The Korean Journal of Applied Statistics
2021, Vol. 34, No. 3, 295–308

DOI: https://doi.org/10.5351/KJAS.2021.34.3.295
Print ISSN1225-066X / Online ISSN 2383-5818

Understanding the semantic change of Hangeul using word embedding

Hyunseok Sun[a], Yung-Seop Lee[b], Changwon Lim[1,a]

[a]Department of Applied Statistics, Chung-Ang University; [b]Department of Statistics, Dongguk University

Abstract

In recent years, as many people post their interests on social media or store documents in digital form due to the development of the internet and computer technologies, the amount of text data generated has exploded. Accordingly, the demand for technology to create valuable information from numerous document data is also increasing. In this study, through statistical techniques, we investigate how the meanings of Korean words change over time by using the presidential speech records and newspaper articles public data. Using this, we present a strategy that can be utilized in the study of the synchronic change of Hangeul. The purpose of this study is to deviate from the study of the theoretical language phenomenon of Hangeul, which was studied by the intuition of existing linguists or native speakers, to derive numerical values through public documents that can be used by anyone, and to explain the phenomenon of changes in the meaning of words.

Keywords: semantic change, word2vec, procrustes align, corpus linguistics

1. 서론

최근 들어 자연어 처리 분야에서 엄청난 각광을 받고 있는 분야가 단어 임베딩에 관한 연구이다. 기존의 텍스트를 표현하는 방법은 bag-of-words 기반의 방법들이 사용되었다. 단어 임베딩이 나오기 전까지 단어를 표현함에 있어 원-핫 벡터가 사용되었고 문맥이나 문서를 표현함에 있어 나타난 단어의 출현 빈도를 이용하였다. 이 방법을 기반으로한 지속적인 연구를 통해 문서 분류에 있어 문서들의 특징을 구분하기 위한 방법으로 TF-IDF 벡터 표현이 등장하였고, 이 방법은 정보 검색(information retrieval)이나 자연어 처리(natural language process)에서 큰 성과를 거두었다. 하지만 이는 bag-of-words 표현의 연장선이었고 bag-of-words 방법의 한계점인 희소성(sparsity)과 고차원 문제를 극복할 수 없었다.

원-핫 벡터는 전체 사전 크기의 벡터로, 해당 단어의 인덱스만 1의 값을 가지고 나머지 인덱스에서는 0을 갖게 된다. 이를 통해 단어의 구분은 가능하지만, 해당 단어 벡터에 대한 의미 유추는 불가능하다. 또한 실제로 사람들이 사용하는 단어의 종류는 기하급수적으로 커지게 되고 사전의 크기가 늘어나면서 해당 단어 벡터를 표현함에 있어 비효율적인 문제를 겪는다. 마지막으로 문서를 단어의 원-핫 벡터를 기반으로 표현함에 있어, 문서에 나타난 단어를 제외하고는 대부분의 단어가 0의 값을 갖게되는 희소성 문제가 발생하여 이를 기반으로 모형을 만들기에 적합하지 않은 문제가 발생한다.

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7083281).

[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: clim@cau.ac.kr

Published 30 June 2021 / journal homepage: http://kjas.or.kr

위와 같은 문제를 해결하기 위해, 단어를 표현함에 있어 이산적인 값(0 또는 1)이 아닌 연속적인 값으로 단어를 표현하면서 단어 벡터에 대한 차원을 줄이고 해당 벡터가 단어의 의미를 내포할 수 있도록 표현하기 위해 나타난 방법이 단어 임베딩(word embedding)이다. Bag-of-words 기반의 표현보다 단어를 효율적으로 표현하기 때문에 최근 자연어 처리 연구에서 다양하게 응용되고 있다. 특히, 기계학습을 통한 문서 분류, 감성분석 및 기계번역 등에서 우수한 성능을 보이고 있다. Kim 등 (2014)은 convolution neural network (CNN)의 입력층에 단어의 임베딩 값을 활용하여 문장 분류에 있어 정확도를 향상시켰다. Mikolov 등 (2013)은 기계번역에서 단어의 임베딩을 활용하여 영어와 스페인어의 번역 문제에서 90%의 높은 정확도를 보였다.

본 연구에서는 단어 임베딩 기법이 단어의 의미를 내포한다는 점에서 이를 응용하여 단어의 의미변화 파악에 활용하였다. 2장에서는 한글의 의미변화를 파악하기 위한 연구 개요를 살펴보고, 3장에서는 이를 수행하기 위한 방법론을 다룬다. 4장에서는 방법론을 적용한 실제 자료에 대한 설명과 이를 통해 얻은 분석 결과를 설명하고, 5 장에서는 연구의 결론과 향후 연구 방향에 대하여 논한다.

2. 의미변화 탐지 연구 개요

2.1. 단어 임베딩

단어 임베딩은 단어의 분산 표현 기법(distributed representation)의 방법으로, 개별의 단어들을 연속적인 공간 안에서의 실수 벡터로 표현하는 방법이다. 실수 벡터 표현을 학습하기 위해 전제가 되는 가정은 비슷한 맥락에 등장하는 단어들은 유사한 의미를 지닌다는 언어학의 의미론인 분산가정(distirbutional hypothesis)에 기반한다 (Harris, 1954). 분산가정을 기반으로 한 단어의 벡터 표현은 주변 단어를 통해 얻어지고, 이는 밀집된 실수 값을 갖는 축소된 차원 크기의 벡터이다 (Sahlgren, 2008).

