는 [Fig. 2.]에서 사전만을 참고하며, 오른쪽 하단의 표의 사전의 번역 결과를 어근화한 형태로 표현된 (a),(b),(c) 중 하나를 무작위로 선택하여 한국어 어휘를 0.5 확률로 교체한다.

3.2 Single Match Greedy Intersect

Single Match Greedy Intersect 전략에서는 원시 언어 단락의 어휘를 어휘집에서 검색한 후, 검색 결과인 목적 언어의 어휘들 중에서 목적 언어의 단락에 존재하는 어휘로 교체한다. 단, 해당 전략은 목적 언어의 어휘가 여러 개의 단어로 이루어진 경우 여러 개의 단어 중 가장 긴 단어 한 개를 선택하여 교체한다. 본 전략에서는 Fig. 2의 어근화한 형태로 표현된 영어 문장과 사전을 참고하여 한국어 어휘를 0.5 확률로 교체한다. Fig. 2의 '국군'은 영어로 번역된 문장을 참고하여 영어로 번역된 문장에 존재하는 'nation_arm_ forces'를 교체 대상으로 하고 nation, arm, forces 중 가장 긴 단어를 선택한다. Fig. 2에서는 nation과 forces의 길이가 같은데, 이 경우에는 어휘 안에서 먼저 등장한 단어를 우선순위로 하여 한국어 어휘의 교체 어휘로 선택한다.

3.3 Multiple Match Greedy Intersect

Multiple Match Greedy Intersect 전략은 두 번째 전략인 SMGI와 동일한 방법을 사용하여 원시 언어 단락의 어휘를 어휘집에서 검색한 후, 해당 검색 결과 중에서 목적 언어 단락에 존재하는 어휘를 교체 대상으로 삼는다. 원시 언어 단락의 한 단어로 이루어진 어휘가 목적 언어 단락의 여러 개로 이루어진 어휘와 대응될 경우, 해당 어휘를 교체 대상으로 한다. 본 전략에서는 SMGI와 동일하게 어근화한 형태로 표현된 영어 문장과 사전을 참고하여 한국어 어휘를 0.5 확률로 교체한다. 본 전략을 적용한 [Fig. 2.]를 보면 SMGI와 같이 'nation_arm_forces'를 교체 대상으로 하고, SMGI에서는 'nation'를 교체 대상으로 하였지만 본 전략의 경우에는 여러 개로 이루어진 어휘인 'nation_ arm_forces' 형태 그대로 '국군'을 교체할 대상으로 선택한다.

4. 실험

4.1 Dataset

11,655개의 한국어-영어로 정렬(parallel- aligned)된 법률 문서를 한국 법률 웹사이트에서 1948.8.30.부터 2018.7.1.까지의 기간에 대해서 크롤링(crawling)하였다. 그리고 해당 법률 문서를 전처리(preprocessing)하여 총 868,163개의 병렬(parallel- aligned) 단락을 이중 언어 워드 임베딩을 학습시킬 말뭉치로 하였다. 또한 한국어-영어 번역 쌍을 구축하는 데에는 국립국어원에서 제공하는 어휘집을 사용하였다. 국립국어원에서 제공하는 사전은 체언 및 용언으로 나뉘어져 있었으며, 한국어-영어 번역 쌍의 정확률을 향상시키기 위해 학습에 사용한 법률 말뭉치를 형태소 분석하여 국립국어원 사전을 나뉜 구조 그대로 적용할 수 있도록 하였다.

본 연구에서 사용된 한국 법률 문서는 국방, 이민, 식품관리, 문화유산 등을 포함한 다양한 주제로 구성되어 있는 말뭉치이다.

제안한 전략 3가지로 학습시킨 이중 언어 워드 임베딩으로부터 얻어진 번역 쌍의 정확도를 평가하기 위해서, 본 연구에서는 학습에 사용된 한국어 법률 데이터를 바탕으로 150개의 단일 단어로 구성된 검증 어휘집과 50개의 다중 단어로 구성된 검증 어휘집을 수작업으로 구축하여 사용하였다.

