한국컴퓨터정보학회 동계학술대회 논문집 제30권 제1호 (2022. 1)

차원감소 단어벡터 시각화를 통한 어휘별 관계 분석

고광호*, 백주련O
*평택대학교 스마트자동차학과,
O평택대학교 데이터정보학과
e-mail: kwangho@ptu.ac.kr*, jrpaik@ptu.ac.krO

Analysis of Vocabulary Relations by Dimensional Reduction for Word Vectors Visualization

Kwang-Ho, Ko*, Juryon, PaikO
*Dept. of Smart Automobile, Pyeongtaek University,
ODept. of Data Information & Statistics, Pyeongtaek University

● 요 약 ●

LSTM과 같은 딥러닝 기법을 이용해 언어모델을 얻는 과정에서 일종의 부산물로 학습 대상인 말뭉치를 구성하는 어휘의 단어벡터를 얻을 수 있다. 단어벡터의 차원을 2차원으로 감소시킨 후 이를 평면에 도시하면 대상 문장/문서의 핵심 어휘 사이의 상대적인 거리와 각도 등을 직관적으로 확인할 수 있다. 본 연구에서는 기형도의 시(詩)을 중심으로 특정 작품을 선정한 후 시를 구성하는 핵심 어휘들의 차원 감소된 단어벡터를 2D 평면에 도시하여, 단어벡터를 얻기 위한 텍스트 전처리 방식에 따라 그 거리/각도가 달라지는 양상을 분석해 보았다. 어휘 사이의 거리에 의해 군집/분류의 결과가 달라질 수 있고, 각도에 의해 유사도/유추 연산의 결과가 달라질 수 있으므로, 평면상에서 핵심 어휘들의 상대적인 거리/각도의 직관적 확인을 통해 군집/분류 작업과 유사도 추천/유추 등의 작업 결과의 양상 변화를 확인할 수 있었다. 이상의 결과를 통해, 영화 추천/리뷰나 문학작품과 같이 단어 하나하나의 배치에 따라 그 분위기와 정동이 달라지는 분야의 경우 텍스트 전처리에 따른 거리/각도 변화를 미리 직관적으로 확인한다면 분류/유사도 추천과 같은 작업을 좀 더 정밀하게 수행할 수 있을 것으로 판단된다.

키워드: 단어벡터(Word vector), 딥러닝(Deep learning), 분류(Classification), 유사도(Similarity)

I. Introduction

딥러닝 기법으로 언어모델을 생성하는 방식에는 단어의 전후 인접 여부만 평가하는 CBOW(Continuous Bag of Words)과 후방에 이어지는 단어들의 순서를 기준으로 학습하는 LSTM(Long-Short Term Memory) 등이 있다[1]. 인접 여부와 단어 순서를 학습하는 이러한 방식에 의해 해당 말뭉치(corpus)를 구성하는 어휘(vocabulary)들의 임베딩(embedding) 과정이 진행되고, 학습 종료 후에는 각 어휘별로 단어벡터(word vector)를 얻을 수 있다[2]. 일반적으로 수백 차원의 단어벡터가 생성되는데 이를 주성분 분석(PCA : Principal Component Analysis)이나 특잇값 분해(SVD : Singular Value Analysis) 등의 기법으로 주요 특성을 보존하면서도 2차원 수준으로 줄일 수 있다[3]. 이렇게 2D 수준으로 줄이고 나면 해당 텍스트에서 주요 어휘들의 상대적인 위치와 방향 등을 평가할 수 있다. 즉, 2D 평면에 위치시킨 각 어휘들의 상관관계를 가시화하여 직관적으로 각 어휘들의 기능과 역할 및 연관성 등을 비교할 수 있다.

이러한 접근법은 특히 영화에 대한 후기/리뷰나 문학작품처럼 하나하나의 어휘들이 환기시키는 분위기와 정동(affection)이 중요한 분야에서 유용할 수 있다. 본 연구에서는 이러한 접근법의 유용성을 평가하기 위한 하나의 시도로써 기형도의 시를 분석하고자 한다. 80년대 우울한 도시 풍경을 잘 묘사한 작품으로 유명한 기형도 작가는 단 한권의 시집을 유작으로 남겼다(『입 속의 검은 잎』 기형도, 문학과지성사, 1989년). 100페이지가 채 되지 않는 분량이기 때문에 본 연구의 분석 대상으로 적절한 것으로 판단된다. 유작의 전체 작품에 대해 LSTM 기법을 적용하여 얻을 수 있는 단어벡터를 사용하여 핵심 어휘들의 상대적인 위치와 거리 및 방향 등을 분석하였다.

