The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 5, pp.489-495, September 30, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.5.489

JCCT 2022-9-60

딥러닝을 위한 텍스트 전처리에 따른 단어벡터 분석의 차이 연구

Study on Difference of Wordvectors Analysis Induced by Text Preprocessing for Deep Learning

고광호*

Kwang-Ho Ko*

요 약 언어모델(Language Model)을 구축하기 위한 딥러닝 기법인 LSTM의 경우 학습에 사용되는 말뭉치의 전처리 방식에 따라 그 결과가 달라진다. 본 연구에서는 유명한 문학작품(기형도의 시집)을 말뭉치로 사용하여 LSTM 모델을 학습시켰다. 원문을 그대로 사용하는 경우와 조사/어미 등을 삭제한 경우에 따라 상이한 단어벡터 세트를 각각 얻을 수 있다. 이러한 전처리 방식에 따른 유사도/유추 연산 결과, 단어벡터의 평면상의 위치 및 언어모델의 텍스트생성 결과를 비교분석했다. 문학작품을 말뭉치로 사용하는 경우, 전처리 방식에 따라 연산된 단어는 달라지지만, 단어들의 유사도가 높고 유추관계의 상관도가 높다는 것을 알 수 있었다. 평면상의 단어 위치 역시 달라지지만 원래의 맥락과 어긋나지 않았고, 생성된 텍스트는 원래의 분위기와 비슷하면서도 이색적인 작품으로 감상할 수 있었다. 이러한 분석을 통해 문학작품을 객관적이고 다채롭게 향유할 수 있는 수단으로 딥러닝 기법의 언어모델을 활용할 수 있다고 판단된다.

주요어 : 딥러닝, 단어벡터, 기형도, 텍스트 전처리, 텍스트 생성, 유사도

Abstract It makes difference to LSTM D/L(Deep Learning) results for language model construction as the corpus preprocess changes. An LSTM model was trained with a famouse literaure poems(Ki Hyung-do's work) for training corpus in the study. You get the two wordvector sets for two corpus sets of the original text and eraised word ending text each once D/L training completed. It's been inspected of the similarity/analogy operation results, the positions of the wordvectors in 2D plane and the generated texts by the language models for the two different corpus sets. The suggested words by the silmilarity/analogy operations are changed for the corpus sets but they are related well considering the corpus characteristics as a literature work. The positions of the wordvectors are different for each corpus sets but the words sustained the basic meanings and the generated texts are different for each corpus sets also but they have the taste of the original style. It's supposed that the D/L language model can be a useful tool to enjoy the literature in object and in diverse with the analysis results shown in the study.

Key words : Deep Learning, Wordvector, Ki Hyung-do, Text Preprocessing, Text Generation, Similarity

*정회원, 평택대 스마트자동차학과 부교수 (제1저자)
접수일: 2022년 7월 6일, 수정완료일: 2022년 8월 10일
게재확정일: 2022년 8월 31일

Received: July 6, 2022 / Revised: August 10, 2022
Accepted: August 31, 2022
*Corresponding Author: kwangho@ptu.ac.kr
Dept. of Smart Automobile, Pyeongtaek Univ, Korea

Ⅰ. 서 론

딥러닝을 이용하여 언어모델을 구축하는 대표적인 기법에는 LSTM(Long Short Term Memory) 이 있다. LSTM 기법은 단어들의 나열 순서(sequence)를 학습하는데 이 과정에서 학습용 말뭉치(corpus)를 구성하는 어휘(vocabulary)에 대한 단어벡터(word vector)를 생성한다. 이 단어벡터를 이용하면 단어별로 유사도가 높은 단어나 유사한 관계를 형성하는 단어를 찾을 수 있다. 이러한 단어벡터의 유사도(similarity) 및 유추(analogy) 연산을 통해 다양한 텍스트 분석이 가능하다[1].

특히, 이러한 방식으로 문학작품을 분석한 선행 연구도 있다[2]. 이 연구에서는 기형도 작가의 시(詩)를 분석하였는데, 기형도 작품의 핵심 주제를 표현하는 키워드를 선정하고 이 키워드와 유사도가 높은 단어를 유사도 연산으로 구해 그 타당성을 분석하였다. 연구 결과 유사도가 높은 시어(詩語)를 다양하게 얻을 수 있었으며, 다소 거리가 멀어 보이는 시어라 하더라도 해당 시어가 사용된 특정 맥락에서는 그 의미가 서로 통하는 것을 알 수 있었다. 이러한 분석을 통해 놓치기 쉬운 시어들을 풍부하게 향유할 수 있는 방법을 얻을 수 있었다. 따라서 시와 같은 문학작품의 분석에 딥러닝 기법의 단어벡터를 활용하는 것은 유효하다고 판단했다.

