https://doi.org/10.7840/kics.2018.43.11.1903

트위터 데이터를 위한 효과적인 단어 임베딩

김 인 환◆, 장 백 철°

Effective Word Embedding for Twitter Data

Inhwan Kim◆, Beakcheol Jang°

요 약

본 논문에서는 트위터 데이터를 기반으로 한 효과적인 단어 임베딩을 위하여 학습 모델 선택과 학습 매개 변수 조정에 대한 지침을 제공한다. 기존 단어 임베딩의 모델과 매개 변수에 대한 연구는 뉴스 및 위키피디아와 같이 정형화된 데이터를 기반으로 연구가 진행되어 트위터와 같은 비정형화 데이터에 적용되기 어렵다. 따라서 본 연구는 트위터 데이터 분석을 위해 최신 단어 임베딩 기술인 Word2Vec을 사용하여 학습 모델 선택과 학습 매개 변수 조정에 따른 성능 변화를 분석하는 실험을 실시하였고 트위터 데이터를 위한 좋은 성능의 단어 임베딩을 위한 효과적인 학습 모델과 매개 변수 값을 제공한다.

Key Words : Word Embedding, Natural Language Processing, Artificial Intelligence, Word2Vec, Skip-gram, CBOW, Twitter

ABSTRACT

In this paper, we provide guidelines for selections of learning model and learning parameter values for effective word embedding for Twitter data. The precedent studies on the model and parameters of the word embedding have been studied based on structured data such as news and Wikipedia, so it difficult to apply them to unstructured data such as Twitter. In this paper, we conducted experiment to analyze the performance change by selecting the learning model and adjusting the learning parameters using state-of-the-art word embedding technology, Word2Vec, for Twitter data and provide effective learning models and parameter values for good word embedding for twitter data.

Ⅰ. 서 론

자연어 처리 분야에서 단어 임베딩은 문서 내 단어의 의미론적 유사성을 확인할 수 있는 단어 표현을 위한 중요한 연구이다. 최근 단어 임베딩 연구는 유사한 의미를 가진 단어는 비슷한 분포를 가진다는 분포 가설을 기반으로 한 연구가 진행되어왔다[1,2]. 분포 가설과 인공 신경망을 활용한 단어 임베딩 연구로서 Mikolov et al.[3]은 주변 문맥과 특정 단어의 관계를 통해 출현 확률을 학습하는 Word2Vec을 소개하였고, Pennington et al.[4]은 문서 내 단어 간 동시출현확률을 학습하는 GloVe를 소개하였다. 소개된 단어 임베딩 기술 중 Word2Vec은 CNN (Convolutional Neural Network)와 함께 사용되어 문장 분석[5]과 문장 분류[6] 등 다양한 분야에서 활용되고 있다.

단어 임베딩을 활용한 연구는 단어 임베딩의 성능에

※ This work was supported by Sangmyung University Research Grant.
◆ First Author : (ORCID:0000-0002-2621-386X)Sangmyung University Department of Computer Science, moreih29@gmail.com, 학생회원
° Corresponding Author : (ORCID:0000-0002-3911-5935)Sangmyung University Department of Computer Science bjang@smu.ac.kr, 정회원
논문번호 : 201808-251-C-RN, Received August 17, 2018; Revised October 19, 2018; Accepted October 25, 2018

따라 전체 성능이 영향을 받기 때문에 같은 학습 데이터 안에서 단어 임베딩 학습에 영향을 주는 요소들을 비교하여 효과적인 단어 임베딩을 위한 연구가 진행되고 있다[8,9]. Lai et al.[8]은 학습 데이터의 크기와 종류 그리고 학습 매개 변수의 조정에 따른 단어 임베딩의 성능을 비교였고 특히, 학습 데이터의 크기보다 데이터의 종류가 결과에 더 큰 영향을 준다는 것을 확인하였다. Chiu et al.[9]는 의학 생명 분야 학습 데이터를 사용하여 학습 매개 변수의 조정이 단어 임베딩의 성능에 큰 영향을 미치는 것을 확인하였다. 따라서, 효과적인 단어 임베딩을 위해서는 학습 데이터와 그에 맞는 적절한 학습 매개 변수의 선정이 중요하다. 하지만 어떠한 지침도 없이 적절한 학습 매개 변수를 찾기 위해 학습 매개 변수를 조정하는 것은 많은 시간이 소요된다. 또한 기존 연구들은 뉴스와 위키피디아, 의학 저널과 같이 정형화된 데이터를 기준으로 학습 매개 변수와 단어 임베딩의 성능의 관계를 분석하였기 때문에 해당 연구에서 선택한 학습 매개 변수는 비정형화 데이터를 사용한 단어 임베딩의 학습에서 적절하지 않을 수 있다.

