다중-어의 단어 임베딩을 적용한 CNN 기반 원격 지도 학습 관계 추출 모델

남상하○, 한기종, 김은경, 권성구, 정유성, 최기선
한국과학기술원
{nam.sangha, han0ah, kekeeo, fanafa, wjd1004109, kschoi}@kaist.ac.kr

CNN-based Distant Supervision Relation Extraction Model with Multi-sense Word Embedding

Sangha Nam○, Kijong Han, Eun-Kyung Kim, Key-Sun Choi
KAIST

요 약

원격 지도 학습은 자동으로 매우 큰 코퍼스와 지식베이스 간의 주석 데이터를 생성하여 기계 학습에 필요한 학습 데이터를 사람의 손을 빌리지 않고 저렴한 비용으로 만들 수 있어, 많은 연구들이 관계 추출 문제를 해결하기 위해 원격 지도 학습 방법을 적용하고 있다. 그러나 기존 연구들에서는 모델 학습의 입력으로 사용되는 단어 임베딩에서 단어의 동형이의어 성질을 반영하지 못한다는 단점이 있다. 때문에 서로 다른 의미를 가진 동형이의어가 하나의 임베딩 값을 가지다 보니, 단어의 의미를 정확히 파악하지 못한 채 관계 추출 모델을 학습한다고 볼 수 있다. 본 논문에서는 원격 지도 학습 기반 관계 추출 모델에 다중-어의 단어 임베딩을 적용한 모델을 제안한다. 다중-어의 단어 임베딩 학습을 위해 어의 중의성 해소 모듈을 활용하였으며, 관계 추출 모델은 문장 내 주요 특징을 효율적으로 파악하는 모델인 CNN과 PCNN을 활용하였다. 본 논문에서 제안하는 다중-어의 단어 임베딩 적용 관계추출 모델의 성능을 평가하기 위해 추가적으로 2가지 방식의 단어 임베딩을 학습하여 비교 평가를 수행하였고, 그 결과 어의 중의성 해소 모듈을 활용한 단어 임베딩을 활용하였을 때 관계추출 모델의 성능이 향상된 결과를 보였다.

주제어: Relation Extraction, Distant Supervision, Word Embedding, Convolution Neural Network

1. 서론

관계 추출(Relation Extraction) 이란 문장 내 등장한 두 개체(Entity) 사이의 관계(Relation)를 알아내는 작업을 일컫는다. 예를 들어, “마크 저커버그는 페이스북 설립자이다.” 라는 문장으로부터 Founder(페이스북, 마크_저커버그) 와 같은 관계를 추출하는 것이다. 최근 들어 지식베이스의 중요성이 대두되고 DBpedia, YAGO, Wikidata 등의 대규모 지식베이스 구축을 위한 연구들이 활발히 진행 중이며, 그에 따라 웹 규모 말뭉치(Web Scale Corpus)에서 지식을 추출하고자 하는 연구들이 진행중이다. 그러나 많은 연구들에서 관계 추출 시스템을 설계하기 위해 기계학습(Machine Learning) 방식을 활용하고 있기 때문에, 많은 양의 지도 학습 데이터(Supervised Learning Data)를 생성하기에는 고비용(High-cost) 문제가 발생하였고, 이를 해결하기 위해 [1]의 논문에서 원격 지도 학습(Distant Supervision) 방식을 소개하였다. 원격 지도 학습 방식은 “두 개체가 지식베이스에서 특정 관계로 연결되어 있고 이 두 개체가 함께 포함된 문장을 말뭉치에서 모두 수집하면, 수집된 문장들은 두 개체간의 특정 관계를 설명하고 있을 것이다.”라는 가정을 기반으로 한다.