딥러닝을 위해서는 학습용 텍스트의 전처리가 필요하다. 구두점이나 줄바꿈 등의 기호를 공백으로 처리하고, 교착어의 특성인 조사와 어미 등을 따로 처리하기 위한 과정이 필요하다. 원문의 느낌을 보존하기 위해서는 조사와 어미 등을 유지하는 것이 좋을 수 있으나 이 경우 어휘의 수가 크게 늘어나 연산 속도와 처리 메모리 사용량 등에서 불리하다. 반면에 조사와 어미 등을 삭제하는 경우에는 어휘 수가 줄어들어 처리 속도 면에서는 유리하나 원문의 느낌과 의미를 일부 놓치게 될 수 있다[3].

본 연구에서는 기형도의 시(『입 속의 검은 잎』, 문학과지성사, 1989년)를 중심으로 텍스트 전처리 방식에 따라 달라지는 단어벡터를 이용한 작품 분석의 변화 양상을 살펴보았다. 기형도 시집의 경우 약 100페이지 정도의 분량으로, 텍스트 스캔, 전처리, 딥러닝 학습에 소요되는 시간이 짧고 각 시에 해당하는 텍스트별로 비교하기가 용이하다.

동일한 문학작품의 텍스트에 대해 상이한 전처리를 통해 얻은 단어벡터를 이용하여 유사도 및 유추 연산을 수행하였다. 이렇게 얻은 단어들이 작품의 이미지를 표현하는 수준을 분석하였고, 특정 작품을 구성하는 단어들의 벡터 성분을 2차원 평면에 표시하여 그 양상을 비교하였다. 또한 학습과정에서 얻을 수 있는 언어모델(LM : Language Model)을 이용하여 생성된 텍스트를 비교하기도 하였다. 이러한 비교 분석을 통해, 문학작품의 미묘한 뉘앙스와 분위기를 객관적으로 수치화할 수 있는 방안을 제시하였다.

Ⅱ. 본 론

1. 적용 모델 및 전처리

자연어처리에 사용되는 대표적인 딥러닝 기법인 LSTM에서는 말뭉치를 구성하는 단어들의 순서를 학습한다. 본 연구에 사용된 LSTM 구성은 그림 1과 같다. 단어벡터의 차원수가 150, 한번에 처리하는 단어의 개수(window size)는 10 이다. 최적화해야 하는 매개변수(parameter)를 줄이기 위해 임베딩 레이어와 어파인 레이어에 동일한 파라미터를 사용하였고(weight tying), 과대적합을 방지하기 위해 드롭아웃(비율 = 0.5) 레이어를 3군데 적용하였다.

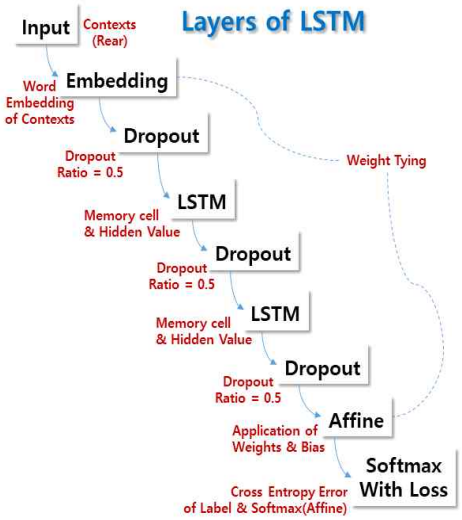

그림 1. LSTM 모델 계층
Figure 1. Layers of LSTM Model

학습용 말뭉치에 대해서는 원문을 그대로 유지한 경우(Text #1 - Original), 조사나 어미는 생략하고 형용사/

부사/동사 등은 원형('~이다')으로 대체한 경우(Text #2 - Abbreviated)로 나누어 사용하였다. 각 말뭉치의 어휘수(vocabulary), 말뭉치크기(corpus) 및 그 비율(corpus/vocabulary) 등을 표 1에 정리하였다. 원형으로 대체한 Text #2에서 그 비율이 증가하기 때문에 학습에 다소 유리하다는 것을 알 수 있다. 이것을 레고 블럭으로 비유하자면, 사용되는 블럭의 종류(어휘)가 적고 블록을 쌓아서 만든 구조물(말뭉치)이 클수록 그 블록의 적용 순서를 학습하기 쉬운 것으로 설명할 수 있다.