비정형화 데이터의 대표적인 예시는 소셜 네트워크 서비스(Social Network Service, SNS)이다. 트위터는 사람들이 많이 사용하는 SNS 중 하나로서 사람들은 트위터를 통해 다양한 정보를 공유하고 있으며 해당 데이터를 분석하여 오피니언 마이닝 및 감정분석[10], 집단 의견차이 분석[15], 이벤트 추출[11]과 같은 연구가 활발히 진행되고 있다. 그래서 본 연구는 트위터 데이터를 기반으로 최신 단어 임베딩 기술인 Word2Vec의 학습 매개 변수의 조정에 따른 성능 변화를 분석하여 학습 매개 변수 조정의 지침을 제공함으로써 이후 연구에서 적절한 학습 매개 변수를 찾는데 소요되는 시간을 감소시킨다.

본 논문은 Word2Vec의 두 가지 학습 모델인 CBOW와 Skip-gram과 학습에 영향을 미치는 7가지 학습 매개 변수를 선정하고 학습 모델과 매개 변수를 조정해가며 특정 이벤트가 발생한 날짜의 트위터 데이터를 학습시킨다. 학습된 모델에서 이벤트 키워드 주변 단어를 확인하고 연관성을 확인하여 Skip-gram와 CBOW의 성능을 비교하고 가장 높은 성능을 보인 학습 매개 변수의 값을 확인한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 설명하고, 3장에서는 제안하는 모델을 소개한다. 4장에서는 제안 방법을 통한 실험의 결과와 성능을 분석한다. 5장에서는 결론 및 향후 연구 목표를 제시한다.

Ⅱ. 관련 연구

2.1 Word2Vec vs GloVe

최신 단어 임베딩 기술인 Word2Vec과 GloVe의 성능을 비교하기 위해 Muneeb et al.[7]은 생물 의학 분야 학습 데이터에서 다수의 설문을 통해 정해진 566개 단어 쌍의 의미론적 유사성[12]을 단어 임베딩 모델의 유사성과 비교하여 Word2Vec의 성능이 높다는 것을 확인하였고 Lai et al.[8]은 위키피디아 데이터를 기반으로 의미론적 유사성 비교[13]와 Mikolov et al.[3]이 평가 방법으로 사용한 유추를 사용하여 위 실험과 동일하게 Word2Vec의 성능이 높게 나온 것을 확인했다.

2.2 Word2Vec

Word2Vec은 문맥에서 유사한 의미를 갖는 단어들은 가까운 거리를 갖는다는 가정[14]을 기반으로 한다. 특정 단어와 일정 크기 내에 근접한 단어들이 해당 단어와 함께 등장할 확률을 조정하고 학습함을 통해 단어를 벡터화 한다. 단어의 벡터화를 통해 단어의 유사성을 코사인 유사도를 통해 나타낼 수 있으며 단어 간 유사도의 차이를 통해 의미론적 추론을 가능하게 한다. 예를 들어, [그림 1]과 같이 중국과 베이징의 관계와 일본과 도쿄의 관계는 국가와 수도라는 관계로서 유사하기 때문에 Vec("중국") - Vec("베이징") + Vec ("일본")과 같은 연산을 통해 Vec("도쿄")를 추론할 수 있다. Word2Vec의 학습 모델은 CBOW(Continuous Bag-of-Words)와 Skip-gram를 사용한다. [그림 2]는 CBOW와 Skip-gram모델의 구조를 표현한다. CBOW 모델은 목적 단어의 주변 단어로부터 목적 단어의 출현 확률을 계산하여 오차를 조정하는 방식을 사용하고 Skip-gram모델은 목적 단어로부터 주변 단어의 출현 확률을 계산하고 오차를 조정하는 방식을 사용한다.