그림 1은 자동으로 주석 데이터(Labeled Data)를 수집하는 원격 지도 학습 방식 예시이다. 원격 지도 학습 방식은 대용량 말뭉치, 대규모 지식베이스 간 학습 데이터를 자동으로 생성해준다는 점에서 상당히 효율적이지만, 자동 수집된 학습 데이터의 품질이 항상 좋은 것은 아니라는 문제점을 가지고 있다. 그림 1에서 보는바와 같이 ‘페이스북’과 ‘마크 저커버그’가 동시에 들어있는 문장을 수집해보면 1번 문장과 같이 실제 두 개체 간의 관계(예시: founder)를 의미하는 문장도 수집되지만, 2번 문장과 같이 두 개체가 포함되어 있을 뿐 두 개체간의 명확한 관계를 의미하지 않는 문장도 학습데이터로 수집될 수 있다.

이러한 단점을 해결하기 위해 자동 수집된 학습 데이터의 품질을 향상시키기 위한 다양한 연구들이[2, 3, 4] 소개되었으나, 관계 추출 시스템을 설계할 때 전통적인 자연언어처리 분야에서 사용하던 특징들(Feature)은 자연언어처리 도구에서 발생하는 오류로 인해 에러 전파 및 축적의 문제가 발생하였다. 그에 따라, 전통적인 특징들을 사용하지 않고 단어 임베딩(Word Embedding)과 DNN(Deep Neural Network) 기반의 관계 추출 연구들이 [5, 6] 소개되었고, 기존 연구들보다 향상된 관계 추출 성능을 보였다. 특히 [6]에서 소개된 PCNN(Piecewise max pooling Convolution Neural Network) 모델은 CNN 학습 모델을 관계 추출에 더 적합한 형태로 확장한 것으로, 입력 벡터에 ‘문장 내 주어 및 동사의 위치’를 추가시킨 것과 ‘문장 내 두 개체의 위치를 기준으로 3개의 Max pooling 연산을 수행’하도록 확장한 것이 큰 특징이

다.

그러나 기존 연구들에서는 모델 학습의 입력으로 사용되는 단어 임베딩에서 단어의 정확한 의미를 반영하지 못한 단점이 있다. 예를 들어, '부르다' 라는 단어는 '입으로 소리를 내는 것' 과 '속이 꽉 찬 느낌이 드는 것' 과 '무엇인가 퍼뜨리고 펼치는 것' 등의 여러 의미가 존재하고, 또 세부적으로는 '노래를 부르는 것', '어떤 행동을 동참하도록 유도하는 것' 등 더욱 다양한 의미로 사용된다. 따라서 한 단어에 대해 하나의 벡터값으로 학습을 진행하게 되면 동형이의어의 특성을 제대로 반영하지 못한 결과가 발생할 수 있다. 그에 따라 영어권에서는 다중-어의 단어 임베딩(Multi-sense Word Embedding)에 대한 연구들이[7, 8] 진행 중이지만, 지금까지 다중-어의 단어 임베딩을 관계 추출 모델에 적용한 결과는 발견하지 못했다.

본 논문에서는 원격 지도 학습 기반 관계 추출 모델에 다중-어의 단어 임베딩을 적용한 모델을 제시하고 그 결과에 대해 소개한다. 관계 추출 모델은 [5]에서 제시한 CNN 모델과 [6]에서 제시한 PCNN 모델 두 가지를 활용하였으며, 어의 중의성 해소 모듈을 활용한 다중-어의 단어 임베딩을 관계추출 모델의 입력으로 사용하였다. 단어 임베딩은 Word2Vec[9]의 Skip-gram 모델을 기반으로 단어와 형태소, 그리고 단어 의미 번호(word sense)를 함께 토큰화하여 학습을 진행하였으며, 본 논문에서 제안한 방식의 우수성을 입증하기 위해 비교대상으로 (1) 단어 단위 임베딩, (2) 형태소 단위 임베딩을 추가 학습하여 관계 추출 성능에 대한 비교 평가를 수행하였고 그 결과를 소개한다.