이렇게 전처리한 텍스트의 양상을 살피기 위해 「안개」라는 시에 대한 전처리 결과를 표 2에 정리하였다. 조사와 어미 등을 생략하고 원형으로 처리한 경우(Text #2), 원래 시어의 느낌과 뉘앙스가 거의 사라졌다. 일반적인 자연어 처리를 위한 조사/어미 생략과 원형 대체의 전처리 방식이 문학작품 분석에서는 다소 불리할 수 있음을 알 수 있다.

표 1. 전처리된 텍스트의 어휘수와 말뭉치 크기
Table 1. The number of vocabulary and corpus of the preprocessed texts

Size	Text #1 (Original)	Text #2 (Abbreviated)
Vocabulary	4,447	2,688
Corpus	11,346	10,059
Corpus / Vocabulary	2.551	3.742

표 2. 전처리된 텍스트의 비교(「안개」)
Table 2. Comparison of the preprocessed texts

Text #1 (Original)	Text #2 (Abbreviated)
아침 저녁으로 샛강에 자욱이 안개가 낀다	아침 저녁 샛강 자욱이 안개 끼다
이 읍에 처음 와본 사람은 누구나 거대한 안개의 강을 거쳐야 한다	이 읍 처음 오다 사람 누구 거대한 안개 강 거치다 하다
앞서간 일행들이 천천히 지워질 때까지	앞서다 일행들이 천천히 집다 때
(...)	(...)
몇 가지 사소한 사건도 있었다	몇 가지 사소한 사건 있다
한밤중에 여직공 하나가 겁탈당했다	한밤중 여직공 하나 겁탈
기숙사와 가까운 곳이었으나 그녀의 입이 막히자 그것으로 끝이었다	기숙사 가깝다 곳 그녀 입 막히다 그것 끝
(...)	(...)
안개는 그 읍의 명물이다	안개 그 읍 명물
누구나 조금씩은 안개의 주식을 갖고 있다	누구 조금씩 안개 주식 갖다 있다

2. 유사도/유추/텍스트생성 결과 비교

LSTM 기법을 적용하여 상이하게 전처리한 말뭉치 텍스트 두 종류를 학습시켜 언어모델과 단어벡터를 얻을 수 있었다. 학습 과정을 통해 임베딩 레이어에서 150차원의 단어벡터를 각 어휘별로 구했다. 말뭉치를 구성하는 단어의 순서를 학습시킨 언어모델을 통해 기형도 시와 유사한 텍스트를 생성시킬 수도 있다.

기형도 시를 분석한 선행 연구의 결과에 의하면, '공장화/도시화의 황량함', '관료적 인간의 우울과 공감', '유년에 대한 향수' 등이 그 핵심 이미지이다[2]. 각 이미지를 잘 표현한 시어(詩語)로써, '도시/공장', '관공서/서류', '엄마/누이' 등을 선정하여 이 키워드와 코사인 유사도(cosine similarity)가 가장 높은 단어를 연산하였다. 코사인 유사도는 두 벡터의 방향이 비슷할수록 그 연산 결과값이 증가한다[4]. 따라서 코사인 유사도가 높은 단어끼리는 그 성격과 느낌 및 분위기 등이 잘 어울린다고 할 것이다. 전처리 방식에 따라 핵심 이미지에 해당하는 키워드와의 유사도가 어떻게 달라졌는지 표 3에 정리하였다. 예를 들어 '도시'라는 키워드와 유사도가 높은 단어로, 원문을 그대로 유지한 경우(Text #1) '세우다', '내리다', '지치다' 등이 연산되었고, 조사/어미를 생략하고 원형으로 대체한 경우(Text #2) '경고', '단풍', '걷히다' 등의 단어가 연산되었다. 도시화가 진행되는 주변부의 시골 동네가 황폐해지는 이미지의 키워드인 '공장'의 성격을 음미해보면, 각 전처리 방식에 따라 유사도 연산의 결과로 얻은 단어가 상이하더라도 그 분위기가 일맥상통하는 것을 알 수 있다.

