Word2vec 모델의 단어 임베딩 특성 연구

강형석, 양장훈
서울미디어대학원 대학교 뉴미디어학부
e-mail: jiangjs@naver.com, jhyang@smit.ac.kr

On Characteristics of Word Embeddings by the Word2vec Model

Hyungsuc Kang, Janghoon Yang
Dept. of New Media, Seoul Media Institute of Technology

요 약

단어 임베딩 모델 중 현재 널리 사용되는 word2vec 모델은 언어의 의미론적 유사성을 잘 반영한다고 알려져 있다. 본 논문은 word2vec 모델로 학습된 단어 벡터가 실제로 의미론적 유사성을 얼마나 잘 반영하는지 확인하는 것을 목표로 한다. 즉, 유사한 범주의 단어들이 벡터 공간상에 가까이 임베딩되는지 그리고 서로 구별되는 범주의 단어들이 뚜렷이 구분되어 임베딩되는지를 확인하는 것이다. 간단한 군집화 알고리즘을 통한 검증의 결과, 상식적인 언어 지식과 달리 특정 범주의 단어들은 임베딩된 벡터 공간에서 뚜렷이 구분되지 않음을 확인했다. 결론적으로, 단어 벡터들의 유사도가 항상 해당 단어들의 의미론적 유사도를 의미하지는 않는다. Word2vec 모델의 결과를 응용하는 향후 연구에서는 이런 한계점에 고려가 요청된다.

1. 서론

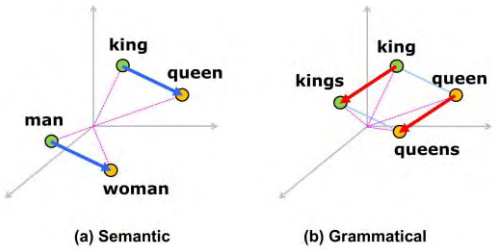
다양한 제어 시스템에 음식 인식 기반의 서비스가 도입되면서, 자연어 처리에 대한 중요성이 증가하고 있다. 자연어 처리를 위해 단어를 벡터 공간상의 실수 벡터로 매핑하는 단어 임베딩(word embedding)은 기본적으로, 비슷한 의미의 단어는 비슷한 문맥에서 등장한다는 언어학의 분산 가설(distributional hypothesis)[1]에 근거한다. 최근 단어 임베딩 모델로 각광받고 있는 word2vec 모델[2, 3]도 동일한 가설에 근거해서 단어 벡터를 결정한다. 즉, 비슷한 문맥(앞뒤에 존재하는 단어들)에서 등장하는 단어들은 word2vec 모델에 의해 벡터 공간상의 가까운 위치에 임베딩된다.

한국어 감성어 분류를 위한 기초 연구를 위해, word2vec 모델로 학습된 한국어 단어 중 감성어(e.g. 놀람, 기쁨, 슬픔 등)가 벡터 공간상에서 감성 차원(e.g. 긍정-부정)에 따라 군집화 가능한지를 확인해 본 바 있다. 이 기초 연구의 결과, word2vec 모델로 학습된 감성어는 감성 차원에 따른 군집화가 불가능했다. 즉, 긍정적인 감성을 표현하는 단어와 부정적인 감성을 표현하는 단어가 벡터 공간에서 구분되어 임베딩되지는 않는 것으로 보인다.

이 결과에서 착안해서, 본 연구는 word2vec 모델로 학습된 단어들이 벡터 공간상에 어떻게 군집화되어 임베딩되는지를 확인하고자 한다. Word2vec 모델은 분산 가설을 따르므로, 비슷한 단어는 벡터 공간상의 가까운 위치에 임베딩될 것이다. 하지만 실제로 유사한 단어들이 벡터 공간상에 얼마나 군집화되는지는 명확하지 않다. 본 연구의 목적은 word2vec 모델로 학습된 단어 임베딩의 군집화 특성을 분석하고 그런 특성이 나타나는 이유를 설명하는 것이다.