4.2 실험 방법

본 논문에서 제안된 이중 언어 워드 임베딩 학습 전략들을 평가하기 위해서 Acc@k 평가법을 사용하였다. Acc@k 평가법은 ground truth 집합 T_topK@d의 목적 언어 번역 결과 중 top k 안에 예측된 번역 결과가 나타나는지 안 나타나는지에 대한 여부를 측정한다. k는 매개변수로 1, 3, 5 값 중 하나의 값으로 설정된다. pseudo cross-lingual 접근법을 사용하기 때문에 목적 언어의 워드 벡터 쿼리할 때에 목적 언어가 아닌 원시 언어가 결과로 나타날 수 있다. 그러므로 top-k를 평가하기 위해서 워드 벡터의 쿼리 결과에서 목적 언어가 아닌 원시 언어로 나타난 결과는 제외시켰다.

또한 글로벌 워드 임베딩(global word embedding)과 본 논문의 로컬 워드 임베딩(local word embedding)의 차이를 알아보기 위해서 글로벌 워드 임베딩인 위키피디아(Wikipedia) 이중 언어 워드 임베딩을 베이스라인

(baseline)으로 사용하였다. 정확한 비교를 위해 해당 위키피디아 이중 언어 임베딩의 학습에도 본 논문과 같이 pseudo cross-lingual 접근법을 사용하였다.

법률 말뭉치에서 전반적으로 사용되는 형태학적으로 복잡한 한국어의 동사 어형의 다양성 때문에 이중 언어 연결강도(bilingual signal)를 늘리고 데이터의 희박성(sparsity)를 줄이기 위해서, 원시 언어와 목적 언어에 어근화(stemming) 및 품사부착(POS tagging)을 하였다. 주어진 원시 언어의 선택적인 교체를 하는 동안, 알파 파라미터 a는 0.5로 고정되어 원시 언어의 어휘 중 절반이 원시 언어의 각 어휘에 대응하는 대역어로 교체된다.

단어 d 및 ground truth 번역 $T_{@d}$, 그리고 이중 언어 워드 임베딩 모델로부터 예측된 번역 $t_{p_{\text{topk@d}}}$ 주어졌을 때, $Acc_{d@}k$ 은 다음과 같이 계산 된다:

$$Acc_{d@}k = \begin{cases} 1 & if \quad T_{@d} \in t_{p_{\text{topk@d}}} \\ 0 & else. \end{cases}$$

단어 번역 결과의 정확도 $Acc@k$는 본 논문에서 제안된 이중 언어 워드 벡터의 쿼리로부터 반환된 결과인 $Acc_{d@}k$의 합을 ground truth로부터 반환된 결과인 $Acc_{d@}k$의 합으로 나누어 계산한다. 모든 실험은 Ubuntu Linux 18.04.에서 Python 3로 수행되었다.

4.3 실험 결과 및 분석

Fig. 3은 본 논문에서 제안하는 3가지 전략을 사용하여 학습시킨 법률 데이터 로컬 이중 언어 임베딩과 베이스라인인(baseline) 위키피디아(Wikipedia) 글로벌 이중 언어 워드 임베딩에 대해서 단일 단어 번역 성능을 비교한 그래프이다.

성능 비교는 벡터 쿼리 결과에서 얻는 번역 결과에 대해 매개변수 k를 1,3,5로 조정하여 각 k개의 번역 결과에 정답이 포함되는지에 따라 측정된 결과로 이루어져 있다. 두 번째 전략은 단일 단어 연결에 최적화되어 있는 전략이기 때문에 위와 같은 단일 단어 번역 실험에서 가장 높은 성능을 얻음을 볼 수 있다. 또한 세 가지 전략 간의 성능 차이는 크지 않았지만 베이스라인인 위키피디아 이중 언어 워드 임베딩과의 차이는 약 2배 차이가 나는 것을 볼 수 있다. 이는 3가지 제안된 전략은 법률 데이터를 이용한 로컬 임베딩이지만 위키피디아 워드 임베딩의 경우에는 글로벌 임베딩이므로 이러한 영역 특수적인 목적의 데이터에서는 로컬 임베딩이 성능이 훨씬 높게 나타난다는 것을 의미한다.