전처리 방식에 따라 유사도 연산의 결과는 달라지지만 여전히 그 분위기와 뉘앙스 등이 통하는 단어를 구할 수 있음을 알 수 있다. Text #2의 '단풍'은 다소 어색하여 '공장'이라는 단어와 어울리지 않으나 기형도 시집에 유일하게 '단풍'이라는 단어가 등장하는 「병」이라는 작품을 분석하면 유사도가 높은 단어임을 알 수 있다. 「병」에서 단풍은 '잔인하게 죽어간 붉은 세월이 접힌 몸통 위에, 아 사람아 단풍든다'에 등장하는 단어이다. 따라서 '단풍'은 늙은 나무가 죽어가는 모습을 시각적으로 묘사한 것으로, 그 연상되는 이미지는 '공장'의 효과와 비슷하다고 볼 수 있다. 이처럼 유사도 연산으로 획득한 단어들을 통해 작품에 대한 새롭고 정밀한 감상이 가능해진다고 할 수 있다.

나머지 단어들도 그 느낌이 비슷하다. 따라서 유사도 연산을 통해 분위기/스타일/뉘앙스가 비슷한 단어를

얻을 수 있으며, 이를 통해 기형도 시의 의미를 새롭게 음미할 수 있지만, 그 전처리 방식에 따라 서로 다른 단어들이 연산됨을 알 수 있다. 이는 전처리 방식에 따라 어휘의 종류와 개수가 달라지고, 이 어휘들의 중복을 허용하는 순열(sequence)인 말뭉치의 구조가 변하기 때문인데, 말뭉치의 순서를 기억하는 LSTM 학습에 의해 결정되는 단어벡터의 방향(특성)도 크게 달라진다는 것을 알 수 있다[5].

표 3. LSTM 단어벡터로 구한 핵심 이미지의 키워드와 유사도가 높은 단어
Table 3. Similar words to the keywords for core image recommended by word vectors of LSTM

Image	Keyword	Text #1 (Original)	Text #2 (Abbreviated)
공장화 및 도시화	도시	세우다 내리다 지치다	경고 단풍 걷히다
	공장	수은주 실패 물	괭이 그을음 투명하다
관료적 인간의 우울과 공감	관공서	자격 예배당 삼촌	짓궂다 노여움 추리닝
	서류	사각 석탄 보행	열무 단호하다 홍당무
유년에 대한 향수	엄마	가늠하다 김장 틈	침대 코밑 덩어리
	누이	햇살 망토 휴식	험악하다 자르다 폐수

다음으로 단어벡터의 유추(analogy) 연산을 수행하여 정리한 것이 표 4이다. 각 핵심 이미지에 해당하는 키워드 사이의 관계와 유사한 제시어의 짝(pair)을 찾는 것이 유추 연산이다[6]. 예를 들어 '공장화/도시화'의 이미지에 해당하는 두 키워드인 '공장-여직공'의 관계와 유사한 관계를 형성하는 단어의 짝을 찾는 것이 유추 연산이다. 표 4에서 '공장-여직공'에 해당하는 유추 연산을 '밭' 이라는 제시어에 대해 수행했는데, 각 전처리 방식에 따라, '성냥불/독서/기도'(Text #1)와 '바라보다/달아나다/모르다'(Text #2)의 단어들을 얻을 수 있었다.

'공장'에서 고생하거나 학대당하기도 하면서 자식을 키우는 일상을 살아가는 주체인 '여직공'의 관계를 고려하면서 그 유추 연산의 결과를 살펴보면 기형도 시를 다양한 관점에서 이해할 수 있게 된다. 예를 들어 Text #1에서는 '밭 - 성냥불/독서/기도' 등이 유추 연산되었다. '밭'이 등장하는 「위험한 가계 육십구년」을 살펴보면, '밭'은 아픈 아버지가 썩은 감자를 캐는 장소로 표현된다. 이 시에서 아픈 아버지의 약값을 두고 공장에 다니는 누이와 엄마가 다투고, 이웃의 닭에게 모이를 주는 아버지와 화자가 어긋나는 대화를 나누는 절망적인 어린 시절을 묘사한다. 따라서 '밭'은 썩은 감자를 캐는 희망이 없는 장소이고, 이러한 장소의 주체는 아픈 아버지라고 볼 수 있다.