단어 임베딩을 위한 고전적인 방법으로 잠재의미분석(latent semantic analysis; LSA)이라는 방법이 존재한다. LSA에서는 단어의 출현 빈도를 기반으로 한 행렬 M을 구축하고 이를 특이값 분해(singular value decomposition) 통해 차원을 축소하는 과정으로 단어 임베딩을 수행한다 (Golub과 Reinsch, 1970). 이때 행렬 분해가 되는 대상 행렬 M에 대한 다양한 연구가 진행되었다. 가장 초기의 방법으로서, 단어의 출현 빈도를 표현한 행렬 M은 용어-문서 행렬로 문서와 단어에 대한 분산 표현을 얻고자 하였다 (Deerwester 등, 1990). 용어-문서 행렬의 희소성 문제를 극복하기 위해 전체 코퍼스를 이용하는 용어-문서 행렬이 아닌 윈도우 기반의 용어-문맥 공기 행렬(window based co-occurence matrix)에 대한 특이값 분해를 통해 단어에 대한 임베딩을 수행하였다 (Naptali 등, 2009). 이후의 연구를 통해, 일반적으로 positive pointwise mutual information (PPMI) 행렬에 대한 특이값 분해에서 단어에 대한 임베딩 성능이 좋은 것으로 알려져 있다 (Matveeva 등, 2007).

이와 다르게, 빈도 기반이 아닌 예측 기반의 신경망 기반 언어 모델은 주변 단어를 통해 중심 단어의 등장을 예측하는 과정을 통해 단어 임베딩을 수행하는 방법이다. 대표적인 예측 기반의 임베딩 모형인 워드투벡터(word2Vec)는 효율적인 단어 벡터 표현이 가능하고 단어 벡터의 연산을 통한 의미 추론 문제에 있어 기존 특이값 분해를 기반으로 한 LSA보다 좋은 임베딩 성능을 보이고 있다 (Mikolov 등, 2013). 학습 방식에 따라 continuous bag-of-words (CBOW)와 continuous skip-gram 방법이 있으며 두 모형의 학습 방식은 Figure 1과 같다. CBOW 모형은 주변 단어(context word)를 이용해 중심 단어(target word)를 예측하는 방법이고 skip-gram 모형은 중심 단어를 이용하여 주변 단어를 예측하는 방법이다.

CBOW 모형은 주변 단어를 기반으로 중심 단어가 나타날 로그 가능도를 최대화되도록 학습하고 이를

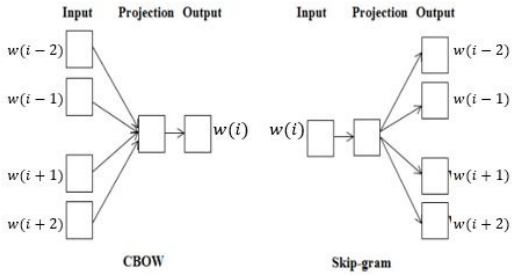

Figure 1: *Illustration of CBOW and skip-gram models.*

식으로 표현하면 식 (2.1)과 같다.

$$L = \sum_{c \in C} \log P(w_t|w_c), \quad P(w_t|w_c) = \frac{e^{v_t v_c}}{\sum_{t' \in V} e^{v'_t v_c}}. \tag{2.1}$$

워드투벡터는 분산 가정을 토대로 주변단어를 통해 각 단어에 대한 축소된 벡터 표현을 학습하고 학습된 벡터가 단어의 의미를 내포한다는 점에서, 이를 이용하여 한글의 의미변화 파악 연구에 활용하였다.

2.2. 의미변화 탐지 연구

단어의 의미변화(semantic change)는 한 언어의 기호와 관련하여 관습적으로 인정된 의미가 다른 의미로 바뀌는 것을 의미한다 (Cho, 2004). 의미변화의 원인으로는 역사적인 사건에 의해서 사건을 명명하는 단어에 의해 나타나기도 하고, 사회적인 원인으로는 사회를 구성하고 있는 계층이나 조직에 따라서 사용하는 말의 의미가 달라지기도 한다. 또한 심리적인 원인으로서 대상 자체가 가지고 있는 속성에 사람의 심리가 반영되어 은유적인 표현으로 의미변화가 일어나는 경우도 있다 (Yoon, 2013). 예를 들어, “사이다”는 원래 탄산음료의 명칭으로 생겨난 단어이지만, 사이다의 속성인 시원함이 반영되어 은유적인 표현으로 “통쾌하다”의 의미를 갖는 단어로 사용되기도 한다.

이러한 언어적 현상인 의미변화를 통계적인 기법을 이용하여 설명하려는 연구가 진행되었지만, 이는 주로 영어권의 언어 데이터를 사용하여 영단어에서의 의미변화를 설명하고자 하였다. 대표적으로 Kim 등 (2014)은 구글의 Ngram corpus 데이터 셋을 이용하여 1900년대초부터 2000년대까지 나타난 단어에서 의미변화가 일어난 단어를 발굴하였다. 발굴한 대표적인 의미변화 단어로 “cell”이 1900년대 초에는 “closet”의 의미로 사용되었지만 1990년대 이후에는 “phone”이라는 의미로 사용됨을 나타내었고 “gay”라는 단어가 “cheerful”이라는 의미에서 “bisexual”을 뜻하는 단어로 사용이 변함을 실제 자료를 통해 증명하였다. 또 다른 의미변화 탐지 연구로 Kulkarni 등 (2015)은 시점별로 워드투벡터를 통해 임베딩 행렬을 구축하고 이를 선형변환을 통해서 정렬하는 과정을 통해 의미변화를 탐지하였다. 대표적인 단어로 “plastic”은 “flexible”에서 “polymer”란 뜻으로 변화함을 발견하였다. 마지막으로 Hamilton 등 (2016)은 단어 임베딩 방법으로 PPMI, SVD, Skip-Gram을 이용하고 의미변화 평가 데이터 셋을 구축하여 세 가지 임베딩 방법에 따른 의미변화 포착 정도를 비교하였다. 또한 불어와 독어, 중국어로 구축된 오픈 소스 데이터를 활용하여 다른 언어권에서의 적용 가능성을 제시하였다.