2. 관련 연구

2.1 단어 임베딩 스킵-그램 모델

스킵 그램(Skip-gram) 모델은 그림 2와 같은 구조로, 타겟 단어를 기준으로 주변에 등장할 단어의 여부를 유추하는 것으로 학습을 진행한다. Skip-gram 모델에서는 주어진 단어(w_t)와 w_t 주변에서 등장한 단어들(c)의 벡터값을 아래 목적 함수(Objective Function)를 최대화하는 방향으로 학습을 진행한다. 아래 식에서 w_t는 학습 코퍼스 내 타겟 단어를 의미하고, c_t는 w_t단어 주변에서 실제 나타난 단어를 의미한다. 그리고 c'_t는 w_t주변에 등장하지 않은 단어들 중 랜덤하게 선택한 Negative sampling을 의미한다. 즉, 타겟 단어 주변에 실제 등장한 단어들을 예측할 확률과 실제 등장하지 않은 단어들을 예측하지 않을 확률을 최대화하는 방식으로 학습이 진행된다.

$$J(\theta) = \sum_{(w_t,c_t)\in D+} \sum_{c\in c_t} logP(D=1|v(w_t),v(c)) + \sum_{(w_t,c'_t)\in D-} \sum_{c'\in c'_t} logP(D=0|v(w_t),v(c'))$$

2.2 PCNN 관계 추출 모델

Convolutional Neural Network(CNN)은 이미지 분류 및 문장 감성 분류(Sentiment Classification) 등에서 우수한 성능을 보이는 모델이다. CNN의 특징이자 장점 중 하나인 입력 데이터 내 주요 특징을 효율적으로 찾아낸다는 점에 착안하여 [5, 10]의 논문에서는 CNN을 이용한 관계 추출 모델을 제안하였다. 그 중 [5]의 논문에서는 위치 임베딩(Position Embedding) 개념을 도입하여, 문장 내 두 개체의 위치를 입력 벡터에 추가해줌으로써 관계 추출 모델의 성능 향상을 보였다. 위치 임베딩이란 문장 내 두 개체와 개체가 아닌 단어들 간의 상대적 거리를 n 차원의 벡터로 임베딩 한 것이다. 예를 들어, 그림 3에서 보는바와 같이 '설립' 이라는 단어가 '마크_저커버그' 개체로부터 3단어만큼 떨어져있고, '페이스북' 개체로부터 1단어만큼 떨어져있다. 이 상대적 거리를 n 차원의 벡터로 임베딩하고, 그 값을 모델 학습의 입력 벡터 중 일부로 사용한다.

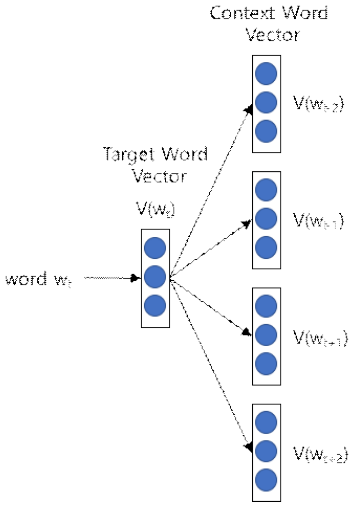

그림 2 Skip-gram 모델 구조 (Window size = 2)

그림 3 위치 임베딩의 상대적 거리 예시

PCNN(Piecewise max pooling Convolutional Neural Network)은 [6]의 논문에서 제안한 관계 추출 모델의 하나로 [10]에서 제안한 CNN 모델을 확장한 것인데, CNN에서 흔히 사용하는 Single Max Pooling Layer를 Piecewise Max Pooling Layer로 확장하였다는 것이 큰 차이점이다. CNN에서 맥스 풀링(Max Pooling)은 Activation Map 혹은 Feature Map이라 불리는 Convolution Layer의 출력 매트릭스에서 가장 큰 값 즉, 가장 중요한 특징을 추출하는 방법이다. 그러나 Single Max Pooling Layer는 은닉 층(Hidden Layer) 결과값 중 최대값 하나만을 선택함으로써 관계 추출에 필요한 특징을 세밀하게 파악하기 힘들다. PCNN은 이러한 단점을 해결하기 위해 Single Max Pooling Layer를 3덩이로 나눈 Piecewise Max Pooling Layer를 제안하였다. 이 층(Layer)의 큰 특징은 문장을 총 3덩이로 나누어서 각각 맥스 풀링을 수행한다는 점이다. 관계 추출에 사용되는 문장은 항상 두 개체를 포함하고 있기 때문에 두 개체를