표 4. 대표적인 시어의 관계에 해당하는 LSTM 기준 유추 단어
Table 4. Corresponding analogue words to the relations of the representative poetry words

Image / Keyword	Text #1 (Original)	Text #2 (Abbreviated)
(공장 및 도시화) 공장 - 여직공 ‖ 밭 - ?	성냥불 독서 기도 상점 망토 골목	바라보다 달아나다 모르다 내리다 빠르다 터지다
(관료적 인간의 우울) 사무실 - 서류 ‖ 집 - ?	소리 여자 아이들 우리 기억 풀잎	떠나다 주인 여전하다 땅속 손 입
(유년에 대한 향수) 누이 - 꽃 ‖ 행인들 - ?	더럽다 빵 폐허 죽다 장마통 무수하다	소리 습관 크다 우리 사람 자정
(남녀 주체의 차이) 사내 - 눈물 ‖ 여자 - ?	발자국 유리 공기 빛깔 어둠 육체	오다 보다 빠지다 접히다 건물 뼈

'공장-여직공'의 관계와 유사하게 아픈 아버지가 밭에서 하는 행동을 연상시키는 '성냥불/독서/기도' 등의 단어가 연산되었음을 알 수 있다. 이를 자세히 검토하면, '성냥불'이 등장하는 「성탄목 겨울 판화3」에서 얼음으로 꽉 찬 방 안의 성탄목을 보며 '성냥불'을 켜보지만 어떻게 견딜 수 있을까 반문하고 있다. 즉, '성냥불'은 빈한한 주체의 절망적인 행동의 예시로 볼 수 있어 그 분위기가 통하는 것을 알 수 있다. 따라서 유추 연산에서도 전처리 방식에 따라 서로 다른 단어를 얻었지만

원래의 관계를 함축하는 시어들의 짝을 적절히 얻을 수 있음을 알 수 있다. 또한 이는 시를 음미하는 새로운 관점을 구할 수 있는 하나의 방식으로 볼 수 있다.

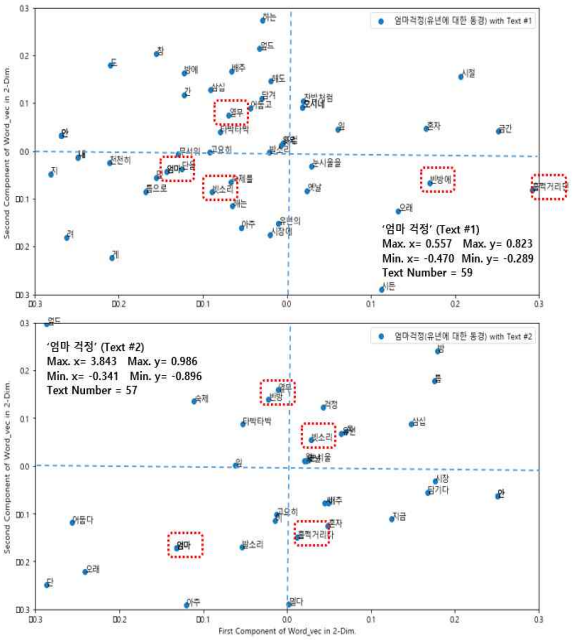

그림 2. 차원감소된 단어벡터 도시(「엄마걱정」)
Figure 2. Display of the wordvectors in 2-dimension

다음으로 150차원의 단어벡터를 2차원으로 감소시킨 후 각 단어들의 위치를 비교해보았다. 벡터의 차원감소는 주성분분석 기법으로 수행하였다. 주성분분석에서는 전체 어휘의 단어벡터의 공분산행렬를 구한 뒤 그 행렬의 특성치(eigenvalue)와 특성벡터(eigenvector)에 해당하는 단어벡터 성분(벡터 내적)을 구하여 차원을 감소시킨다[7]. 특성치가 클수록 특성의 분산이 크기 때문에 가장 큰 분산의 두 방향으로 단어벡터 내적을 취하면 원래 벡터의 성격을 유지하면서도 2차원으로 그 차원을 줄일 수 있다[8]. 이렇게 감소된 2차원 단어벡터의 경우 평면상에 그 위치를 도시할 수 있어 직관적으로 단어 사이의 거리와 방향 등을 파악할 수 있다. 따라서 주요 시어들 사이의 관계를 새롭게 인식할 수 있다. 그림 2에 「엄마걱정」이라는 시의 단어들을 이런 방식으로 도시하였다. 이 시에는 열무 삼십 단을 이고 시장에 간 엄마를 기다리는 어린 시절의 화자가 등장한다. 빗소리 들리는 빈방에 엎드려 훌쩍이면서 시든 배추 잎같은 엄마의 발소리를 기대하는 아이의 가련한 모습이 절묘하게 표현된 작품이다.