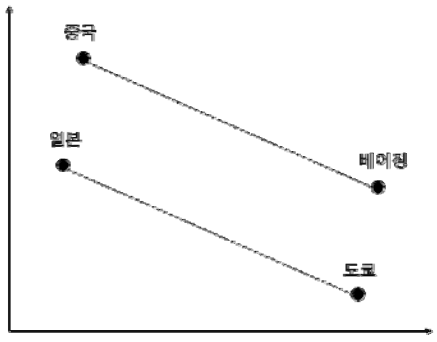


그림 1. Word2Vec의 단어 표현
Fig. 1. Word representations of Word2Vec

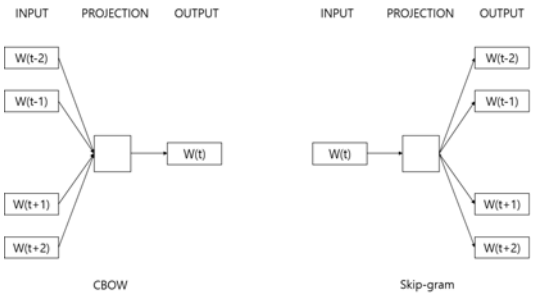


그림 2. Word2Vec의 모델 구조
Fig. 2. Model architecture of Word2Vec

2.3 학습 매개 변수 조정

Word2Vec을 학습할 때 필요한 학습 매개 변수는단어 임베딩 성능에 영향을 미친다. Chiu et al.[9]은 의학 생명 분야 학습 데이터에서 CBOW와 Skip-gram모델에 대하여 566개의 단어 쌍 유사성 비교를 통해 성능을 비교하여 Skip-gram모델이 더 나은 성능을 보인다는 것을 확인하였고, Skip-gram모델에 대하여 6가지 학습 매개 변수 (negative sampling, sub-sampling, minimum-count, learning rate, vector dimension, context window size)의 변화에 따른 성능 변화를 분석하였고 학습 매개 변수의 조정이 단어 임베딩 성능에 큰 영향을 미친다는 것을 확인하였으며 각 학습매개 변수의 최적 값을 제시하였다. 하지만 학습 데이터로 정형화된 의학 저널을 사용했기 때문에 제시된 학습 매개 변수 값은 트위터와 같은 비정형화 데이터에서 적절하게 적용되지 않을 수 있기 때문에 트위터 데이터를 위한 학습 매개 변수 조정 연구가 필요하다.

Ⅲ. 실험 방법

3.1 실험 데이터

본 연구에서 사용한 데이터는 2017년 8월부터 10월까지 질병과 관련된 키워드를 통해 수집한 트위터 데이터 중 특정 이벤트가 발생한 날의 트윗이다. 데이터 수집을 위해 사용된 키워드는 일반 단어(질병, 전염, 바이러스 등) 14개, 증상 관련 단어(발열, 기침, 두통 등) 13개, 법정감염병명(A형간염, 메르스, 수두 등) 21개로 총 48개이며, 데이터 수집 기간 동안 발생한 질병 관련 이벤트는 [표 1]과 같다.

3.2 학습 모델 및 매개 변수

질병 관련 이벤트가 발생한 날의 트위터 데이터를 Word2Vec의 Skip-gram과 CBOW모델을 사용해 학습시킨다. [표 3]은 학습에 사용된 학습 매개 변수 종류와 값을 의미하며 굵은 글씨로 표시된 값은 한 매개 변수 값을 조정할 때 고정된 다른 매개 변수 값을 의미한다. 각 학습 매개 변수의 의미는 다음과 같다.

Minimum Frequency (min-freq) : 단어의 출현 횟수가 최솟값을 넘지 못할 때 해당 단어를 학습에서 제

표 1. 2017년 8월부터 12월 안에 발생한 질병 관련 이벤트의 날짜와 핵심 단어 및 내용
Table 1. Dates, keywords and content of disease-related events that occurred in August 2017 through December 2017

날짜	이벤트 단어	이벤트 내용
20170817	족발	족발·편육 제품 중 일부에서 식중독균 및 대장균 대량 검출
20170818	계란	계란에서 피프로닐, 비펜트린과 같은 살충제 성분 검출
20170825	소시지	유럽산 햄·소시지로 인한 E형 간염 감염 우려로 잠정 판매 중단
20170902	맥도날드	불고기 버거 섭취 후 집단 장염 발생
20170917	일본	대구서 올해 첫 일본뇌염 환자 발생
20170928	생리대	생리대에서 유기화학물(VOCs)이 검출되어 식약처에서 유해성 검사
20171002	개미	'살인 개미'로 불리는 맹독성 붉은 독개미 부산항 상륙
20171013	진드기	'살인 진드기'에 물린 80대 환자 중증열성혈소판감소증후군(SFTS) 의심
20171020	고래회충	학교 급식서 위 벽에 침투해 복통을 유발하는 고래회충 발견
20171024	녹농균	기업 대표 사망 원인인 패혈증의 원인으로 녹농균 의심
20171104	레지오넬라증	청송군에 위치한 '솔샘' 온천을 이용한 이용객 2명 레지오넬라증 감염
20171115	인플루엔자	경기도서 올해 첫 B형 인플루엔자 발견
20171205	중국	중국 윈난성 서 AI 인체감염 환자 발생으로 보건 당국은 중국 여행시 AI 인체 감염 주의 당부
20171212	오리	전남 영암 오리 농가에서 고병원성 AI 확진
20171218	신생아	이대목동병원 사망 신생아 3명 세균 감염 의심