이와 같은 선행연구를 토대로 본 연구에서는 통계적인 방법과 기계학습을 이용하여 한글에서의 의미변화 현상을 포착하고, 이를 한글의 통시적 변화 연구에 활용하는 방안을 제시하고자 한다. 이를 위한 방법론으로는 워드투벡터를 통해 시점에 따라 단어의 의미를 반영할 수 있는 벡터를 학습하고, 시점별로 프로크러스티즈

직교 정렬(procrustes orthogonal alignment)을 통해 시간에 따라 단어의 의미를 수치화한다. 수치화된 시점별 단어 벡터를 통해 단어의 의미변화 정도를 시계열 자료로 구성한다. 이를 통해 기존 언어학자나 원어민의 직관에 의해 연구되던 한글의 이론적 언어 현상 연구에서 벗어나 누구나 사용할 수 있는 공공데이터를 통해 의미변화 현상을 설명한다.

3. 의미변화 탐지를 위한 방법론

본 연구에서 단어의 의미변화를 포착하기 위한 방법은 크게 단어의 축소된 벡터 표현 학습 과정과 임베딩 행렬을 정렬하는 과정, 마지막으로 의미변화를 정량화하고 변화를 탐지하는 과정으로 구성된다. 이를 위해서 우선적으로 시간 정보가 있는 문서 집합을 시간단위에 따라 분할한다. 분할된 문서 집합 내에서 워드투벡터를 통해 임베딩된 단어 벡터를 통해 시점별 임베딩 행렬을 구축한다. 이후 각 시점별 임베딩 행렬을 단어에 따라 정렬하는 과정을 거치고 정렬된 단어의 특징 값을 통해 유사도를 측정한다.

3.1. 워드투벡터를 이용한 시점별 단어 임베딩

의미변화를 탐지하기 위해 우선 분산 가정을 토대로 어떤 시점에 나타난 단어의 의미는 해당 시점의 주변 단어에 의해 결정된다는 가정을 기반으로 시점별로 단어에 대한 임베딩을 수행한다. 시점별 단어를 임베딩하기 위해 워드투벡터의 학습방식인 Skip-Gram을 이용하였다. 이를 위해 우선 전체 코퍼스인 D를 시점 단위에 따라 전체 T개의 시점으로 분할한다. 이후 각 시점 t에서의 코퍼스 D_t에서 단어 w마다 크기 d의 실수 벡터 $u_w^{(t)}$를 워드투벡터를 통해 구한다. 최종적으로 시점별 코퍼스에서 나타난 단어들의 임베딩 값을 구하고 시점별 임베딩 행렬 $W^{(1)}, \ldots, W^{(T)}$을 구축한다. 아래의 기호 설명을 통하여 위의 과정에 대한 설명을 보충하였다.

- D=($D_1, \ldots, D_T$), D: 전체 코퍼스, D_t: t 시점의 코퍼스
- $w \in V_t$, w: 단어, V_t: t 시점에 나타난 단어 집합
- d: 임베딩 크기
- $u_w^{(t)} \in R^d$, $u_w^{(t)}$: t 시점에 나타난 단어 w의 단어 벡터
- $W^{(t)} \rightarrow R^{|V_t| \times d}$, $W^{(t)}$: t 시점의 임베딩 행렬

이때 시점을 나누어 독립적으로 학습하게 되는 경우, 두 가지 문제가 발생하게 된다. 첫번째로 랜덤한 초기치 설정으로 인하여 서로 다른 시점에서 같은 단어의 임베딩 결과가 다르게 나타나는 문제가 발생한다. Figure 2은 3개의 시점을 나누어 8개의 단어에 대해 2차원의 임베딩 벡터를 그림으로 표현한 결과이다. 시점 내에서 단어의 의미는 학습하였지만, 시점별 비교시 같은 단어라도 서로 다른 좌표 상에 임베딩된 것을 알 수 있다. 이러한 문제로 인해 의미변화를 포착하기 위해 한 단어에 대한 시점별 비교를 하는 경우, 단어의 의미는 변하지 않았지만 임베딩 결과는 서로 다르게 되어 정확한 의미변화를 포착할 수 없게 된다.

두 번째 문제는 워드투벡터 모형은 학습된 단어에 대해서만 추론이 가능하기 때문에 모든 시점에서 공통적으로 나타나는 단어만 비교할 수 있다는 문제가 발생한다. 이를 out-of-vocabulary (OOV)라하며, 시점별로 독립적으로 학습하여 얻은 단어 벡터를 비교할 경우 어떤 단어가 특정 시점에 나타나지 않으면 해당 단어는 시점별 비교가 불가능하게 된다. Figure 3는 서로 다른 시점에서 서로 다른 단어들이 임베딩된 결과의 예시로, 3개의 시점에서 공통적으로 나타난 “사과”의 경우만 시점별 비교가 가능하고 나머지 단어들은 OOV 문제로 인해 비교가 불가능하게 된다.