특히 텍스트 전처리의 영향력을 평가하기 위해 원문의 조사와

어미 등을 그대로 살린 경우(Text #1)와 조사/어미 등을 삭제하고 동사나 형용사의 원형으로 처리한 경우(Text #2)에 해당하는 말뭉치를 각각 사용하였다. 각 경우 학습에 사용된 어휘 수와 말뭉치 수를 Table 1에 정리하였는데, 원형으로 대체하는 경우 말뭉치와 어휘 수의 비율이 3.74로 높아 언어모델 학습에 좀 더 유리함을 알 수 있다[4].

Table 1. Text Preprocessing Comparison

Text Preprocessing	Text #1	Text #2
No. of Vocabulary	4447	2688
No. of Corpus	11346	10059
Ratio of Corpus / Vocabulary	2.55	3.74

한글의 경우 조사와 어미에 따라 그 의미가 미묘하게 달라지는 교착어의 특성을 가지기 때문에 일반적으로 형태소 분석을 통한 단어 태깅(tagging)을 수행하여 분석 목적에 적당한 품사만을 선별하는 경우가 많다[5]. 본 연구에서는 이러한 품사 선별 과정과 정확하게 일치하지는 않더라도 그 전처리 양상이 달라지는 경우 주요 어휘들의 상대적인 위치와 방향을 이차원 평면에 도시하여 전처리 과정의 전체적인 효과를 정성적으로 파악할 수 있는 가능성을 모색해보고자 한다.

II. Result

본 연구에서 언어모델 생성을 위해 사용한 딥러닝 기법은 LSTM이다. 이에 적용한 신경망을 Fig.1에 도시하였다. 학습에 사용하는 연속된 단어(Time-series words)는 10개씩 적용되었고, 단어벡터 크기는 150차원이며, 드롭아웃(dropout ratio = 0.5)을 적용하여 과적합을 방지하였다. 학습 소요 시간을 줄이기 위해 임베딩과 어파인(Affine) 계층의 가중치는 같은 값을 적용하였다.

이렇게 얻은 단어벡터에 특잇값 분해법으로 차원 감소시켜 2D 평면상에서 주요 어휘들을 도시하였다. 우선, 주요 어휘에 해당하는 핵심 시어(詩語)의 선택이 필요한데, 선행 연구에 의하면 기형도 시의 핵심 이미지는 크게 3가지로 나눌 수 있다[6]. 이는 '도시화/빈민화', '관료적 인간의 우울/공감', '유년에 대한 향수' 등인데, 각 이미지에 해당하는 주요 어휘로는 '도시/공장', '사무실/서류', '엄마/누이' 등을 제시할 수 있다.

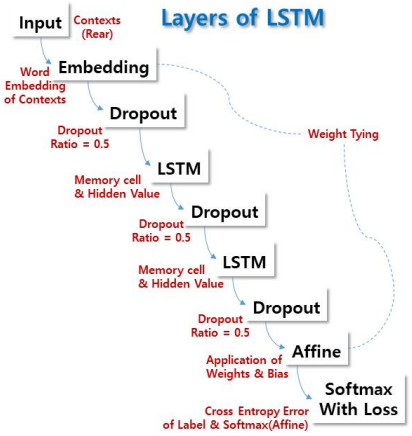

Fig. 1. Layers of LSTM Model

이중 특히 자본주의 관료체제에서 우울한 인간의 형상과 이에 대한 공감을 잘 표현한 「기억할만한 자나침」이라는 작품의 핵심 어휘에 대한 단어벡터의 가시화를 시도하였다. 참고로 해당 시의 전문을 전처리 방식에 따라 각각 아래에 정리하였다. 특히 Text#2 방식처럼 조사/어미 등을 제거하고 동사/형용사의 원형으로 대치하는 경우 시적 의미와 분위기 등은 거의 사라지는 것을 알 수 있다.

「기억할만한 자나침」

그리고 나는 우연히 그곳을 지나게 되었다 눈은 퍼부었고 거리는 캄캄했다 움직이지 못하는 건물들은 눈을 뒤집어쓰고 희고 거대한 서류뭉치로 변해갔다 무슨 관공서였는데 희미한 불빛이 새어나왔다 유리창 너머 한 사내가 보였다 그 춥고 큰 방에서 그는 혼자 울고 있었다 눈은 퍼부었고 내 뒤에는 아무도 없었다 침묵을 달아나지 못하게 하느라 나는 거의 고통스러웠다 어떻게 해야 할까 나는 중지시킬 수 없었다 나는 그가 울음을 그칠 때까지 창밖에서 떠나지 못했다 그리고 나는 우연히 지금 그를 떠올리게 되었다 밤은 깊고 텅 빈 사무실 창밖으로 눈이 퍼붓는다 나는 그 사내를 어리석은 자라고 생각하지 않는다 - ***(Text #1 : 조사/어미를 그대로 살리는 경우)***