2. 연구 가설

Word2vec 모델의 유추 검사(analogy test)는 특정 단어 쌍의 의미론적/문법적 유추 관계를 확인하는 검증 방법이다. 즉, (그림 1)에서처럼 남녀 관계에 해당하는 단어 쌍(king-queen, man-woman)이 벡터 공간에서 동일한 거리와 방향으로 떨어져 있는지, 단수 복수 관계에 해당하는 단어 쌍(king-kings, queen-queens)이 동일한 거리와 방향으로 떨어져 있는지를 확인하는 것이 유추 검사이다. 특정 범주의 단어 쌍에 속하는 단어들이 이런 유추 관계에 부합하도록 벡터 공간에 임베딩된다면, 해당 임베딩 모델은 해당 단어 쌍의 유추 관계를 잘 반영했다고 볼 수 있다.



(그림 1) 단어 벡터의 의미론적/문법적 관계

하지만 이 유추 검사는 단어 쌍의 관계((그림 1)에서 파란색과 빨간색 벡터가 이 관계에 해당하고, 편의상 이 벡터를 '유추 벡터'라고 명명하겠음)를 확인하는 것이지, 각 범주의 단어(e.g. 남자 범주의 king 및 man, 여자 범주의 queen 및 woman)가 벡터 공간에서 얼마나 가까운지는 확인하지 않는다. 다시 말해, (그림 1)에서 단어 벡터 king과 man 사이의 거리는 유추 검사의 고려 대상이 아니다. Wor2vec 모델이 분산 가설을 잘 따른다면, 단어 벡터 king은 단어 벡터 man보다 단어 벡터 queen에 더 가까이 임베딩될 것이다. 왜냐하면, king과 queen이 비슷한 문맥에서 등장할 확률이 king과 man이 비슷한 문맥에서 등장할 확률이 더 높을 것이지 때문이다. 또한 영어 명사의 단수형과 복수형은 일반적으로 비슷한 문맥에서 등장할 확률이 높을 것이므로, (그림

1)의 빨간색 유추 벡터의 크기는 상당히 짧을 것이라고 예상할 수 있다. 따라서 (그림 1)의 6개 단어 벡터를 2개의 클러스터(cluster)로 군집화한다면, 하나의 클러스터는 king, kings, queen, queens로 구성되고 나머지 클러스터는 man 및 woman으로 구성될 가능성이 높다. 본 연구를 통해, 이런 연구 가설이 사실인지를 검증하고자 한다.

3. 데이터 및 방법

한국어 word2vec 모델로 학습된 단어 임베딩의 군집화 특성을 확인하기 위해 적절한 샘플 단어를 선정해야 한다. 이전 연구에서 한국어 word2vec 모델의 유추 검사를 위해 개발한 KATS(Korean Analogy Test Set) [4]의 일부 범주를 본 연구에서 사용하고자 한다. 즉, KATS의 범주 중 '국가', '수도', '언어', '통화', '남자' 및 '여자' 범주에 속하는 단어를 군집화 샘플로 사용했다. 이 6개 범주에 더해, 감정을 표현하는 단어도 군집화 샘플에 추가해다. <표 1>은 군집화에 사용된 샘플 단어(총 125개)의 목록을 보여준다(밑줄 친 단어는 각 범주를 대표하는 단어이다).