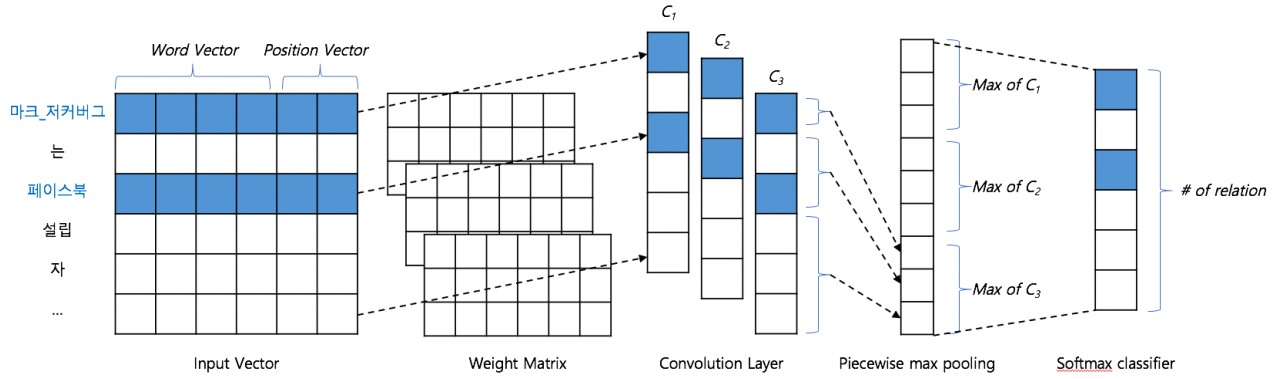

그림 4 PCNN 구조 및 예시

기준으로 문장을 총 3덩이로 나눌 수 있고 각 덩이마다 최대값을 추출하여 학습에 반영한다. PCNN의 구조는 그림 4와 같다. 입력 벡터(Input Vector)는 단어 벡터(Word Vector)와 위치 벡터(Position Vector)로 구성되며 3개의 Convolution Layer, Piecewise Max Pooling Layer, Softmax Output 단계로 전체 구조가 이루어진다.

3. 방법론

본 논문에서는 CNN 및 PCNN 기반의 관계 추출 모델을 활용하여 어의 중의성 해소(Word Sense Disambiguation) 기반 다중-어의 단어 임베딩을 적용한 관계 추출 모델을 제안한다.

3.1 단어 임베딩

단어 임베딩(Word Embedding)은 최근 자연언어 처리 분야에서 상당히 유용하게 사용되고 있다. 일반적으로는 입력 말뭉치를 토큰(Token) 단위로 분할한 다음 의미적 연관성이 높은 토큰들을 유사한 실수 벡터 값으로 생성한다. 이때 영어에서는 일반적으로 단어(Word) 단위 즉, 띄어쓰기 단위로 토큰을 생성한다. 그러나 한국어는 조사, 어미, 그리고 접미사 등 한 단어를 이루는 복수개의 구성 요소들로 인해 띄어쓰기 단위로 단어 임베딩 학습을 진행하면 그 결과가 상대적으로 영어만큼 좋지 않다. 그에 따라 한국어에서는 형태소 단위로 토큰을 구성하여 단어 임베딩을 학습하는 방식이 사용되고 있고, 이때 토큰의 구성 요소로 품사태그가 함께 사용되기도 한다. 예를 들어, '사과/Noun'과 같이 품사태그를 함께 학습할 경우 이 단어와 유사한 단어들로 '사죄/Noun', '애도/Noun', '죄송하다/Adjective' 등이 위치하게 된다. 품사태그를 함께 단어 임베딩 학습에 활용했을 때 좋은 점은 동일한 형태의 단어가 동사로 쓰였는지 명사로 쓰였는지 등에 대한 구분이 가능하다. 예를 들어, '가지'라는 단어는 식물을 뜻하는 명사로 쓰이기도 하고 소유하다는 의미의 동사 어근으로 쓰이기도 한다. 따라서 한국어에서는 단어 임베딩 생성 시 형태소 단위 그리고 품사태그를 함께 학습에 이용하는 것이 효율적이다.