표 2. 기본 학습 매개 변수로 학습한 Word2Vec 모델에서 추출된 이벤트 단어와 근접한 5개 단어
Table 2. Five words close to the event words extracted from the Word2Vec model learned with the default learning parameters

이벤트 단어	Skip-gram	CBOW
족발	애쉬, 추격, 브금, 경로, 인플루엔자	애쉬, 브금, 경로, 인플루엔자, 공부
계란	복지부, 해소, 에는, 강화, 공공	바꾸기로, 치던, 느낀, 사람과, 퍼져서
소시지	있습니다, 치료, **간염,** 성매매, 가열	**간염**, 대한, #국민편익, 폭력, 식약처
맥도날드	전쟁, 대책, 깨끗한, 언약, 했고	@YouTube, 이번, 냄새, 가축, 근데
일본	세계, 하지, 의미, 바로, 어떤	귀여워, 비쥬, 맑아요, 흰색, 무사해서
생리대	존재, 죽음, 홍보, 무수 ,우리나라	추천, 도박, 음식, 조사, 한국
개미	신체, 담배, 슬픔, 심장, 구멍	가축, 플루, 질병, 생각, 농장
진드기	성격, 지사, 개선, 웹툰, 해놓고	되는게, 방문, 요도, 축산, 마니
고래회충	불가, 주인공, 모두, 증진, 진영	주스, 카운트, 소녀, 주인공, 헤세
녹농균	경로, **사망**, 맞다, 다문, 책임	빗나감, 경로, **사망**, 되면, 세상
레지오넬라증	캠페인, 무기, 에너지, 가능, 후지산	캠퍼스, 식물, 입구, 청구권, 전투
인플루엔자	안전, 방법, 설령, **진단**, 체력	기업, 언급, 신문, 규제, 일상
중국	물이, 국산, 오히려, 필링, 츠키	기념일, 국산, 예년, 박스, 손수
오리	**확진**, **전남**, 케어, 이름, 위생	**방역**, 병원, 이름, 순서, 번만
신생아	**사망**, 가능성, 당국, **목동**, 정황	**감염**, **사망**, 건강, **세균**, 사람

외시키도록 하는 값을 의미한다.

Layer Size (layer) : 학습 데이터 안의 각 단어를 벡터화 하였을 때 단어를 표현하는 벡터의 차원수를 의미한다.

Learning Rate (learning) : 단어를 학습하기 위해 신경망에서 벡터의 가중치를 단계적으로 조정하는데 사용되는 값이다.

Iteration (iter) : 전체 학습의 반복 횟수를 결정하는 값이다.

Window Size (window) : 주변 단어와 특정 단어의 근접성을 통해 출현 확률을 계산하기 위하여 주변 단어의 범위를 지정하는 값이다.

Sub-Sampling (sample) : "in", "the", "a"와 같이 문맥에서 많이 반복되는 단어이지만 큰 의미를 갖지 않는 단어를 확률적으로 학습에서 제외시키기 위한 값이다.

Negative Sampling (neg) : 단어의 벡터 값을 업데이트 할 때 중심 단어와 근접 단어 그리고 근접하지 않은 단어를 negative sampling 크기만큼 추출하여 벡터 값을 업데이트 한다.

표 3. Word2Vec에 사용된 매개 변수 값
Table 3. Parameter values used in Word2Vec

학습 매개 변수	값
min-freq	**1**, 3, 5, 10, 15, 20, 30
layer	**100**, 200, 300, 400, 500, 700, 1000
learning	**0.025**, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3
iter	**1**, 5, 10, 30, 50, 100, 200
window	**5**, 10, 30, 50, 100, 200, 500
sample	**0**, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7
neg	**0**, 5, 10, 15, 20, 25, 30

3.3 학습 모델 성능 평가

학습 시킨 모델의 성능을 평가하기 위한 방법으로 정밀도(Precision)을 사용하며 식은 다음과 같다.

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (1)$$

여기서 TP는 해당 모델에서 이벤트 단어와 근접한 10개의 단어를 통해 검색한 트윗 중에서 이벤트 단어를 포함한 트윗의 개수이며, FP는 검색한 트윗 중에서 이벤트 단어를 포함하지 않는 트윗의 개수이다. 따라서 학습을 통해 이벤트 단어와 근접한 단어들이 실제 이벤트 단어가 포함된 트윗에 출현했는지 확인할 수 있다.