'엄마', '빈방', '열무', '빗소리', '훌쩍이다' 등의 단어를 각 전처리 방식에 따라 그 위치와 방향을 살펴보면, Text #1(上)에서는 '엄마-빗소리-열무'의 단어와 '빈방-훌쩍이다'의 단어들이 근접하고, Text #2(下)에서는 '빈방-열무-빗소리' 단어들이 근접하고, '엄마-훌쩍이다'가 서로 가깝게 배치됨을 알 수 있다. 두 경우 모두 '빗소리-열무'가 근접 배치되어 있다. 이에 반해 '훌쩍이다'는 '빈방'(Text #1) 혹은 '엄마'(Text #2)와 근접하여 그 성격이 조금 달라지는 것을 알 수 있다. 빈방에서 훌쩍이는 화자가 엄마를 기다리는 시적 상황을 고려하면 두 경우 모두 의미가 통하면서도 미묘하게 다른 감상의 관점을 제공하는 것으로 해석할 수 있다. 따라서 유사도 및 유추 연산과 마찬가지로, 2차원 단어벡터의 도시를 통해 단어 사이의 관계를 직관적으로 파악하면 그 감상법이 풍부해짐을 알 수 있다. 다만 전처리 방식에 따라 단어 사이의 거리와 방향이 달라지므로 방향을 기준으로 하는 유사도 연산과 거리를 기준으로 하는 군집(clustering) 연산의 결과가 다르게 영향을 받을 것으로 예상된다[9]. 다른 분야의 연구에 있어 이를 고려하여 분석한다면 도움이 될 것으로 판단된다.

표 5. LSTM 언어모델을 이용한 텍스트 생성 결과 비교 #1
Table 5. Comparison of the text generation with LSTM language model #1

Text #1 (Original)	Text #2 (Abbreviated)
엄마 - 사랑 창문 이후 가게에 적신다 다보면 칼라가 쌓으면 꿈들 감기고 축축한 황혼 눕혀두고 물 무서워 경악할 왼손 가서 떠다니	**엄마** - 램프 얼음 유배지 전율 지우다 실수 은백양 얼핏 약간 이마 어김없이 돋우다 가슴 얕다 송판들이 늦다 미래 얼룩이
↓	↓
엄마가 사랑하는 창문에서 이후부터 가게를 적신다 / 내다보면 칼라가 쌓이며 꿈들이 감기고 / 축축한 황혼을 눕혀둔 물이 무서워 / 경악하는 왼손이 가다 떠다닌다	엄마가 램프의 얼음을 유배지처럼 전율하며 지운다 / 실수처럼 은백양이 얼핏 약간은 이마를 어김없이 돋운다 / 가슴이 얇은 송판들이 늦은 미래에 얼룩진다

마지막으로 LSTM 학습을 통해 얻을 수 있는 언어모델을 이용하여 텍스트를 생성시켜 보았다. '엄마'라는 키워드를 주고 언어모델이 생성시키는 단어들을 순차적으로 연결시켜 텍스트를 구성한 것이 표 5이다. 조사와 같은 어미 변화에 따라 의미와 분위기가 달라지는 교착어인 한글에서 LSTM 방식으로 학습된 언어모델로 텍스트를 생성시키면 당연히 조사의 변화가 적절치

못해 문장을 파악하기 어렵다[10]. 특히 조사/어미를 삭제하고 원형으로 대체한 Text #2의 경우 생성된 텍스트를 알아보기는 더욱 어렵다. 따라서 생성된 단어들의 순서를 유지한 채 조사를 수동으로 수정하여 자연스러운 문장으로 바꾸어 아래 박스에 표시하였다. 다소 어색한 문장들이나, 시적 상상력을 발휘하여 생성된 텍스트를 독해하면 이색적인 의미와 분위기를 감상할 수 있다. 특히 모든 시어들이 기형도의 것이기에 기형도의 시에 익숙한 감상자는 생성된 텍스트의 분위기와 느낌을 용이하게 음미할 수 있을 것이다. 기형도의 작품과 유사한 분위기의 시를 간편하게 생성시키고 그 의미가 통하도록 약간의 수정작업(조사 변화)을 추가해서 새로운 작품을 얻을 수 있는 것이다.