그러나 위 방식의 단어 임베딩은 단어의 실제 의미를 반영하지 못한다는 단점이 있고, 이는 한국어뿐만 아니라 영어에서도 마찬가지이다. 예를 들어, 'apple'이라는 단어는 과일로 쓰이기도 하고 회사로 쓰이기도 한다. 단어 임베딩은 앞서 2장에서 언급하였듯이 주변 단어가 어떤 것들로 구성되느냐에 따라 학습되는데, '과일 apple' 주변에 등장하는 단어와 '회사 apple' 주변에 등장하는 단어를 모두 'apple' 단어 주변에 등장하는 단어로 뭉쳐서 학습을 진행하기 때문에 결국 'apple'은 하나의 n 차원 실수 벡터값을 가지게 되고 여러 의미를 구분 지을 수 없는 단어 임베딩이 학습된다. 한국어에서도 마찬가지로 '사과'를 '과일 사과'와 '사죄의 사과'를 섞어 학습하게 된다. 그에 따라 아래 식과 같은 삼각 부등식 문제(Triangle Inequality Problem)가 발생할 수 있다[7].

$$distance(a,c) \leq distance(a,b) + distance(b,c)$$

예를 들어 pollen(꽃가루)와 refinery(정제 공장)간의 유사도(distance)가 pollen(a)과 plant(b)와의 거리 그리고 refinery(c)와 plant(b)와의 유사도 합 보다 작아지게 되는 문제가 발생한다. 즉, plant라는 동형이의어를 중심으로 'pollen'과 'refinery'라는 두 단어의 유사도가 실제 의미적 연관성보다 가까워지는 문제이다. [7]

이러한 문제를 해결하기 위해 단어의 의미를 여러 개의 군집으로 클러스터링 하여 분할하는 방법 [7], 워드넷(WordNet)을 기반으로 단어 의미를 구분하는 방법 [8], 사전 뜻풀이를 학습에 활용하여 단어 의미를 구분하는 방법 [11] 등이 발표되었다. 본 논문에서는 단어의 의미를 구분 짓는 어의 중의성 해소(WSD) 모듈의 결과를 활용하여 단어 임베딩 학습을 진행하였고, 어의 중의성 해소 모듈은 본 연구팀에서 연구 개발한 모듈을 활용하였다.

아즐리 고등학교를 다닐 당시 그는 서양고전학 (classics) 과목에서 우수한 성적을 거두었다. 이후 3학년 때 필립스 엑세터 아카데미로 학교를 옮긴 그는 과학(수학, 천문학 및 물리학)과 서양고전 연구(Classical

그림 5 한국어 위키피디아 개체 태깅 예시. '아즐리 고등학교', '서양고전학', '필립스 엑세터 아카데미'

추가적으로, 관계 추출에 적합한 단어 임베딩을 학습하기 위해 여러 단어로 된 개체(Multi-word Entity)는

하나의 토큰으로 묶는 개체-반영 단어 임베딩 학습을 진행하였다. 그림 4에서 보는 바와 같이 '마크 저커버그'는 두 단어로 구성된 하나의 개체이다. 한 개체는 여러 단어로 구성되었다 하더라도 하나의 단어 임베딩 값을 가지도록 학습하는 것이 단어 임베딩과 관계추출 모델 설계에 적합하다. 만약, 개체에 대한 구분 없이 모든 토큰을 형태소 단위로 학습을 진행하게 되면 여러개의 단어로 구성된 개체에 대한 임베딩은 얻어낼 수 없다. 그 결과, 예를 들어, '코피 아난'과 유사한 단어로는 '국제 연합 사무총장', '반기문', '유엔 사무총장', '연방 준비 제도' 등이 위치한다. 말뭉치 내 개체 판별은 한국어 위키피디아에 태깅되어 있는 개체들을 사용하였으며, 예시는 그림 5와 같다. 파란색으로 표기되는 이 개체들은 위키피디아 내용 작성자들이 손으로 태깅한 것으로 정확도가 매우 높다고 할 수 있다.