Ⅳ. 실험 및 분석

4.1 기본 매개 변수

Skip-gram, CBOW모델과 함께 [표 3]에서 굵게 표

시된 기본 학습 매개 변수 값을 사용하여 이벤트가 발생한 날의 트위터 데이터를 학습시키고 이벤트 단어와 근접한 5가지 단어를 추출한 후 근접한 순서대로 단어들을 나열하였다. 추출된 단어는 [표 2]와 같으며 굵게 표시된 단어는 이벤트와 연관성이 높다고 생각되는 단어이다. 학습된 두 모델에서 추출된 각 모델 별 75개의 단어들 중 실제 이벤트와 연관성이 있다고 판단되는 질병 관련 단어로 Skip-gram 모델에서 '간염', '사망', '진단' 등 7개, CBOW 모델에서 '간염', '사망, '방역' 등 6개를 찾을 수 있었다.

4.2 Min Frequency, Layer Size, Learning Rate

[그림 3]은 Min Frequency의 변화에 따른 정밀도의 변화를 보여준다. Min Frequency는 값이 증가할수록 출현 빈도가 적은 단어를 학습에서 제외시키기 때문에 의미 없는 단어의 학습을 감소시켜 학습에 소요되는 시간이 감소하였으나 학습 데이터가 적은 경우 주요 단어를 학습에서 제외시켜 성능에서 큰 향상을 보이지 않았다. Skip-gram 모델의 경우는 매개 변수 값이 증가할수록 정밀도가 소폭 감소하였으나 CBOW 모델의 경우는 정밀도가 소폭 증가하였다.

[그림 4]는 Layer size의 변화에 따른 정밀도의 변화를 보여준다. 두 모델 모두 매개 변수 값 변화에 따른 정밀도의 변화가 크지 않았다. Skip-gram 모델의 경우 layer size가 500일 때 가장 높은 성능을 보였고, CBOW 모델의 경우 layer size가 200일 때 가장 높은 성능을 보였다.

[그림 5]는 Learning rate의 변화에 따른 정밀도의 변화를 보여준다. 앞의 두 매개 변수의 조정과는 다르게 learning rate의 조정은 성능에 큰 영향을 미쳤다. 두 모델 모두 learning rate의 증가에 따라 성능이 증가하는 모습을 보였으나, Skip-gram의 경우 learning rate

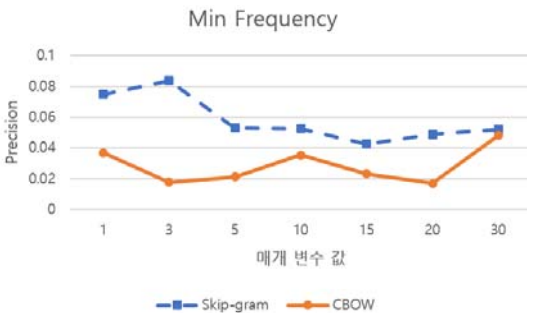


그림 3. Min frequency 값 조정에 따른 정밀도 변화
Fig. 3. Precision changes as a function of min frequency

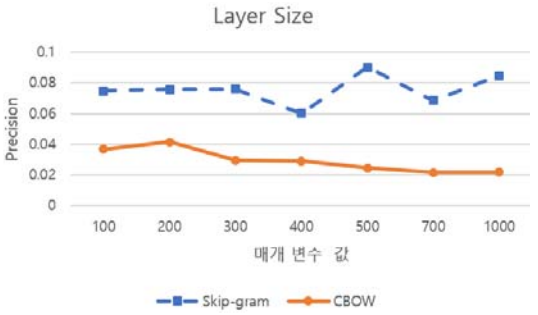


그림 4. Layer size 값 조정에 따른 정밀도 변화
Fig. 4. Precision changes as a function of layer size

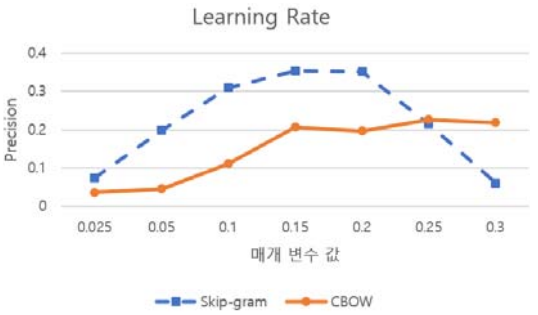


그림 5. Learning rate 값 조정에 따른 정밀도 변화
Fig. 5. Precision changes as a function of learning rate

가 0.15일 때 가장 높은 성능을 보인 후 급격히 성능이 하락하였으나, CBOW 모델은 매개 변수 값의 증가에 따라 꾸준히 성능이 향상되었다.