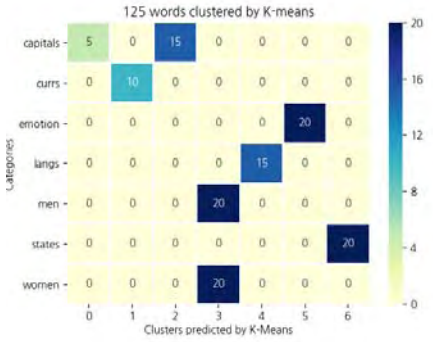
<표 1> 군집화에 사용된 샘플 단어

범주 (개수)	단어
국가 (20)	국가, 한국, 일본, 중국, 베트남, 태국, 인도네시아, 필리핀, 인도, 이란, 이라크, 이집트, 그리스, 러시아, 미국, 영국, 프랑스, 독일, 이탈리아, 스페인
수도 (20)	수도, 서울, 도쿄, 베이징, 하노이, 방콕, 자카르타, 마닐라, 뉴델리, 테헤란, 바그다드, 카이로, 아테네, 모스크바, 워싱턴, 런던, 파리, 베를린, 로마, 마드리드
언어 (15)	언어, 한국어, 일본어, 중국어, 힌디어, 아랍어, 그리스어, 러시아어, 영어, 프랑스어, 독일어, 이탈리아어, 스페인어, 포르투갈어, 태국어
통화 (10)	통화, 원, 엔, 위안, 달러, 유로, 루블, 페소, 루피, 크로네
남자 (20)	남자, 남성, 아버지, 아빠, 아들, 형, 오빠, 남동생, 할아버지, 손자, 신랑, 남편, 소년, 왕, 왕자, 삼촌, 아저씨, 신사, 아비, 아범
여자 (20)	여자, 여성, 어머니, 엄마, 딸, 누나, 언니, 여동생, 할머니, 손녀, 신부, 아내, 소녀, 여왕, 공주, 숙모, 아주머니, 숙녀, 어미, 어멈
감정 (20)	감정, 놀람, 경악, 행복, 발랄, 환희, 분노, 두려움, 짜증, 좌절, 슬픔, 우울, 침울, 따분, 싫증, 기쁨, 차분, 느긋, 고요, 만족

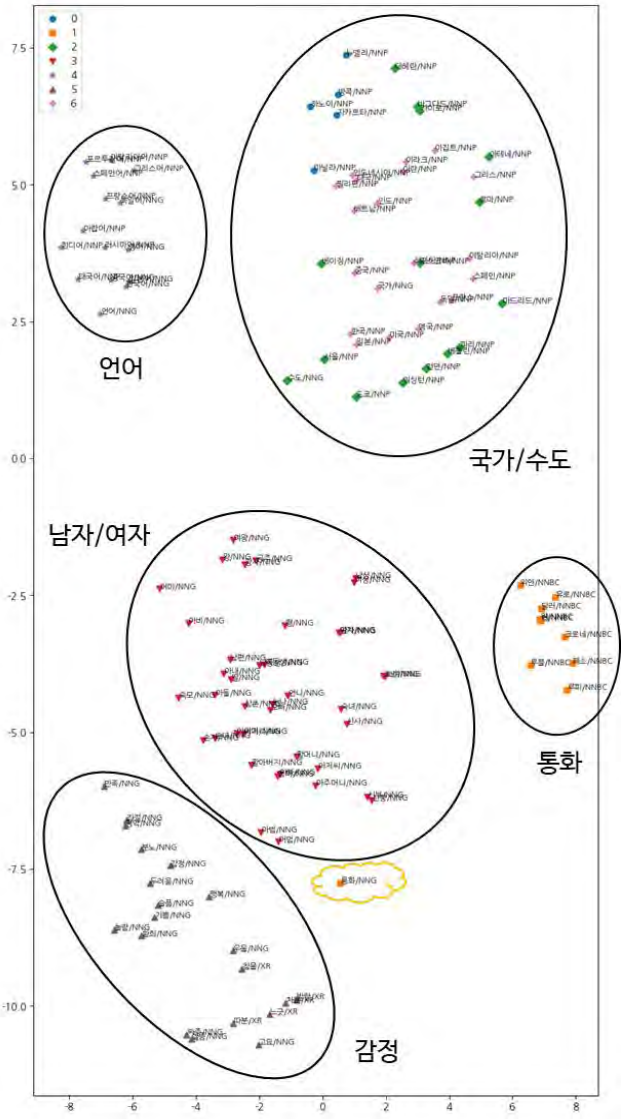
위 샘플 단어의 단어 벡터는 이전 연구[4]의 word2vec 모델에서 차용했다. 사용된 형태소 분석기/품사 태거는 파이썬 패키지 KoNLPy[5]에서 제공하는 mecab 태거이었고, word2vec 모델의 학습에 사용된 한국어 말뭉치는 나무위키(https://namu.wiki/)와 위키백과(https://ko.wikipedia.org/)의 대용량 덤프 파일이었다. Word2vec 모델링을 위해 gensim 라이브러리[6]를 사용했고, 이때 사용된 하이퍼파라미터로 벡터 차원은 300, 학습 알고리즘은 skip-gram 그리고 윈도 크기는 10이었다. 이렇게 구해진 총 125개의 단어 벡터를 K-means 알고리즘으로 군집화했다.