그리고 나 우연하다 그곳 지나다 되다 눈 퍼붓다 거리 캄캄하다 움직이다 못하다 건물들 눈 뒤집어쓰다 희다 거대하다 서류뭉치 변하다 무슨 관공서 희미하다 불빛 새다 유리창 너머 한 사내 보이다 춥다 크다 방 혼자 울다 있다 눈 퍼붓다 내 뒤 아무 없다 침묵 달아나다 못하다 하다 나 거의 고통 어떻다 하다 나 중지시키다 없다 나 울다 그치다 때 창밖 떠나다 못하다 그리고 나 우연하다 지금 떠올리다 되다 밤 깊다 텅 비다 사무실 창밖 눈 퍼붓다 사내 어리석다 생각하다 않다 - ***(Text #2 : 조사/어미를 삭제하고 원형으로 대체)***

이 시의 내용을 간략하게 살펴보면, 화자가 눈 내리는 밤에 우연히 발견한 관공서처럼 보이는 건물의 사무실에서 홀로 울고 있는 남자를 보게 된다. 놀랍고도 측은한 마음으로 그 남자를 오랫동안 조용히 지켜보게 된 이후, 어느 날 비슷한 사무실에 앉아 창밖으로 눈 내리는 밤풍경을 바라보던 화자가 그 남자를 떠올리며 기이한 공감을 느끼게 된다. 눈을 뒤집어쓴 건물을 '희고 거대한 서류뭉치'로 표현하였고, 그 건물은 '관공서'였으며, 여자도 아니고 여러 명도 아닌 한명의 '남자'가 '울고' 있는 것을 발견했고, 이후 비슷한 상황에 처하게 된 화자가 남자를 떠올리며 '어리석지 않다'고 생각하는 것이다.

울고 있던 남자와 지켜보는 화자는 모두 관료주의적인 사회 시스템에서 괴로움을 겪게 되는 전형적인 카프카형 인간이라고 할 수 있다. 이는 목표 없이 부유하지만 높은 공감력 때문에 아름답고도 어리석은 인간들이라고 밴야민이 묘사한 인간형이다[7]. 따라서 위 시는 관료주의적 자본주의 사회 시스템에서 우울함을 겪게 되는 현대인들의 전형적인 모습과 그 모습에 대한 측은지심이나 공감을 잘 표현한 시라고 할 수 있다. 또한 그러한 정서를 잘 표현한 핵심 시어로는 '사무실', '관공서', '서류뭉치', '사내', '울다' 등을 제시할 수 있다.

Fig. 2에 이렇게 선정된 주요 어휘들의 차원 감소된 단어벡터를 2D 도면에 도시하였는데, 조사와 어미를 그대로 살린 Text#1과

조사/어미를 삭제하고 원형으로 대체한 Text#2를 각각 위아래에 위치시켰다. 전처리 방식에 따른 상대적인 비교를 위해 '사무실'을 중심으로 상대적인 위치를 살펴보면, 예를 들어 '서류뭉치'의 위치가 2/4분면(Text_#1)에서 4/4분면(Text_#2)으로 이동되고 그 거리가 2배 정도 멀어졌지만, 원점을 시점으로 한 단어벡터 사이의 각도('사무실' → '서류뭉치')는 거의 120~130도 수준으로 유사함을 알 수 있다.

이는 k-최근접(Nearest Neighbor) 분류와 같은 머신러닝 기법을 적용하는 경우 텍스트 전처리 방식에 따라 핵심 어휘들의 군집도가 달라지지만, 코사인유사도와 같은 어휘들 사이의 유사도는 유지될 수 있다는 사실을 의미한다[8]. 즉, 벡터의 위치가 중요한 분석기법과 벡터의 방향이 기준인 분석기법의 적용에 있어 그 결과가 전처리 방식에 의해 영향을 받는 정도가 달라진다는 것이다.

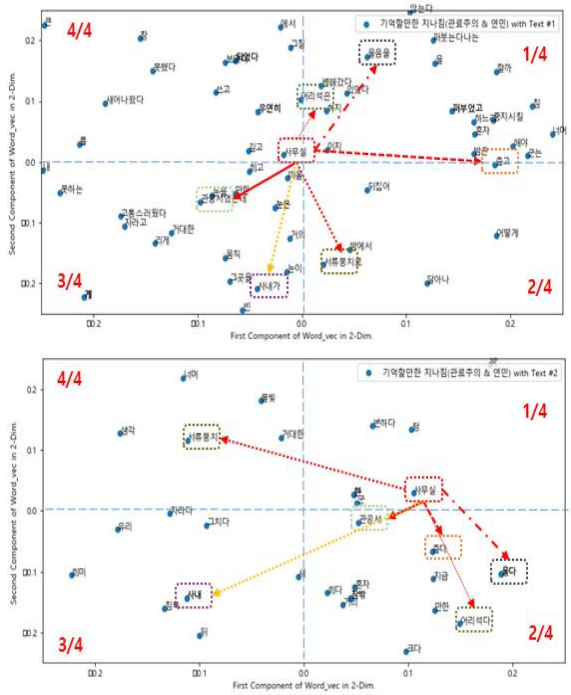

Fig. 2. Core Vocabulary Position & Direction in 2D Domain for a Poem by Hyeung-Do, Ki