4.3 Iteration, Window Size

[그림 6]은 Iteration의 변화에 따른 정밀도의 변화를 보여준다. Iteration의 증가는 학습의 반복 횟수 증가를 의미하기 때문에 성능의 증가를 쉽게 예측할 수 있지만 학습에 소비되는 시간을 증가시키기 때문에 너무 많은 반복 횟수는 학습에 소요되는 시간을 급격하게 증가시킬 수 있다. Skip-gram 모델의 경우 예상된 대로 iteration의 증가에 따라 성능이 향상되는 모습을 보였으나 itearation이 30일 때 이후로는 성능 향상이 정체되는 모습을 보였다. CBOW 모델의 경우 itearation이 30일 때 까지는 Skip-gram 모델과 동일하게 성능이 향상되는 모습을 보였으나, 이후로는 오히려 성능이 감소하는 모습을 보였다.

[그림 7]은 Window size의 변화에 따른 정밀도의 변화를 보여준다. Window size가 증가할수록 각 단어마다 학습해야하는 주변 단어가 많아지기 때문에 전체적인 학습량이 증가하는데, Skip-gram 모델의 경우 많아진 학습량과 비례하여 성능이 지속적으로 향상되는

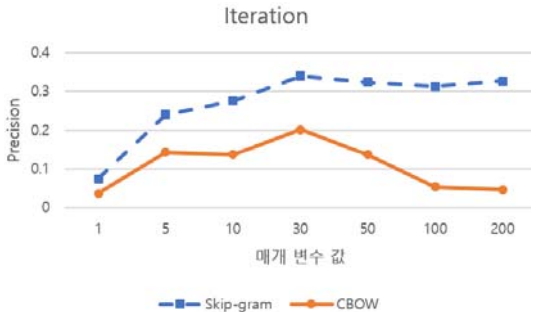


그림 6. Iteration 값 조정에 따른 정밀도 변화
Fig. 6. Precision changes as a function of iteration

그림 7. Window size 값 조정에 따른 정밀도 변화
Fig. 7. Precision changes as a function of window size

모습을 보였다. 하지만 CBOW 모델의 경우 학습량이 많아졌음에도 불구하고 성능에 큰 영향을 미치지 못하였고 오히려 성능이 소폭 감소하는 모습을 보였다.

4.4 Sub-Sampling, Negative Sampling

[그림 8]은 Sub-Sampling의 변화에 따른 정밀도의 변화를 보여준다. CBOW 모델의 경우 sub-sampling 값의 변화에 따라 큰 성능의 차이를 보이지 않았지만 Skip-gram 모델의 경우 sub-sampling 값이 1e-2일 때 가장 높은 성능을 보인후 이후 성능이 급격히 하락하

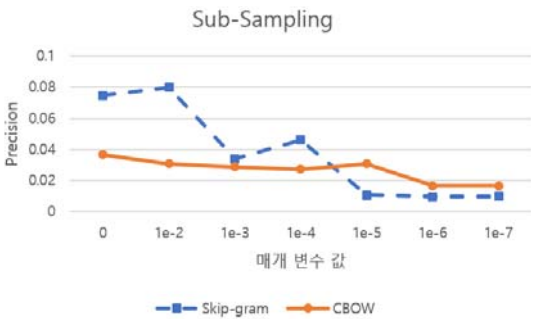


그림 8. Sub-sampling 값 조정에 따른 정밀도 변화
Fig. 8. Precision changes as a function of sub-sampling

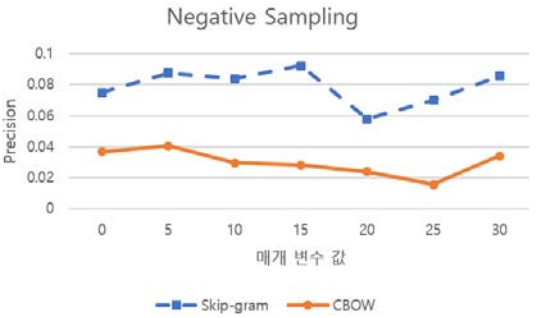


그림 9. Negative sampling 값 조정에 따른 정밀도 변화
Fig. 9. Precision changes as function of negative sampling

여 7가지 매개 변수 변화에 따른 word2vec의 성능에 대한 실험 중 가장 낮은 성능을 보였다.

[그림 9]는 Negative sampling의 변화에 따른 정밀도의 변화를 보여준다. Negative sampling의 증가는 window size와는 반대로 전체적인 학습량을 감소시키지만 Skip-gram 모델의 경우 negative sampling 값이 15일 때까지 성능이 향상되는 모습을 보였다. 반면 CBOW 모델의 경우 지속적으로 성능이 하락하는 모습을 보였다.