4. 결과 및 분석

(그림 2)는 7개 범주에 속하는 총 125개 단어 벡터에 대한 K-means 군집화의 결과를 보여주는 교차 분석표(cross table)이다. (그림 3)은 300 차원의 벡터 공간에 임베딩된 125개 단어 벡터를 t-SNE(Stochastic Neighbor Embedding) 차원 축소 알고리즘을 이용해서 2차원으로 시각화한 분포도이다.



(그림 2) 7 개 범주의 샘플 단어에 대한 군집화 결과



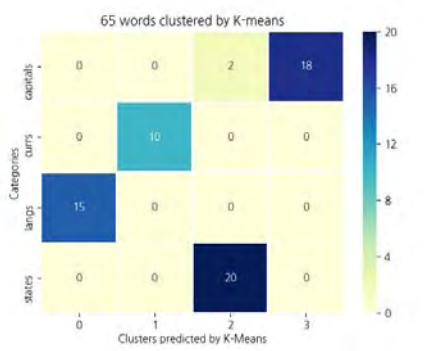
(그림 3) 7 개 범주의 샘플 단어에 대한 t-SNE 분포도

우선 (그림 2)와 (그림 3)에서 확인할 수 있듯이, 통화/언어/감정/국가 범주에 속하는 단어는 K-means 알고리즘에 의해 각각 클러스터 1/4/5/6으로 군집화되었다. 그리고 수도 범주에 속하는 단어는 클러스터 0/2로 나누어져 군집화되었고, 남자 및 여자 범주는 클러스터 3으로 함께 군집화되었다. 일단 전반적으로 보면, 각 범주에 속하는 단어는 300차원의 벡터 공간에서 어느 정도 군집화되어 임베딩되었다고 볼 수 있다. 수도 범주의 단어가 2개의 클러스터로 분할되는 탓에, 남자 및 여자 범주가 하나의 클러스터로 합쳐졌지만, (그림 3)의 분포도에서 직관적으로 확인할 수 있듯이 전반적으로 비슷한 단어들은 벡터 공간상의 가까운 위치에 존재한다고 볼 수 있다.

(그림 3)에서 주목할 만한 점은 통화 범주에 속하는 대표 단어 '통화/NNG'[1]가 나머지 단어와 통일한 클러스터로 군집화되긴 했지만, t-SNE 분포도에서는 꽤 멀리 떨어져 있다는 점이다. 이런 현상에 대해 가능한 한 가지 설명은 해당 단어가 동음이의어(즉, 통화는 currency뿐만 아니라 phone call이라는 의미도 있음)이라는 점이다. 일반적으로 단어 임베딩에서 동음이의어가 서로 구분되어 처리되지 않으면, 해당 단어는 벡터 공간상에 제대로 임베딩되지 않는 문제가 존재한다.

(그림 3)에서 국가/수도 범주와 남자/여자 범주는 나머지 3 범주에 비해 덜 확연하게 구분됨을 확인할 수 있다. 비록 국가/수도 범주에 속하는 단어들을 자세히 살펴보면, 국가 범주에 속하는 단어는 가운데 부분에 그리고 수도 범주에 속하는 단어는 가장자리 부분에 분포되어 있어서 어느 정도 구분된다고 볼 수는 있지만, 남자 및 여자 범주에 속하는 단어들은 이런 구분이 불가능하다. 오히려 남자-여자의 유추 관계에 속하는 단어 쌍들은 서로 가까이 존재함(예를 들어, (그림 3)에서 '남자' 및 '여자', '소년' 및 '소녀'는 거의 붙어 있어서 구별조차 어려울 정도임)을 확인할 수 있다.