표 6. LSTM 언어모델을 이용한 텍스트 생성 결과 비교 #2
Table 6. Comparison of the text generation with LSTM language model #2

Text #1 (Original)	Text #2 (Abbreviated)
여직공 안개 죽음 눈 - 플랫폼 위하여 화차 차마 잡이여요 힘없 점검중이지 그런 죠 날랐다 땅 풀어 썩은 감는다 반짝이 정거장에 풀면 아름다운 엔 거쳐야 사시나무 미리 올봄엔 또한 뛰어	**여직공 안개 죽음 눈** - 절망 뜻밖 문지방 갇히다 수천 꿈만한 맡 얇다 튼튼한 대장장이 옆 여러 채 평안한 엎질러진 기다리 쏟다 코 않다 가늠여 실수 두서너 편
↓	↓
플랫폼을 위하여 화차는 차마 잡지 못하고 / 힘없이 점검중이다 / 그렇죠 날라도 땅이 풀려 썩은 채로 감기겠죠 / 반짝이는 정거장에 풀리는 아름답고 거친 사시나무 / 미리 올봄에 또 뛰어든다	절망스럽지만 뜻밖에도 문지방에 갇혔다 / 수천 개의 꿈만큼 얇고 튼튼한 대장장이 옆 / 여러 채가 평안하게 엎질러져 있다 / 기다리다 쏟아지는 코들은 아니라고 가늠해도 실수는 두서너 편이다

표 6은 '여직공-안개-죽음-눈'의 단어를 순차적으로 입력했을 때 LSTM 언어모델이 생성시킨 텍스트를 예시한 것이다. LSTM 모델의 경우 제시어의 순서에 따라 매개변수 연산을 통해 다음에 연결될 단어를 확률적으로 선정하므로 하나의 단어만 도입했을 때와 그 텍스트 생성 양상이 달라진다[11]. 역시 조사 연결을 수동으로 작업하면 기형도의 기존 작품과 분위기가 유사하면서 이질적인 작품을 얻을 수 있다. 전처리 방식에 따라 생성된 텍스트가 상이하지만 시적 상상력이 허용하는 범위 내에서 음미할 가치가 있는 텍스트들이라고 판단된다. 두 전처리 방식 중에 어느 편이 더 우수한 결과를 얻었다고 말하기는 어려워도 서로 다른 기표 연쇄를 통해 익숙하면서도 미묘하게 다른 감상 지점를 얻을 수 있다. 문학작품을 향유하는 하나의 방법으로써 딥러닝 기법을 적용하는 것은 합리적인 시도라고 판단된다.

III. 결 론

언어모델을 구축하기 위한 대표적인 딥러닝 기법인 LSTM의 경우 어휘의 중복 허용 순열로 구성되는 말뭉치를 학습한다. LSTM 기법이 적용된 딥러닝 모델에서 말뭉치를 구성하는 어휘의 배치 순서를 기억하기 위해서는 단어 임베딩 과정이 필요하고, 학습이 종료되면 임베딩 과정에서 결정된 단어벡터를 얻을 수 있다.

본 연구에서는 기형도의 유작 시집인 『입 속의 검은 잎』을 말뭉치로 사용하였다. 이 말뭉치의 텍스트 전처리 방식에 따라 변화하는 연관 단어의 유사도 및 유추 연산 결과를 비교 분석하였다. 또한 단어벡터에 대해 주성분분석을 적용하여 차원 감소시킨 2차원 단어벡터를 평면에 도시하여 전처리 방식에 따른 주요 단어들의 배치 변화를 분석하였다.

말뭉치를 구성하는 텍스트를 그대로 살린 경우(Terxt #1)와 조사/어미를 삭제하고 동사/형용사/부사 등을 원형으로 대체한 텍스트(Text #2)의 두 가지 전처리 방식의 결과를 비교 분석하였다. 선행 연구에서 제시되었던 기형도 시의 핵심 이미지를 잘 표현한 시어들을 선정해 유사도/유추 연산으로 얻을 수 있는 단어들을 검토했다. 전처리 방식에 따라 서로 다른 단어들을 연산 결과로 얻었지만 그 분위기/느낌/뉘앙스 등은 일맥상통하여 적절한 유사도 및 유추 연산이라고 판단된다. 또한 다소 이질적인 단어들이 연산되기도 했지만 해당 단어가 등장하는 시를 살펴보면 제시어(query)와 유사도가 높고, 유추관계가 형성됨을 알 수 있었다. 이러한 분석과정을 통해 문학작품을 좀더 다양하고 심도깊게 감상할 수 있다고 판단된다.