이러한 두 가지 문제를 해결하기 위한 방법으로 전 시점의 학습 결과를 다음 시점의 초기값으로 이용하여 파라미터를 업데이트하였다. 전 시점의 단어 벡터를 다음 시점의 초기값으로 활용하게 되는 경우, 해당

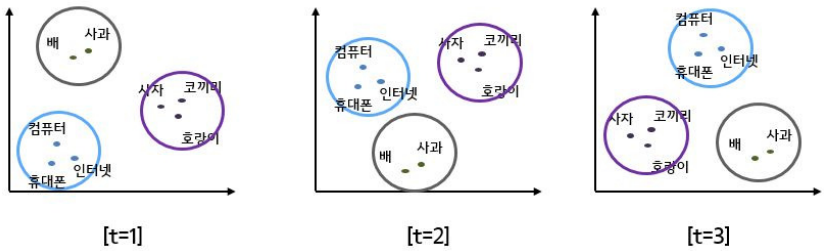

Figure 2: *Word vectors for each viewpoint consisting of random initial value settings (d = 2).*

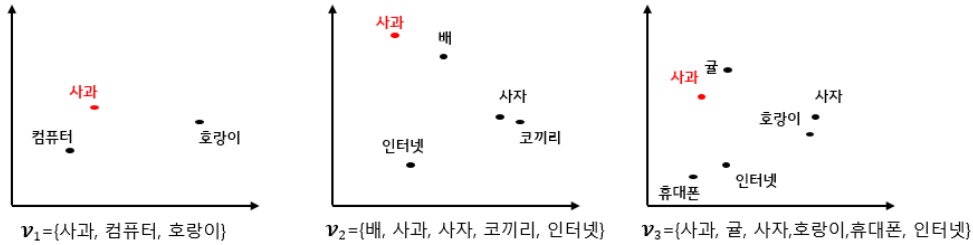

Figure 3: *Viewpoint corpus made up of different words (d = 2).*

시점에서 학습이 이루어지더라도 전 시점의 결과와 비슷한 형태를 갖게된다. 또한 다음 시점에서 이전 시점에 나타난 단어가 등장하지 않더라도 이전 시점의 임베딩 결과를 활용할 수 있게되어 OOV 문제를 해결할 수 있다. 이와 같은 방법으로 구축한 임베딩 행렬의 크기는 독립적인 t 시점에서의 사전 크기 V_t가 아닌 이전 시점까지의 단어들이 쌓여서 만들어진 $V_t \cup V_{t-1} \cup \cdots \cup V_1$ 크기로 정의된다.

3.2. 프로크러스티즈 직교 행렬

임베딩 행렬을 구축한 이후, 각 행렬을 시점에 따라 정렬하여 서로 다른 시점에서 나타난 단어를 비교하기 위해 정렬 문제를 해결하여야 한다. 전 시점을 이용하여 다음 시점의 파라미터를 업데이트 하더라도 의미변화가 일어나는 단어는 무수히 많은 단어 중에서 극히 소수일 것이라 가정하고, 대부분의 단어 벡터가 시점별로 유사하게 나타나도록 임베딩 행렬을 정렬하였다. 정렬 후에 시점별로 단어 벡터가 다르게 나타나는 경우를 의미변화가 일어났다고 판단하기 위하여 프로크러스티즈 직교 정렬(procrustes orthogonal align)을 통하여 시점별 임베딩 행렬을 정렬한다.

프로크러스티즈 분석은 기하적 공간상에서 두 행렬의 형상 비교를 위하여 한 개체를 다른 개체 쪽으로 적합시키는 방법이다 (Choi 등, 2009). 임베딩 행렬에서 각 단어들은 행렬의 형상점(landmark)이고, 이웃한 시점별로 두 임베딩 행렬의 형상점에 따라 축소, 확대, 회전 등을 통해 두 행렬의 차이 크기가 가장 작아지도록 하는 행렬을 찾는다. Figure 4은 형상이 다른 두 개체를 비교하기 위해 한 개체가 다른 개체 쪽으로 적합되는 과정을 나타낸 그림이다 (Klingenberg, 2015).

프로크러스티즈 문제의 최적의 해는 특이값 분해 통해 찾을 수 있는 것으로 알려져 있다 (Schonemann, 1966). t 시점의 임베딩 행렬 $W^{(t)}$와 $t+1$ 시점의 임베딩 행렬 을 정렬하기 위한 방법은 식 (3.1)에서 $W^{(t)}$와 $W^{(t+1)}$을 가장 가깝게 매핑하는 $R^{(t)}$를 찾는 것이다.

$$R^{(t)} = \arg\min_{Q} \left\| QW^{(t)} - W^{(t+1)} \right\|_F^2 . \tag{3.1}$$

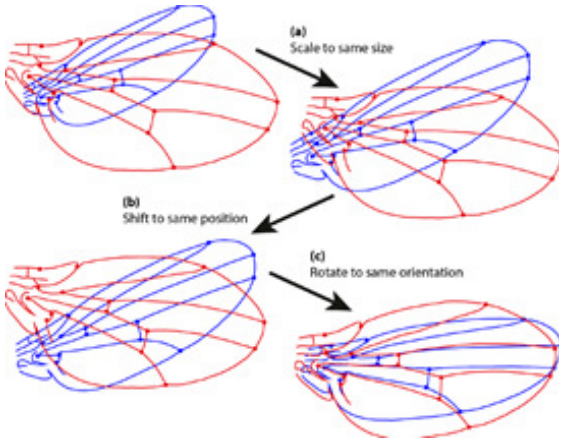

Figure 4: *Procrustic shape analysis: 3-step transformation process (Klingenberg, 2015).*

이때, $R^{(t)}$를 찾는 과정은 다음과 같다:

$$\begin{aligned}
R^{(t)} &= \arg\min \left\| QW^{(t)} - W^{(t+1)} \right\|_F^2 \\
&= \arg\min \left\langle QW^{(t)} - W^{(t+1)}, QW^{(t)} - W^{(t+1)} \right\rangle \\
&= \arg\min \left\| W^{(t)} \right\|_F^2 + \left\| W^{(t+1)} \right\|_F^2 - 2 \left\langle QW^{(t)}, W^{(t+1)} \right\rangle \\
&= \arg\max \left\langle Q, W^{(t+1)} W'^{(t)} \right\rangle \\
&= \arg\max \langle Q, U\Sigma V' \rangle \\
&= \arg\max \langle U'QV, \Sigma \rangle \\
&= U \left(\arg\max \langle Q', \Sigma \rangle \right) V' \\
&= UV'
\end{aligned}$$