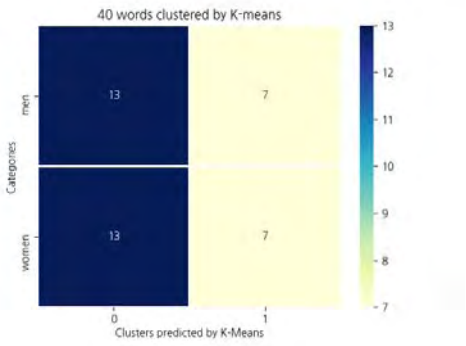
여기서 한 가지 주의할 점은 군집화의 결과는 군집화에 사용된 샘플의 구성에 따라 달라질 수 있다는 점이다. 실험 과정에서 한두 단어의 추가와 삭제로 군집화의 결과가 달라짐을 확인한 바 있었다. 그래서 수도 범주가 2가지 클러스터로 군집화되고, 남자/여자 범주가 1가지 클러스터로 군집화되는 현상을 좀더 면밀히 살펴보기 위해서, 군집화 샘플 단어의 구성을 변경해 보았다.



(그림 4) 국가/수도/언어/통화 범주의 샘플 단어에 대한 군집화 결과

[1] (그림 3)에서 단어 뒤에 표시된 약어는 품사를 의미한다. 즉, NNG 는 보통명사, NNP 는 고유명사, NNBC 는 단위성 의존명사, XR 은 어근을 의미한다.

우선 (그림 4)는 국가/수도/언어/통화의 4가지 범주에 속하는 단어만으로 샘플을 구성한 후 군집화한 결과를 나타낸다. 그 결과 수도 범주에 속하는 '로마' 및 '마드리드"라는 2개 단어를 제외하고, 나머지 단어들은 원래 범주에 해당하는 단어들끼리 제대로 군집화되었다. 특히 '로마'라는 단어가 국가 범주로 군집화된 이유는 동음이의어의 경우와 유사하다고 볼 수 있다. '로마'라는 단어는 현재 이탈리아의 수도인 도시를 의미하기도 하지만, 많은 경우 고대 로마 제국을 의미하기도 하므로, word2vec 모델링 과정에서 국가라는 범주로 학습되었을 가능성이 있다.



(그림 5) 남자/여자 범주의 샘플 단어에 대한 군집화 결과

(그림 5)는 남자/여자의 2가지 범주에 속하는 단어만으로 군집화한 결과이다. 총 20쌍의 단어 중 클러스터 1로 군집화된 단어 쌍은 남자-여자, 남성-여성, 신랑-신부, 소년-소녀, 왕-여왕, 왕자-공주, 신사-숙녀의 7쌍이다. 클러스터 0로 군집화된 나머지 13쌍의 단어들과 비교해 보면, 대체적으로 클러스터 0는 가족 관계의 명칭이고 클러스터 1은 가족 관계의 명칭이 아님을 확인할 수 있다. 물론 신랑-신부, 왕-여왕 및 왕자-공주의 단어 쌍도 가족 관계 명칭이라고 볼 수는 있지만, 나머지 클러스터 0의 단어에 비해 가족 관계성이 떨어진다고 볼 수 있다. 그리고 분산가설을 고려해 보면, 이 3쌍의 단어들이 클러스터 0에 속하는 단어들고 비슷한 문맥에서 발생할 가능성이 적다고도 볼 수 있다. 요약하면, 예상과는 달리 남자/여자 범주에 속하는 단어들은 벡터 공간상에서 남자 및 여자로 군집화되기보다는 가족 관계 명칭 여부에 따라 군집화된다고 볼 수 있다.