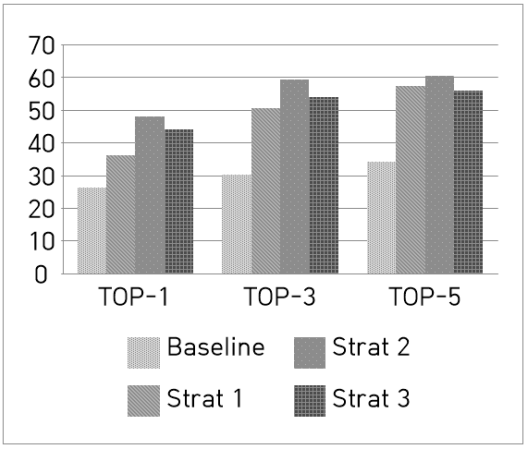


Fig. 3. Performance (Y-axis : %) of various strategies vs. baseline on the single word translation task

Fig. 4는 본 논문에서 제안하는 3가지 전략 중 첫 번째와 세 번째 전략으로 각각 학습시킨 법률 데이터 로컬 이중 언어 임베딩에 대해서 다중 단어 번역 성능을 비교한 그래프이다.

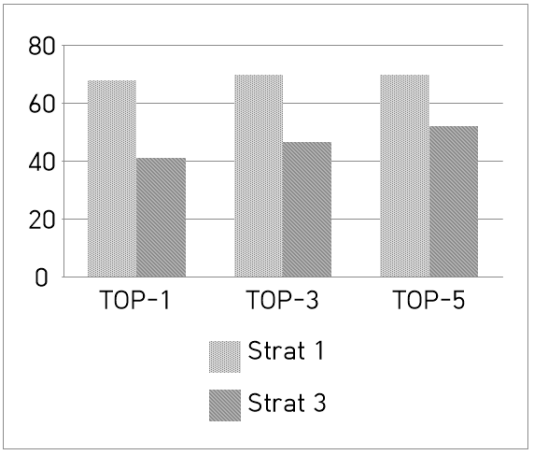


Fig. 4. Performance (Y-axis : %) of first and third strategies on the multiple word translation task

다중 단어 번역 결과 성능 비교의 경우에는 두 번째 전략이나 위키피디아 글로벌 이중 언어 임베딩은 단일 단어를 목적으로 학습된 임베딩이기 때문에 첫 번째와

세 번째 전략에 대해서만 수행되었다. 본 성능 비교의 경우에도 매개변수 k를 1,3,5로 조정하여 측정하였다. 세 번째 전략의 경우 목적 언어 말뭉치를 참조하여 학습된 결과이고 첫 번째 전략의 경우 목적

언어 말뭉치의 참조 없이 어휘집만을 사용하여 학습된 결과임에도 불구하고 첫 번째 전략이 성능이 더 높게 나오는 것을 볼 수 있다. 이는 본 연구에서 사용한 법률 데이터의 희박성(data sparsity)로 인한 결과로 예측된다.

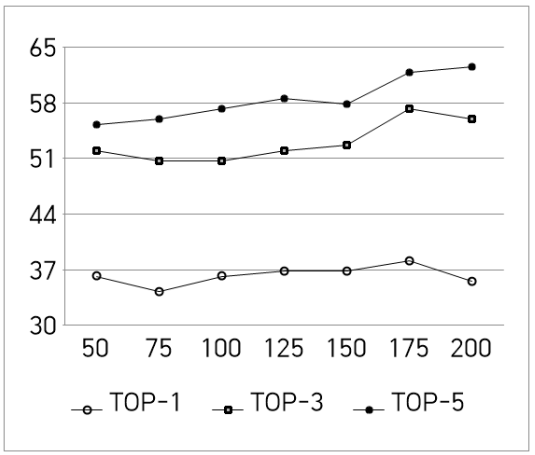


Fig. 5. Performance (Y-axis : %) of strategy 1 on the single word translation task when varying embedding dimension

[Fig. 5.]는 제안하는 3가지 전략 중 첫 번째 전략에 대해 임베딩 차원 수에 따른 단일 단어 번역 성능을 나타내는 그래프이다.