3.3. 의미변화 정량화

이웃한 연도별로 공통으로 나타난 단어들을 통해 직교 프로크러즈티즈 문제의 최적 해를 구함으로써 각 시점의 임베딩 행렬을 정렬 문제를 해결한 후, 단어의 의미변화를 정량화하기 위해 두 벡터의 유사성을 계산하는 방법인 코사인 유사도(cosine similiarity)를 이용하였다. 의미변화를 탐지하기 위한 척도로 한 단어에 대한 시점별 코사인 유사도와 각각의 시점 내에서 한 단어와 다른 단어들의 코사인 유사도를 계산하여 의미변화의 정도를 수치화하여 결과를 도출하였다. 우선 단어의 의미변화를 정량화하기 위해 단어 w_i의 t 시점의 단어 벡터와 $t+1$ 시점의 단어 벡터의 코사인 유사도를 계산한다. 이를 단어의 자기유사도라 하고 식 (3.2)와 같이 이웃한 연도별로 자기 유사도를 계산한다.

$$s^{(t)}(w_i) = \text{cos-sim}\left(w_i^{(t)}, w_i^{(t-1)}\right). \tag{3.2}$$

또한, 같은 시점 내에서 특정 단어가 어떤 단어와 비슷한 의미를 갖는지를 파악하기 위해 단어 간 코사인 유사도를 계산한다. t 시점 내에서 단어 w_i와 단어 w_j와의 유사도를 계산하여 두 단어 간의 시간에 따른 유사성을 살펴본다. 이를 쌍별 유사도라 하고 식 (3.3)과 같이 각 시점별로 단어 간 유사도를 계산한다.

$$s^{(t)}\left(w_i, w_j\right) = \text{cos-sim}\left(w_i^{(t)}, w_j^{(t)}\right). \tag{3.3}$$

Table 1: Number of documents and words at the time of the presidential speech recording

Presidential speech recording (τ = 7)			
Time point (t)	Time (year)	# documents	# words (min = 3)
t = 1	1950 ~ 1959	873	7,059
t = 2	1960 ~ 1969	764	8,145
t = 3	1970 ~ 1979	528	7,630
t = 4	1980 ~ 1989	834	8,790
t = 5	1990 ~ 1999	1,462	10,431
t = 6	2000 ~ 2009	1,687	11,412
t = 7	2010 ~ 2018	815	8,965
Total	1950 ~ 2018	7,088	19,240

4. 실제 자료 분석 및 결과

단어의 의미변화를 파악하기 위해서는 오랜 기간 동안 축적된 텍스트 데이터를 필요로 한다. 단어의 의미가 변화하는 과정은 신조어와 다르게 단기간에 이루어지기보다는 새로운 의미로 서서히 사용되기 때문에 이런 변화 과정을 담고 있는 텍스트 자료가 필요하다. 따라서 본 연구에서는 두 가지 데이터를 사용하고 각 데이터에서 방법론을 적용하여 나타난 결과를 제시하였다.

4.1. 자료 설명

한글의 의미변화를 파악하기 위해 사용된 자료는 대통령 연설 기록문 데이터와 신문기사 데이터이다. 대통령 연설 기록문은 국가기록원 대통령 사이트(https://www.pa.go.kr)와 대한민국 정책브리핑 연설문 아카이브 사이트(https://www.korea.kr)에서 1950년부터 2018년까지 수록된 대통령 연설문 텍스트를 크롤링하여 자료를 구축하였다. 자료 기간 동안의 총 문서 수는 7,088개이고 출현빈도가 3 이상인 단어의 개수는 19,240개이다. Table 1은 대통령 연설 기록문에 대한 설명을 정리한 표이다. 1950년부터 2018년까지의 기간을 10년 단위로 시점을 구분하여 총 7개의 시점별 문서 집합으로 분할하였다.

신문기사 데이터는 공공데이터 포털의 뉴스 빅데이터 자료를 이용하였다(www.data.go.kr/dataset/15012945/fileData.do). 특정 키워드가 포함된 기사에 대한 메타 자료로서 기사의 게시일과 기사 본문의 내용을 이용하여 한글의 의미변화를 파악하기 위한 자료로 구축하였다. 자료 기간은 1990년부터 2018년이고 총 기사 수는 2,264,284개이고 출현빈도가 10 이상인 단어 수는 81,992개이다.

Table 2는 신문기사 데이터에 대한 설명을 정리한 표이다. 대통령 연설 기록문에 비하여 자료 기간이 짧고 문서와 단어의 수는 훨씬 많았기 때문에 3년 단위로 시점을 구분하여 총 10개의 시점별 문서 집합으로 분할하였다.

4.1.1. 하이퍼파라미터 설정

대통령 연설 기록문과 신문기사 데이터를 이용하여 단어의 의미가 이전과 달라진 단어들이 어떻게 변화하였는지 살펴보았다. 이를 위하여 우선 각 자료에서의 임베딩 성능을 평가하기 위하여 Park 등 (2018)이 구축한 한글 임베딩 성능 평가 데이터셋을 이용하였다. 한글 임베딩 성능 평가 데이터셋은 한글 단어 벡터에 대한 의미 분석 능력을 평가하기 위한 테스트 셋으로 구문적 특징 문항 5,000개와 의미적 특징 문항 5,000개를 합한 총 10,000개의 문항으로 이루어져있다. 이중에서 5가지 카테고리가 있는 5,000개의 의미적 특징 문항을 통해 각 자료에 대한 임베딩 성능을 평가하고 이를 바탕으로 워드투벡터 모형의 하이퍼파라미터를 설정하였다.