4.5 결과 분석

[표 4]는 실험을 통해 찾은 Word2Vec의 두 모델 중 높은 성능을 보인 모델과, 학습 매개 변수 별 가장 높은 성능을 보인 변수 값이다. Skip-gram과 CBOW 모델 중 평균적으로 높은 성능을 보인 모델은 Skip-gram 모델이었으며, Skip-gram 모델에서 min frequency는 3일 때, layer size는 500일 때, learning rate는 0.15일 때, iteration은 30일 때, window size는 500일 때, sub-sampling은 1e-2일 때, negative sampling은 15일 때 각 학습 매개 변수 별로 가장 높은 정밀도를 보였다. [그림 10]은 Skip-gram 모델에서 각 학습 매개 변수 별로 가장 높은 성능을 보인 매개 변수를 적용하여

표 4. 가장 좋은 성능을 보인 학습 모델과 학습 매개 변수 값
Table 4. The best-performing learning model and learning parameter values

학습 모델	Skip-gram
min-freq	3
layer	500
learning	0.15
iter	**30**
window	500
sample	1e-2
neg	15

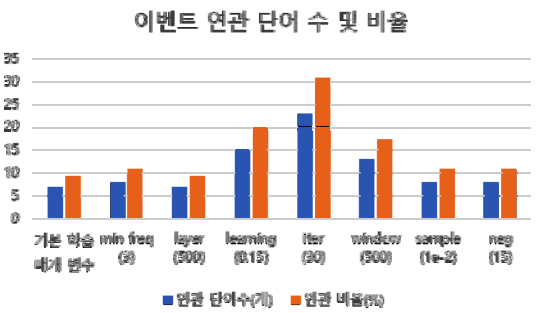


그림 10. 기본 학습 매개 변수와 [표 4]에 표시된 학습 모델 및 학습 매개 변수를 적용하여 추출된 단어들 중 이벤트와 연관성이 있는 단어의 수와 단어의 비율
Fig. 10. The number of words related to the event and the ratio of the words among the extracted words by applying the default learning parameters and the learning model and learning parameters shown in [Table 4]

학습시킨 word2vec 모델에서 각 날짜별 이벤트 단어와 근접한 5개의 단어들을 합한 75개의 단어들 중 이벤트와 연관성이 있다고 판단되는 단어의 개수 및 비율을 나타낸 것이다. 기본 학습 매개 변수를 사용하여 학습한 word2vec 모델에서 연관 단어는 7개, 연관 비율은 9.33%이었다. min frequency, layer size, sub-sampling, negative sampling에서 가장 높은 성능을 보인 값들을 사용한 word2vec 모델들은 기본 학습 매개 변수를 사용한 모델과 큰 차이를 보이지 않았으나, learning rate, iteartion, window size에서 가장 높은 성능을 보인 값들을 사용했을 때는 기본 학습 매개 변수를 사용한 모델보다 성능이 큰 폭으로 향상된 것을 확인할 수 있다. 이 중 가장 높은 성능을 보였을 때는 iteration이 30일 때이며, 해당 변수를 사용한 word2vec 모델에서 연관 단어는 23개, 연관 비율은 30.67%로 증가했다. [표 5]는 Skip-gram모델에서 7가지 학습 매개 변수 중 가장 높은 성능을 보인 itearation을 30으로 하고 나머지 변수를 기본 값으로 설정하여 학습시킨 후 해당 모델에서 이벤트 단어와 근접한 순서대로 5개의 단어를 나열하고 실제 이벤트와 연관성이 있을 것으로 판단되는 단어는 굵게 표시한 것이다. '맥도날드', '일본', '생리대', '인플루엔자' 단어는 해당 이벤트가 발생한 날짜에 수집된 트위터 데이터에서 질병과 무관한 광고성 트윗을 과도하게 학습하여 이벤트와 관련된 근접 단어를 추출하지 못하였다.

표 5. Skip-gram모델, iteration 값 30, 기본 학습 매개 변수를 사용하여 학습시킨 후 추출한 이벤트 근접 단어
Table 5. Extracted event near words after learning using skip-gram model, iteration value 30, default learning parameters

이벤트 단어	근접 단어
족발	**편육**, 햄버거, 소홀, 친환경, 나선
계란	**살충제**, 늘었다, 전국, 불매운동, 닭고기
소시지	**유럽**, 섭씨, **간염**, 중단, 고온
맥도날드	반미, 코카콜라, 깊은, 운동권, 정석
일본	무수, 쉬쉬, 지구인, 처절한, 정치인
생리대	활동가, 출범, 열림, 일회용, 담소
개미	**방역**, 당국, **살인**, 들어와, **비상**
진드기	**살인**, **중태**, **SFTS**, 나노, 항바이러스제
고래회충	**위벽**, **복통**, 울산, 유발, 서도
녹농균	병원, **사망**, 드물게는, **패혈증**, 뉴스
레지오넬라증	**솔샘**, **청송**, **온천**, 이용, 괴멸
인플루엔자	시사, 당부, 민우, 준수, 수칙
중국	인체, 본부, **가금류**, 자치구, **원난성**
오리	**영암**, **전남**, 반경, 당국, 속보
신생아	**사망**, 정황, **세균**, 중환자실, 부검