3.2 관계 추출 모델

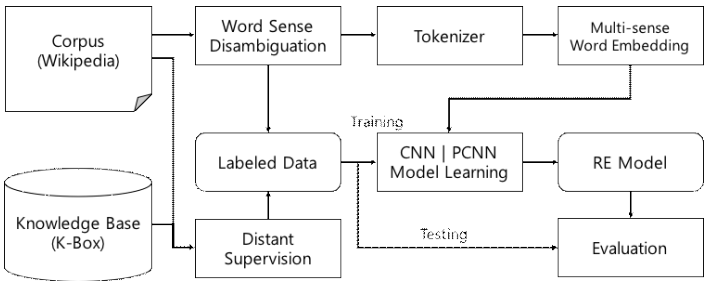

그림 6 관계 추출 시스템 구성도

본 논문의 관계 추출 시스템 구성도는 그림 6과 같고, 크게 단어 임베딩 학습과 원격 지도 학습 관계 추출 모델 학습 및 평가 부분으로 구성된다. 먼저 코퍼스를 입력으로 받아 어의 중의성 해소(Word Sense Disambiguation)를 수행한다. 어의 중의성 해소 단계는 본 연구팀의 비-지도 학습 방식 기반 MRF WSD 모듈을 활용하였으며, 이 모듈은 코어넷[13]의 개념체계를 기준으로 중의성을 해소한다. 그 다음 토큰화(Tokenizer) 단계를 수행하는데, 이때 형태소 분석기는 [12]의 Twitter 형태소 분석기를 사용하였다. 그리고 3.1절에서 설명한 것과 같이 개체-반영 토큰화를 수행하여 여러 단어로 구성된 하나의 개체는 하나의 토큰으로 인정하였다. 그 다음 Skip-gram 모델로 다중-어의 단어 임베딩을 학습하여 단어별로 의미 번호가 부착된 형태의 토큰들이 각각 임베딩 값을 가질 수 있게 하였다.

그 다음, 지식베이스(Knowledge Base)와 말뭉치(Corpus) 간 원격 지도 학습 데이터 수집(Distant Supervision)을 수행하고, 이때 수집된 문장은 단어 임베딩 학습과 동일한 방법으로 토큰화를 수행한다. 그리고 이 데이터(Labeled Data)를 두 그룹으로 나누어 한 그룹은 학습, 나머지 한 그룹은 평가에 사용한다. 관계 추출 모델은 [5]의 CNN 모델과 [6]의 PCNN 모듈 2개를 각각 학습 및 평가에 사용하였고, 두 모델 모두 가중치 매트릭스를 3개씩 사용하여 3개의 Convolution Layer를 생성한 다음 매트릭스 연결(concat)방식으로 결합(merge) 하였다. CNN 모델은 Single Pooling Layer 그리고 PCNN 모델은 Piecewise Max Pooling Layer 방식으로 구현하였으며, 나머지 층은 모두 동일하게 구성하였다.

4. 실험

4.1 실험 데이터

실험을 위해 한국어 위키피디아 2017년 7월 1일 버전의 말뭉치의 6,941,760 문장과 디비피디아(DBpedia) 2016-10 버전 기반의 K-Box를 지식베이스로 사용하였다. K-Box는 한국어 관계(Local Property)로 정의된 트리플을 디비피디아 공용 관계(Ontological Property)로 변환 및 확장한 지식베이스이다. 예를 들어, 'prop-ko:탄생지'와 같은 한국어 관계를 'dbo:birthPlace'와 같은 디비피디아 공용 관계로의 변환을 말하며, 관계 변환 테이블(Mapping Table)은 온톨로지 전문가 3명이 수작업으로 생성하였다. 원격 지도 학습 데이터 수집 결과 총 451개의 관계를 기준으로 358,464개의 Labeled Data를 수집하였고, 이때 대부분의 관계들이 수집된 데이터의 양이 적은 Long tail 문제에 해당한다. 다중 분류기(Multi-class Classifier) 모델에서 클래스 별 학습할 데이터가 적으면 제대로 학습이 진행되지 않기 때문에, 원활한 관계 추출 모델 학습을 위해 각 관계별로 수집 데이터의 개수가 1000개 이상인 68개 관계 기준 200,323개의 데이터를 학습 및 평가에 사용하였고, 수집 데이터 수 기준 Top 10개의 관계 통계는 표 1과 같다.