이상의 군집화 결과를 종합해 보면, 언어/통화/감정에 속하는 단어들은 범주별로 아주 뚜렷이 구분되어 군집화된다. 이는 각 범주에 속하는 단어들은 서로 비슷한 문맥에서 발생할 가능성이 높지만, 한 범주에 속하는 단어가 나머지 범주에 속하는 단어와 비슷한 문맥에서 발생할 가능성은 아주 작다는 의미로 해석할 수 있다. 예를 들어, 학습에 사용된 말뭉치에서 '한국어'와 '일본어'라는 단어가 비슷한 문맥에서 등장할 가능성은 크지만, '한국어'라는 단어가 '달러'나 '놀람'이라는 단어와 비슷한 문맥에서 등장할 가능성은 매우 작을 것이다.

대조적으로 국가/수도 및 남자/여자 범주의 구분은 뚜렷하지 않은데, 그 이유도 마찬가지 방식으로 설명할 수 있다. 즉, '한국'과 '서울'이라는 단어는 비슷한 문맥에서 발생할 가능성이 크고, '소년'과 '소녀'도 비슷한 문맥에서 발생할 가능성이 크다. 따라서 (그림 3)에서 확인할 수 있듯이, 국가/수도 범주에 속하는 단어 쌍들은 나머지 단어 쌍에 비해 가까이 위치하고, 남자/여자 범주에 속하는 단어 쌍들도 매우 가까이 위치하게 된다. (그림 1)에서 언급한 유추 벡터의 관점에서 설명하면, 국가-수도 단어 쌍과 남자-여자 단어 쌍

의 유추 벡터는 국가-언어 단어 쌍과 국가-통화 단어 쌍의 유추 벡터에 비해 훨씬 짧다고 할 수 있다. 그 결과 나머지 3가지 범주에 비해, 국가/수도 및 남자/여자 범주는 벡터 공간에서 범주별로 잘 구분되어 임베딩되지 않는 것으로 볼 수 있다.

5. 결론

특정 범주에 속하는 100여 개 단어에 대한 K-means 군집화를 통해, 본 연구는 word2vec 모델에 의한 단어 임베딩의 군집화 특성을 확인했다. 예상했던 대로, 비슷한 단어들은 벡터 공간상의 가까운 위치에 임베딩됨을 확인했다. 하지만 좀더 면밀히 면밀히 살펴보면, 이질적인 범주의 단어들은 벡터 공간에서 뚜렷이 구분되어 군집화되지만, 일상 언어 생활에서 서로 다른 범주로 인식되는 일부 범주의 단어들은 뚜렷이 구분되어 군집화되지는 않음을 확인할 수 있었다. 즉, 의미론적으로 대조되는 단어들이 가까이 임베딩되는 경우가 있었다.

따라서 word2vec 모델을 통해 벡터 공간에 임베딩된 단어 벡터는 항상 의미론적 유사성이라는 기준으로 군집화되는 것은 아니다. 얼마나 비슷한 문맥에서 특정 단어들이 자주 등장하는가에 따라, 해당 단어들의 상대적 위치가 결정되는 것이다. 심지어 동일한 문맥에서 자주 등장하는 단어라 할지라도(e.g. '소년' 및 '소녀'), 일상 언어 생활에서는 동의어로 인지되지 않을 수 있다. 결국, 단어의 의미론적 유사성을 잘 반영하는 것으로 알려진 word2vec 모델에도 어느 정도의 한계가 존재함을 확인할 수 있다. 그러므로 word2vec 모델로 학습된 단어 벡터를 활용하는 연구에서 이런 한계를 고려하여 관련된 알고리즘이나 서비스의 개발이 요구된다.

6. 사사

이 논문은 2019년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(과제번호: NRF-2017R1A2B4007398)

참고문헌

[1] Sahlgren, Magnus. "The distributional hypothesis." *Italian Journal of Disability Studies* 20 (2008): 33-53.

[2] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv: 1301.3781*(2013).

[3] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

[4] 강형석, 양장훈. "한국어 단어 임베딩 모델의 평가에 적합한 유추 검사 세트." *한국디지털콘텐츠학회 논문지* 19.10 (2018): 1999-2008.

[5] 박은정, 조성준. "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지." *제26회 한글 및 한국어 정보처리 학술대회 논문집*(2014).

[6] Rehurek, Radim, and Petr Sojka. "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 2010.