V. 결 론

기존 단어 임베딩의 효과적인 학습에 관련된 연구는 뉴스와 위키피디아와 같이 정형화된 데이터를 기반으로 연구가 진행되었기 때문에 최근 많은 연구가 진행되고 있는 트위터와 같은 비정형화 데이터에 적용되기는 어려웠다. 본 연구는 최신 단어 임베딩 기술인 Word2Vec을 사용하여 트위터 데이터를 기반으로 효과적인 단어 임베딩을 위한 학습 모델 선택 및 학습 매개 변수 조정을 연구하였다. 본 연구는 Word2Vec의 두 가지 학습 모델인 CBOW와 Skip-gram을 바탕으로 7가지 학습 매개 변수와 기본 값을 선정하고 각 매개 변수를 조정하며 성능을 분석하였다. 그 결과, Skip-gram모델의 성능이 CBOW모델 보다 더 좋은 성능을 가진다는 것을 확인하였고 다른 학습 매개 변수를 조정했을 때보다 학습 반복 횟수를 일정 횟수 증가시켰을 때 성능이 급격히 상승하는 것을 확인하였으며 이를 통해 트위터 데이터를 기반으로한 효과적인 Word2Vec학습에 필요한 학습 매개 변수 조정에 대한 지침을 제공한다.

References

[1] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning,"

in *Proc. 25th Int. Conf. Machine Learning*, pp. 160-167, Portland, OR, USA, May 2008.

[2] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137-1155, Feb. 2003.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Workshop 2013,* Scottsdale, Arizona, May 2013.

[4] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, pp. 1532-1543, Doha, Qatar, Oct. 2014.

[5] C. dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proc. COLING 2014, the 25th Int. Conf. Computational Linguistics: Technical Papers*, pp. 69-78, Dublin, Ireland, Aug. 2014.

[6] Y. Kim, "Convolutional neural networks for sentence classification," in *EMNLP 2014*, pp. 1746-1751, Doha, Qatar, Oct. 2014

[7] M. TH, S. Sahu, and A. Anand, "Evaluating distributed word representations for capturing semantics of biomedical concepts," in *Proc. BioNLP 15*, pp. 158-163, 2015.

[8] S. Lai, K. Liu, S. He, and J. Zhao, "How to generate a good word embedding," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 5-14, 2016.

[9] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo, "How to train good word embeddings for biomedical NLP," in *Proc. 15th Workshop on Biomed. Natural Lang. Process.*, pp. 166-174, Berlin, Germany, Aug. 2016.

[10] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *LREc*, vol. 10, pp. 1320-1326, 2010.

[11] B. Jang and J. Yoon, "Characteristics analysis of data from news and social network services," *IEEE Access*, vol. 6, pp. 18061-18073, 2018.

[12] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. B. Melton, "Semantic similarity and relatedness between clinical terms: an experimental study," in *AMIA Annu. Symp. Proc.*, vol. 2010, pp. 572-576, 2010.

[13] L. Finkelstein, et al., "Placing search in context: The concept revisited," in *Proc. 10th Int. Conf. World Wide Web*, pp. 406-414, Hong Kong, Hong Kong, May 2001.

[14] M. Sahlgren, "The distributional hypothesis," *Ital. J. Disabil. Stud.*, vol. 20, pp. 33-53, 2008.

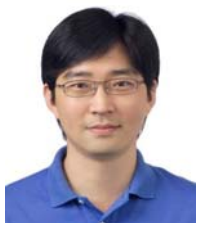
[15] H. Jung, J. Bae, S. Hong, C. Park, and M. Song, "Analysis of twitter public opinion in different political views : A case study of sewol ferry accident," *Korean J. Journalism & Commun. Stud.*, vol. 60, no. 2, pp. 269-302, Apr. 2016.

김 인 환 (Inhwan Kim)

2016년 3월~현재 : 상명대학교 컴퓨터과학과 학석사 연계과정

<관심분야> 인공지능, 빅데이터, 자연어 처리

장 백 철 (Beakcheol Jang)



2001년 2월 : 연세대학교 컴퓨터과학과 학사

2002년 8월 : 한국과학기술원 컴퓨터과학과 석사

2009년 8월 : North Carolina State University, 컴퓨터과학과 박사

2012년~현재 : 상명대학교 컴퓨터과학과 교수

<관심분야> 무선 네트워크, 사물 인터넷, 빅데이터, 인공지능