표 1 수집 데이터 기준 수 기준 Top 10 관계 통계

Property	# of collected data	Property	# of collected data
country	73,890	team	6,110
isPartOf	49,957	birthPlace	5,962
capital	12,953	successor	5,228
location	11,570	deathPlace	5,191
part	7,947	owner	4,971

4.2 실험 결과

본 논문에서 제안한 방식의 우수성을 입증하기 위해 총 3가지 방식의 단어 임베딩을 학습하였으며, 학습 시 공통으로 설정한 하이퍼파라미터(Hyperparameter)는 표 2와 같다. (1) 어절단위 토큰화, (2) 형태소단위 토큰화, (3) 어의중의성 해소 토큰화

표 2 단어 임베딩 하이퍼 파라미터

Dimension	Window	Min. Word
100	5	1

샘플 단어에 대한 유의어 결과는 표 3과 같다. 표에서 볼 수 있듯이 어의 중의성 해소 결과를 반영한 단어 임베딩은 그렇지 않은 단어 임베딩보다 동일한 단어에 대해 여러 의미로 구분이 가능하며, 각 의미마다 연관 있는 단어들끼리 군집이 이루어짐을 확인할 수 있다. 또한, 개체를 반영한 단어 임베딩을 학습하였기 때문에 여러 단어로 구성된 개체를 하나의 임베딩으로 학습하는

것을 확인할 수 있고 근접한 단어들도 상당히 의미있음을 확인할 수 있다.

학습된 다중-어의 단어 임베딩 결과를 활용하여 관계 추출 모델의 Held-out 성능 평가를 진행하였다. Held-out 평가는 수집된 데이터를 절반으로 나누어 한 그룹으로 학습, 나머지 한 그룹으로 평가를 진행하여 정확도(Precision), 재현율(Recall), 그리고 F1-score를 측정하는 방식이고, 그 결과는 표 5와 같다.

본 논문에서 제안한 관계 추출 모델 학습 시 어의 중의성 해소 임베딩의 효과를 확인하기 위하여 3개의 각기 다른 임베딩 값을 입력으로 사용하여 성능을 측정해보았고, 각 모델의 하이퍼파라미터는 표 4과 같이 설정하였다.

표 3 '시장'과 '사과'의 유사단어

Token	Word	Similar Words
Word/POS	시장	투자, 유통, 수익, 수출, 자산, 대기업, 매출, 수입, 업계, 가격
	사과	묻다, 사죄, 죄송하다, 건네다, 애도, 봉투, 제보, 하소연, 해명, 발언
Word/POS /WSD	시장-0 (물건을 사고파는 장소)	시장, 산업-0, 업계-0, 경쟁력, 중소기업-0, 사업-4, 투자-0, 산화방지제, 금융-0
	시장-1 (지방 자치 단체 장)	교육감-0, 기초자치단체장, 새누리당, 박순자, 고진화, 박영선_(1960년), 송광호, 대한민국_제5회_지방_선거, 전병헌
	사과-3 (자기의 잘못을 인정하고 용서를 빎)	사죄-0, 건네다-0, 고소하겠, ㅠㅠ, 봉투-0, 죄송하다, 모닝스타, 낸시랭, 앉은뱅이-
	사과-4 (사과나무의 열매)	과일-1, 완두-0, 잠자리-1, 밤-1, 포도-0, 뱀장어, 살구-0, 호두, 견과류, 옻나무
Entity	유엔	유럽_연합, UN, 국제_연합, 유럽_공동체, 북대서양_조약기구, 국제_연맹, 국제연합, 안전보장