본 성능 분석 실험 또한 매개변수 k의 값을 1,3,5로 조정하여 이루어졌다. k값이 1인 경우에는 차원 수를 변화시키는 데 성능에 큰 변화가 없었지만, k값이 3인 경우와 5인 경우에는 차원 수에 따른 성능에 큰 변화가 있었다. k가 3일 때는 차원 수가 175일 때 약 5%의 성능 향상이 있었고 k가 5일 때는 차원수가 200일 때 약 7% 정도로 가장 큰 성능 향상이 있었다. 이러한 결과는 차원 수가 올라갈수록 데이터에 대한 특징이 더 잘 표현될 수 있다는 것을 나타낸다. 그러나 k가 1인 경우와 3인 경우에는 차원 수가 200일 때 성능이 저하되는 것을 볼 수 있는데, 이는 희박성(sparsity)이 높은 데이터의 경우 차원 수가 올라갈수록 오히려 성능이 저하될 수 있다는 것을 의미한다.

5. 결론 및 향후 연구

최근 들어서 다중 언어 워드 임베딩(Multilingual Word Embedding)에 관한 관심도 많아졌고 다중 언어 워드 임베딩은 기계번역(Machine Translation), 품사 태깅(POS tagging), 감정분석(Sentiment Analysis) 등 자연어 처리 분야에서 폭 넓게 사용되었음에도 불구하고, 한국어와 특정 언어의 병렬 데이터를 사용한 이중 언어 워드 임베딩의 경우에는 병렬 데이터가 희소하였기 때문에 한국어를 사용한 이중 언어 워드 임베딩의 연구는 활발히 이루어지지 못했다. 또한 그 중에서도 영역 특수적인 목적으로 자연어 처리 관련 분야에서 사용될 수 있는 이중 언어 워드 임베딩 연구는 더 희소하다.

따라서 본 연구에서 한국 법률의 한국어-영어 번역 말뭉치를 사용하여 로컬 워드 임베딩(local word embedding)을 학습시켰으며, 학습에 사용된 말뭉치가 단락 기준 정렬 데이터이고 번역 쌍을 연결할 때 일대일 연결이 어려운 경우를 처리하기 위해서 Random Match, Single Match Greedy Intersect, Multiple Match Greedy Intersect 3가지 연결 전략을 제안하였다.

본 논문의 기여는 첫째, 단락 정렬 말뭉치를 사용한 워드 임베딩과 같은 문장 정렬 말뭉치를 사용한 워드 임베딩 보다 상대적으로 질이 떨어질 수 있는 말뭉치에 대해 다양한 매핑전략을 적용하였다. 둘째, 한국어-영어 번역 쌍을 연결시키는 과정에서 발생하는 하나의 원시 어휘에 여러 개의 단어로 이루어진 목적 어휘가 연결되는 문제를 해결하기 위해 3가지의 전략을 제안하고 각 전략들의 결과를 비교분석하였다. 셋째, 법률 데이터를 사용하여 로컬 워드 임베딩을 학습시킨 결과, 영역 특수적인 자연어 처리 관련 분야에서 위키피디아 데이터를 사용하여 글로벌 워드 임베딩을 학습시킨 이중 언어 임베딩보다 2배의 정확도를 얻음을 보였다.

향후 연구로는 본 연구에서 법률 영역을 대상으로 3가지 전략으로 로컬 워드 임베딩을 실험해보았을 때 글로벌 워드 임베딩보다 성능이 높게 나옴을 확인할 수 있었다. 따라서 다른 영역의 데이터의 경우에도 로컬 워드 임베딩이 글로벌 워드 임베딩보다 성능이 높게 나올 것으로 기대된다. 또한 본 연구에서는 이중 언어 워드 임베딩을 사용하였는데, 다중언어 임베딩인 삼중(tri) 사중(quad)언어 임베딩에서도 로컬 워드 임베딩과 글로벌 워드 임베딩에서도 본 3가지 전략을 사용할 경우 성능의 향상이 있을 것으로 기대된다.

REFERENCES

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems,* 3111-3119.

[2] J. Turian, L. Ratinov, Y. Bengio. (2010). Word representations: a simple and general method for semi-supervised learning. *ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* 384-394.