Table 2: Number of documents and words at the time of the presidential speech recording

Presidential speech recording (τ = 10)			
Time point (t)	Time (year)	# documents	# words (min =10)
t = 1	1990 ~ 1992	11,149	10,185
t = 2	1993 ~ 1995	23,320	11,191
t = 3	1996 ~ 1998	29,713	13,903
t = 4	1999 ~ 2001	73,976	18,622
t = 5	2002 ~ 2004	75,585	18,981
t = 6	2005 ~ 2007	111,948	22,580
t = 7	2008 ~ 2010	338,616	36,383
t = 8	2011 ~ 2013	539,529	45,070
t = 9	2014 ~ 2016	579,411	45,554
t = 10	2017 ~ 2018	481,037	36,094
Total	1950 ~ 2018	2,264,284	81,992

Table 3: Example of semantic features inference questions for Korean words

카테고리	질문(Query)	답안(Answer)
수도-국가	아테네 : 그리스 = 바그다드 : **?**	이라크
남성-여성	남자 : 여자 = 아버지 : **?**	어머니
인물-국적	간디 : 인도 = 나폴레옹 : **?**	프랑스
국가-언어	프랑스 : 불어 = 미국 : **?**	영어
기 타	개 : 강아지 = 소 : **?**	송아지

Table 4: Hyperparameters finally set for each data

Hyperparameter	Presidential speech recording	Newspaper article data
embedding size (d)	100	200
window size	5	5
epochs	20	10
min_count	3	10

Table 3은 각 카테고리의 예시 질문으로, $A : B = C$: ?의 의미 관계를 포착하는 정도에 따라 임베딩 성능을 평가하였다. 단어 A와 단어 B의 관계를 포착하여 단어 C와 이 관계에 대응되는 단어를 유추하는 문제로 정확도가 높을수록 단어의 의미를 잘 내포한다고 판단하였다.

4.2. 분석 결과

의미 추론 능력을 통한 임베딩 성능 평가 결과, 대통령 연설 기록문과 신문기사 데이터 모두 CBOW 방식보다는 skip-gram 방식에서 정확도가 더 높게 나타났고 단어 수가 많은 신문기사 데이터 셋은 임베딩 크기를 200차원으로 설정하는 경우 정확도가 가장 높게 나타났다.

이에 비해 단어 수가 적은 대통령 연설 기록문의 경우 임베딩 크기를 100차원으로 설정한 경우에 정확도가 가장 높게 나타났다. 경험적인 방법을 통해 하이퍼파라미터를 설정하고 최종적으로 사용된 모형의 하이퍼파라미터는 Table 4와 같다.

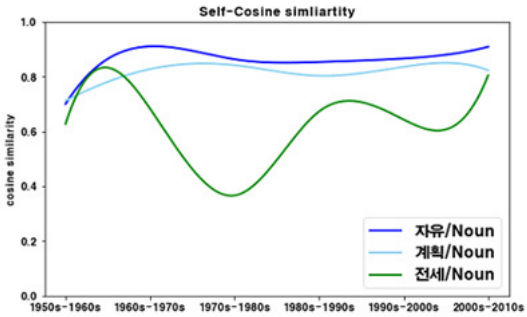

Figure 5: *Self-similarity graph for freedom(자유), plan(계획), and jeonse(전세).*

Table 5: Top 5 words with high similarity to jeonse(전세) by time point

Year	Word list
1950 ~ 1959	서광/Noun, 현상/Noun, 패망/Noun, 시기/Noun, 불리하다/Adjective
1960 ~ 1969	섬멸/Noun, 남쪽/Noun, 적/Noun, 기울다/Verb, 참혹/Noun
1970 ~ 1979	급전/Noun, 역전/Noun, 주월/Noun, 최후/Noun, 상황/Noun
1980 ~ 1989	월세/Noun, 깨지다/Verb, 손해/Noun, 집값/Noun, 무너지다/Verbt
1990 ~ 1999	월세/Noun, 돈/Noun, 시중/Noun, 자금/Noun, 만기/Noun
2000 ~ 2009	주택/Noun, 투기/Noun, 월세/Noun, 임대/Noun, 아파트/Noun
2010 ~ 2018	금리/Noun, 월세/Noun, 시세/Noun, 완화/Noun, 가계부채/Noun

4.2.1. 대통령 연설 기록문 분석 결과

하이퍼파라미터 설정을 통해 최적의 모형을 이용하여 각 자료에서 단어의 의미가 어떻게 변화하는지 살펴보았다. 대통령 연설문 기록에서 나타난 결과로 “전세”라는 단어가 갖는 의미가 1970년대($t = 3$)를 기점으로 달라진 것을 확인하였다. Figure 5을 통해 본 연구에서 단어의 의미가 변화하였다고 판단한 기준을 알 수 있다. “자유”와 “계획”은 이웃한 연도별로 자기 유사도를 살펴보았을 때, 전체적으로 0.75 이상의 높은 값으로 일정하게 유지되고 있음을 볼 수 있다. 반면에 “전세”는 1970년대($t = 3$)와 1980년대($t = 4$)사이에서 자기 유사도가 급격하게 감소하였고 1980년대 이후로 다시 값이 상승하여 일정하게 유지된 것으로 나타났다.