표 4 모델별 하이퍼파라미터

Model	Activation	Optimizer	Dropout
CNN	RELU	ADADELTA	1
PCNN	RELU	ADAM	1

표 5 단어 임베딩 별 관계추출 모델 평가 결과

Model	Embedding	Precision	Recall	F1-score
CNN	Word	0.5537	0.3506	0.4275
	+POS	0.5315	0.4279	0.4739
	++WSD	**0.5921**	**0.5039**	**0.5443**
PCNN	Word	0.457	0.3251	0.3799
	+POS	0.4555	0.3472	0.394
	++WSD	0.4529	0.3713	0.4081

평가 결과, 두 모델 모두 단어만 사용한 임베딩보다 형태소를 함께 사용하는 것의 성능이 향상되었고, 형태소를 함께 사용한 것 보다 어의 중의성 해소 모듈 결과를 반영한 단어 임베딩을 사용하는 것의 성능이 향상됨을 확인할 수 있다. 단, 이때 모든 단어 임베딩 방법에서 개체-반영은 공통적으로 수행하였다. 그리고 PCNN 모델보다 CNN을 사용한 모델이 성능이 약간 더 높았는데, 한국어에서는 문장 내 두 개체의 위치가 문장의 맨 첫 번째에 나오는 경우, 그리고 두 개체가 연접해있는 경우가 많아 PCNN 보다 CNN 방법이 더 높은 성능을 보인 것으로 판단된다. 또한 하이퍼파라미터를 바꿔가며 수행한 여러 번의 반복 실험에서 Dropout은 하지 않는 것이 더 높은 성능을 보여주었다.

5. 결론 및 향후연구

본 논문에서는 원격 지도 학습 기반 관계 추출 모델에 다중-어의 단어 임베딩을 적용한 관계추출 모델 성능향상 방법을 제안하였고, CNN과 PCNN기반의 두 관계 추출 모델에 적용한 실험을 수행하였다. 또한 관계 추출을 위한 단어 임베딩 생성 시 여러 단어로 구성된 개체에 대해 하나의 토큰으로 반영하는 개체-반영 단어 임베딩을 기본적으로 활용하였으며, 그 결과 다중-어의를 해소한 단어 임베딩이 관계 추출 모델의 성능을 향상시킴을 확인할 수 있었다.

향후에는 자연언어의 특성인 시계열성을 반영하여 CNN과 RNN의 결합 모델인 Convolutional RNN 모델을 관계추출 문제에 적용하는 방향의 연구를 수행할 예정이다. 그리고 원격지도학습 데이터 수집 시 발생하는 문제점 중 하나인 에러 데이터 제거 방법에 대한 연구를 진행할 계획이다.

사사

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2013-0-00109, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] Mintz, Mike, et al. "Distant supervision for relation extraction without labeled data." Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th

International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009.
[2] Sebastian Riedel, Sebastian Riedel, Limin Yao, and Andrew McCallum. "Modeling relations and their mentions with- out labeled text." In Proceedings of ECML PKDD, pages 148-163. 2010.
[3] Raphael Hoffmann, et al. "Knowledge-based weak supervision for information extraction of overlapping relations." In Proceedings of ACL, pages 541-550. 2011.
[4] Mihai Surdeanu, et al. "Multi-instance multi-label learning for relation extraction." In Proceedings of EMNLP-CoNLL, pages 455-465. 2012.
[5] Yoon Kim. “Convolutional neural networks for sentence classification.” In Proceedings of EMNLP, pages 1746-1751. 2014.
[6] Zeng, D., et al. "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks." In Proceedings of EMNLP, pages 1753-1762. 2015.
[7] Neelakantan, A., et al. "Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space." CoRR, cs.CL. 2015.
[8] Rothe, S., & Schütze, H. "AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes." arXiv.org. 2015.
[9] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.
[10] Zeng, D., et al. "Relation classification via convolutional deep neural network." In Proceedings of COLING, pages 2335-2344. 2014.
[11] 이주상, 신준철, and 옥철영. "단어 의미와 자질 거울 모델을 이용한 단어 임베딩." *정보과학회 컴퓨팅의 실제 논문지* Vol.23 No.4, pages 226-231. 2017.
[12] 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.
[13] 한국과학기술원 전문용어언어공학연구센터, “다국어 어휘의미망 제1권: 어휘의미망 구축론”, KAIST Press, 2005.