[3] J. Guo, W. Che, H. Wang, T. Liu (2014). Revisiting embedded features for simple semisupervised learning. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* 110-120.

[4] S. Gouws, A. Sogaard (2013). Simple task-specific bilingual word embeddings. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1386-1390.

[5] M. Artetxe, G. Labaka, E. Agirre (2017). Learning bilingual word embeddings with (almost) no bilingual data. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1,* 451-462.

[6] D. Y. Lee, W. H. Yu, H. S. Lim (2017). Bi-directional LSTM-CNN-CRF for Korean Named Entity Recognition System with Feature Augmentation. *Korea Convergence Society, 8(12),* 55-62.

[7] D. Y. Lee, J. C. Jo, H. S. Lim (2017). User Sentiment Analysis on Amazon Fashion Product Review Using Word Embedding. *Korea Convergence Society, 8(4),* 1-8.

[8] S. H. Lee, C. H. Lee, H. S. Lim (2017). Bilingual Word Embedding Using Parallel Corpus. *Korean Institute of Information Scientists and Engineers,* 645-647.

[9] F. Diaz, B. Mitra, N. Craswell (2016). Query Expansion with Locally-Trained Word Embeddings. *arXiv preprints,* 1605.07891.

[10] Y. Goldberg, O. Levy (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprints,* 1402.3722.

[11] T. Mikolov, K. Chen, G. Corrado, J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprints,* 1301.3781.

[12] S. Ruder, I. Vulic, A. Sogaard (2017). A Survey of Cross-Lingual Word Embedding Models. *arXiv preprints,* 1706.04902.

[13] T. Mikolov, Q. V. Le, I. Sutskever, (2013). Exploiting Similarities among Languages for Machine Translation. *arXiv preprints,* 1309.4168.

[14] M. Faruqui, C. Dyer (2014). Improving Vector Space Word Representations Using Multilingual Correlation. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* 462-471.

[15] L. Duong, H. Kanayama, T. Ma, S. Bird, T. Cohn, (2016). Learning crosslingual word embeddings without bilingual corpora. *arXiv preprints,* 1606.09403.

[16] KM. Hermann, P. Blunsom (2013). Multilingual Distributed Representations without Word Alignment. *arXiv preprints,* 1312.6173.

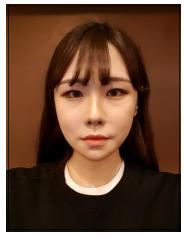
[17] A. Klementiev, I. Titov, B. Bhattarai (2012). Inducing Crosslingual Distributed Representations of Words. *Proceedings of COLING 2012,* 1459-1474.

[18] S. H. Yun, Y. T. Kim (1993). Idiom-Based Analysis of Natural Language for Machine Translation. *Korean Institute of Information Scientists and Engineers, 20(8),* 1148-1158.

임 희 석(Lim, Heui Seok) [정회원]

- 1992년 2월 : 고려대학교 컴퓨터학과 (이학사)
- 1994년 2월 : 고려대학교 컴퓨터학과 (이학석사)
- 1997년 2월 : 고려대학교 컴퓨터학과 (이학박사)
- 2008년 3월 ~ 현재 : 고려대학교 컴퓨터 학과 교수
- 관심분야 : 자연어 처리(NLP), 뇌신경 언어 정보 처리
- E-Mail : limhseok@kroea.ac.kr

최 순 영(Choi, Soon Young) [정회원]



- 2015년 8월 : 건국대학교 컴퓨터공학과 (이학사)
- 2016년 3월 ~ 현재 : 고려대학교 컴퓨터학과 소프트웨어 전공 석사과정
- 관심분야 : 자연어 처리(NLP), 딥러닝(Deep Learning), 인공지능, 워드임베딩
- E-Mail : perfectchl@korea.ac.kr

앤 드 류(Matteson, Andrew) [정회원]

- 2012년 9월 : 미시간 주립 대학교 컴퓨터학과 (이학사)
- 2016년 9월 ~ 현재 : 고려대학교 컴퓨터학과 소프트웨어 전공 석사 과정

- 관심분야 : 자연어 처리(NLP), 딥러닝(Deep Learning), 인공지능, 워드임베딩
- E-Mail : amatteson@korea.ac.kr