이와 같이 두 시점 사이의 자기유사도가 떨어진 시점을 단어의 의미변화가 발생하였다고 판단하고, 각 시점에서 쌍별 유사도를 통해 어떤 단어와 유사하게 나타났는지 살펴보았다. Table 5는 각 시점마다 전세와 유사한 단어로 나타난 상위 5개 단어를 살펴본 결과이다. 1950년대부터 1970년대까지 전세와 유사한 의미를 갖는 단어로 “서광”, “섬멸”, “급전” 등이 나타났고 이러한 단어를 통해 전세가 ‘전쟁이나 경기 따위의 형세나 형편’을 의미하는 단어로 사용됨을 알 수 있다. 1980년대부터는 전세와 가까운 단어로 “월세”, “주택”, “금리” 등이 나타났고 전세가 ‘계약에 의하여 일정 기간 동안 빌려주는 일’을 뜻하는 단어로 사용됨을 알 수 있다. 이를 통해 동음이의어인 전세의 두 가지 뜻 중에서 1970년대까지는 전쟁의 형세를 뜻하는 단어로 쓰이다가

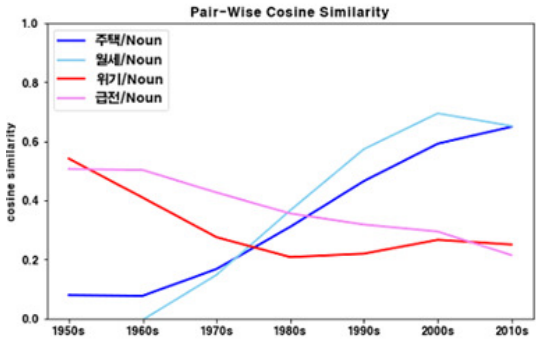

Figure 6: *Words that increase and decrease similarity to jeonse(전세).*

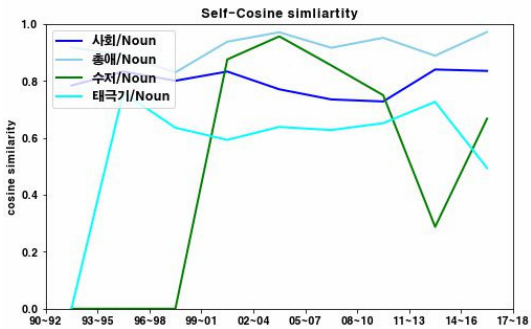

Figure 7: *Self-similarity graph of society(사회), love(총애), spoon(수저), and Taegeukgi(태극기).*

1980년대 이후에는 집값을 뜻하는 단어로 바뀌는 것을 포착하였다. Figure 6는 전세와 4개의 단어 각각의 쌍별 유사도를 시점별로 표현한 그림이다. 그래프를 통해 알수 있듯이, “주택”과 “월세”의 유사도는 점점 증가하고 “위기”와 “급전”은 점점 낮아지는 형태임을 알 수 있다.

4.2.2. 신문기사 데이터 분석 결과

신문기사 데이터에서 의미변화가 나타난 단어로는 “수저”, “태극기” 등이 있었다. Figure 7에서 “사회”와 “총애”의 자기 유사도는 전체적으로 높은 수준으로 일정하게 유지되고 있고, 반대로 “수저”와 “태극기”는 특정 시점에서 자기 유사도가 낮아지는 형태를 보이고 있다. 이를 통해 수저의 경우 2011~2013년(t = 8)까지의 의미와 2014~2016년(t = 9)까지의 의미가 변화하였음을 알 수 있다. 태극기의 경우 2017~2018년(t = 10)에서 자기 유사도가 낮아진 형태를 보이고 이 시점들을 기준으로 쌍별 유사도를 통해 주변 단어의 변화를 살펴보았다. 이러한 단어들이 어떻게 변하였는지 시각적으로 파악하기 위한 방법으로 200차원의 각 단어 벡터를 PCA를 통해 3차원으로 축소하여 시간이 변함에 따라 어떤 단어들과 가까워지는지 살펴보았다. Figure 8은 태극기와 수저의 시점별 단어 벡터들과 각 시점에서 가깝게 나타난 주변 단어 벡터들을 선정하여 3차원의 공간으로 축소하여 그린 결과이다.

태극기의 경우, 초기에는 “현수막”, “걸리다”, “국기”와 같은 단어들과 가깝게 나타나지만 시간이 흐를수록 “집회”, “지지자”, “시위” 등의 단어와 가깝게 유사되는 것을 볼 수 있다. 이를 통해 역사적인 사건과 사회적인 원인으로 인해 태극기가 우리나라의 국기를 뜻하는 의미에서 집회나 시위를 대표하는 단어로 의

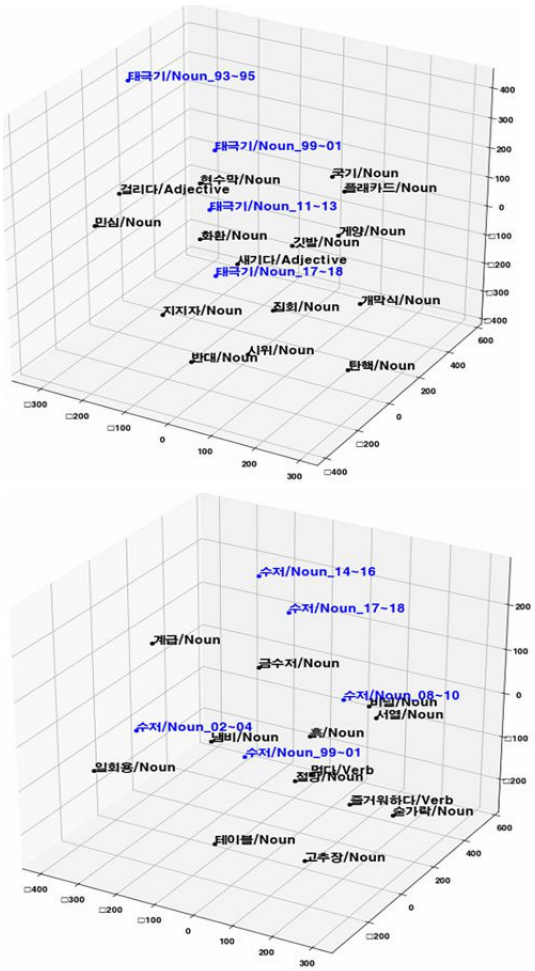

Figure 8: *3D visualization result of the semantic change of Taegeukgi(태극기) and spoon(수저).*

미가 변하고 있는 것을 알 수 있다. 또한 최근 태극기와 유사한 의미를 갖는 단어로 나타난 “촛불”은 “추모”, “염원” 등의 의미에서 태극기처럼 집회와 시위를 대표하는 단어로 쓰임을 포착하였다.