차원감소시킨 단어벡터들을 평면에 도시한 결과 전처리 방식에 따라 주요 시어들 사이의 거리와 방향이 일부 변화하는 것을 확인할 수 있었다. 이는 단어 사이의 거리가 중요한 군집 연산과 방향이 중요한 유사도/유추 연산의 결과가 전처리 방식에 따라 상이하게 영향을 받을 수 있음을 의미한다. 따라서 텍스트 분석에서 주로 사용하는 연산의 종류에 따라 전처리 방식의 효과를 구분해서 접근할 필요가 있다. 또한 이러한 접근법은

단어벡터의 위치를 평면 상에서 직관적으로 인식할 수 있기 때문에 근접한 단어들의 배치 양상을 통해 문학작품을 구성하는 주요 시어들의 성격을 객관적으로 확인할 수 있는 수단도 얻을 수 있다.

LSTM 기법으로 학습시킨 언어모델을 사용하여 텍스트를 생성시켜 비교해 보았다. 조사와 어미의 변화에 따라 그 의미가 크게 변하는 교착어인 한글의 특성에 의해 생성된 텍스트를 그대로 인식/감상하기는 어려우나, 이를 수동으로 적절한 조사/어미로 수정하면 생성된 텍스트를 충분히 이해할 수 있었다. 생성된 텍스트는 모두 학습용 말뭉치인 기형도의 기표로 연쇄된, 원문과 유사하면서도 이질적인 작품으로 감상할 수 있었다. 원래의 문학작품을 잘 이해하고 있는 독자라면 이렇게 생성된 텍스트를 통해 창작의 즐거움을 이색적으로 향유할 수 있을 것으로 판단된다.

이상의 분석 결과를 통해 문학작품을 객관적으로 이해하고 감상할 수 있는 방식으로써 딥러닝 기법을 활용할 수 있음을 알 수 있었다. 특히 유사도 및 유추 연산의 결과로 문학작품을 다채롭게 이해할 수 있고, 2차원 평면에 표시된 단어들의 위치를 통해 직관적인 시어의 성격 이해 및 텍스트 생성 기능을 이용한 이색적인 창작욕의 향유도 얻을 수 있었다. 이러한 결과물은 말뭉치 텍스트의 전처리 방식에 따라 달라지지만 그 효과는 거의 동일하다. 이상의 검토를 통해 전혀 다른 두 분야의 융복합적 접목을 통해 문학의 객관적 이해를 도모하고 이색적인 미적 가치를 향유할 수 있는 유효한 방법론을 본 연구에서 제시한 것으로 판단된다.

References

[1] K. Hyungsuc, Y. Janghoon, “Analyzing Semantic Relations of Word Vectors trained by The Word2vec Model”, Journal of KIISE, 46(10), pp. 1088-1093, 2019
[2] K. Kwangho, “Deep Learning Application for Core Image Analysis of the Poems by Ki Hyung-Do,” Journal of the Convergence on Culture Technology, 7(3), pp. 591 - 598, 2021.
[3] F. Heimerl, M. Gleicher, “Interactive Analysis of Word Vector Embeddings”, Computer Graphics Forum, 37(3), pp. 253-265, 2018
[4] A. Basirat, “Real-valued Syntactic Word Vectors”, Journal of Experimental & Theoretical Artificial Intelligence, 32(4), pp. 557-579, 2020
[5] Y. Chang, et al., “Using Word Semantic Concepts for Plagiarism Detection in Text Documents”, Information Retrieval Journal, 24(4-5), pp.298-321. 2021
[6] K. Sinjae, “Learning Tagging Ontology from Large Tagging Data,” Journal of Korean Institute of Intelligent Systems, 18(2), pp. 157 - 162, 2008.
[7] K. Kwangho, et al., “Input Dimension Reduction based on Continuous Word Vector for Deep Neural Network Language Model,” Phonetics and Speech Sciences, 7(4), pp. 3 - 8, 2015.
[8] A. Gavric, et al., “Real-Time Data Processing Techniques for a Scalable Spatial and Temporal Dimension Reduction”, 21st International Symposium(INFOTEH), pp. 1 - 6, 2022
[9] Y. Lee, et al., “Applying Convolution Filter to Matrix of Word-clustering Based Document Representation”, Neurocomputing, 315, pp.210-220, 2018, doi:10.1016/j.neucom.2018.07.018
[10]L. Hickman, et al., “Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations”, Organizational Research Methods,25(1), pp.114-146, 2022
[11]N. Fatima, et al., “A Systematic Literature Review on Text Generation Using Deep Neural Network Models”, IEEE Access, 10, 53490 - 53503. 2022