Table 2. Text Preprocessing Comparison (Relative Distance & Angle from 'Office')

Property (from 'Office')		Text #1	Text #2
Distance	'Paper'	1.0	2.0
	'Man'	1.0	2.1
	'Cold'	1.0	0.2
	'Weep'	1.0	1.0
Angle	'Paper'	120deg	130deg
	'Man'	90deg	130deg
	'Cold'	170deg	40deg
	'Weep'	110deg	45deg

거리와 각도를 각 단어에 대해 정리한 것이 Table 2인데, '사무실(Office)' 단어를 기준으로 하여 다른 단어와의 거리와 벡터 사이의 각도를 정리한 것이다. Table 2에서 거리(Distance)의 경우 Text#1에 대해서는 1.0으로 간주할 때 상대적인 거리를 Text#2에 대해 정리한 것이다. 각도(Angle)는 원점에서 시작하는 '사무실'까지의 벡터와 원점에서 각 단어로 연결되는 벡터 사이의 각도를 의미한다. 전처리 방식이 Text#1에서 Text#2로 바뀌면, 상대적인 거리 요소는 '서류뭉치(Paper)' 및 '사내(Man)'의 경우 2배 정도 증가하고, '춥다(Cold)'의 경우는 1/4로 감소하며, '울다(Weep)'는 거의 비슷한 값을 가진다. 각도에 대해서는 '서류뭉치'와 '사내'는 거의 비슷하거나 약간 늘어나고, 나머지 두 단어에 대해서는 크게 줄어든다.

거리는 주로 군집/분류(clustering/classification)에 영향을 끼치고, 각도는 유사도(similarity)/유추(analogy) 연산과 밀접한 연관성이 있다는 것을 염두에 두면, 텍스트 전처리 방식에 따른 분석 과정에서 유의해야 할 사항이 파악된다고 할 수 있다. 이는 대상 문서에 포함된 문장이나 단어의 특성 벡터(feature vector)를 이용한 감성분석(sentiment analysis)이나 추천시스템(recommendation system) 구축 등과 같은 일반적인 단어/문장/문서 분석 작업에 있어 텍스트 전처리의 영향이 상이하게 작용할 수 있음을 의미한다. 즉, 단어임베딩을 통한 단어벡터의 생성 양상이 텍스트 전처리 과정에 의해 변화되고, 이에 따라 핵심 어휘별 유사도/유추/군집/분류 분석의 결과가 달라질 수 있다는 것이다. 이 같은 영향력의 변화 양상을 단어벡터 차원감소를 이용한 2D 평면 도시화를 통해 핵심 키워드 사이의 거리와 각도를 직관적으로 확인한다면 좀 더 정밀한 단어/문장/문서/특성 분석이 가능해질 것으로 판단된다.

ACKNOWLEDGEMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1F1A1064073).

REFERENCES

[1] K. Hyungsuc, Y. Janghoon, "Analyzing Semantic Relations of Word Vectors trained by The Word2vec Model", Journal of KIISE, Vol. 46, No.10, pp. 1088-1093, 2019.

[2] F. Heimerl, M. Gleicher, "Interactive Analysis of Word Vector Embeddings", Computer Graphics Forum, Vol. 37, No. 3, pp. 253-265, 2018.

[3] A. Basirat, "Real-valued Syntactic Word Vectors", Journal of Experimental & Theoretical Artificial Intelligence, Vol. 32, No.4, pp. 557-579, 2020.

[4] K. Kwangho, et al., "Input Dimension Reduction based on Continuous Word Vector for Deep Neural Network

Language Model," Phonetics and Speech Sciences, Vol. 7, No. 4, pp. 3-8, 2015.

[5] K. Sinjae, "Learning Tagging Ontology from Large Tagging Data," Journal of Korean Institute of Intelligent Systems, Vol. 18, No. 2, pp. 157-162, 2008.

[6] K. Kwangho, "Deep Learning Application for Core Image Analysis of the Poems by Ki Hyung-Do," The journal of the convergence on culture technology, Vol. 7, No. 3, pp. 591-598, 2021.

[7] B. Cowan, "Walter Benjamin's Theory of Allegory." New German Critique, No. 22, pp. 109-22, 1981.

[8] X. Shen, et al., "Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking," 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3013-3020, 2012.