마찬가지로 수저의 시점별 단어 벡터들과 각 시점에서 유사하게 나타난 단어 벡터들을 3차원으로 축소하여 그린 결과를 살펴보면, 1999년~2001년에는 “먹다”, “설탕”, “일회용”, “고추장” 등의 단어와 가깝게 나타났다가 2014년 이후에서는 “계급”, “금수저”, “서열” 등의 단어와 점점 유사해지는 것을 알 수 있다. 이를 통해 식기류를 뜻하는 본래의 의미에서 계층과 계급을 뜻하는 단어로 변화하고 있음을 알 수 있다. Table 6는 수저, 태극기, 촛불이 각각 어떻게 의미가 변화하고 있는지를 정리한 표이다.

5. 결론

본 논문에서는 단어의 의미변화를 정량화하기 위한 통계적 방법론을 모색하고 이를 한글 데이터에 적용하여 한글의 의미변화를 파악하고자 하였다. 분석 결과로, 언어학적인 의미변화는 아니지만 동음이의어의 사용이

Table 6: Moving paths of semantic changes of spoon, Korean flag, and candlelight

Word	Moving away	Moving towards
수저/Noun	일회용/Noun, 고추장/Noun, 백반/Noun, 비닐봉투/Noun, 즐거워하다/Verb	금수저/Noun, 흙/Noun, 아등바등/Adverb, 계급/Noun, 헬조선/Noun
태극기/Noun	게양/Noun, 새기다/Verb, 깃발/Noun, 응원/Noun, 플래카드/Noun	집회/Noun, 시위/Noun, 촛불집회/Noun, 박사모/Noun, 탄핵/Noun
촛불/Noun	기도/Noun, 염원/Noun, 콘서트/Noun, 추모/Noun, 등불/Noun	촛불집회/Noun, 규탄/Noun, 시국선언/Noun, 투쟁/Noun, 민심/Noun

시대에 따라 다르게 사용된 것과 사회적 현상을 반영하는 새로운 비유적인 표현을 포착할 수 있었다.

대통령 연설 기록문의 경우, 기간은 길지만 문서의 길이가 짧고 충분히 학습될 수 있는 단어의 수가 부족하였다. 반대로 신문기사 데이터의 경우 문서의 길이와 학습되는 단어 수가 많아 임베딩 성능은 좋게 나타났지만 문서의 기간이 짧아 의미변화를 포착하기에는 어려움이 있었다. 영어권의 경우, 단어의 의미변화를 탐지하기 위해 다양한 연구가 진행되고 있다. Corpus of historical American (COHA) 데이터셋의 경우, 1800년대부터 2000년대까지 총 4억개의 단어가 수록되어있다 (Davies, 2010). Google books N-gram 데이터셋의 경우, 1800년부터 2000년대까지 총 300만 권의 책을 통해 2,000억개의 단어가 수록되어 있다 (Lin 등, 2012). 이러한 오픈소스로 제공되는 풍부한 데이터를 활용하여 의미변화를 탐지하기 위한 연구가 진행 중이다. 영어권의 의미변화 탐지 연구와 같이 한글에서 나타난 의미변화를 파악하여 더 의미 있는 결과를 제시하기 위해 추후 연구에서는, 데이터 수집 차원에서 영어권의 데이터와 같이 긴 기간의 데이터 수집과 다양한 단어를 포함한 코퍼스 셋을 구축하여 연구를 개선시키고자 한다. 또한 임베딩 측면에서 개선시키기 위해 작은 데이터에서 임베딩 성능이 좋은 모형을 적용하여 개선시키고자 한다.

마지막으로 기존의 방법론에 추가적으로, 자기 유사도에서 값이 급격하게 감소하는 지점을 찾기 위해 이상점 탐지(outlier detection) 기법을 응용하여 의미변화가 일어난 단어들을 선별하고자 한다. 자동적인 의미변화 추출 방안을 제시하고 이에 대한 정확도 평가를 위해 의미변화 단어 목록을 구축하여 연구를 개선시키는 방향으로 향후 연구를 진행하고자 한다.

References

Cho NH (2004). Acceptance and development of the theory of semantic change, *Linguistics*, **43**, 461–485.

Choi TH, Choi YS, and Shin SM (2009). A study on the relationship between player characteristic factors and competitive factors of tennis grand slams competition using canonical correlation biplot and procrustes analysis, *Korean Journal of Applied Statistics*, **22**, 855—864.

Davies M (2010). *The Corpus of Historical American English: COHA*, BYE, Brigham Young University.

Deerwester S, Dumais ST, Furnas GW, Landauer TK, and Harshman R (1990). Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, **41**, 391–407.

Golub GH and Reinsch C (1970). Singular value decomposition and least squares solutions, *Umerische Mathematik*, **14**, 403–420.

Hamilton WL, Leskovec J, and Jurafsky D (2016). *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*.

Harris ZS (1954). Distributional structure, *Word*, **10**, 146–162.

Kim Y, Chiu YI, Hanaki K, Hegde D, and Petrov S (2014). *Temporal Analysis of Language through Neural Language Models*.

Klingenberg CP (2015). Analyzing fluctuating asymmetry with geometric morphometrics: concepts, methods, and applications, *Symmetry*, **7**, 843–934.

Kulkarni V, Al-Rfou R, Perozzi B, and Skiena S (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, 625–635.

Lin Y, Michel JB, Aiden EL, Orwant J, Brockman W, and Petrov S (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, 169–174.

Matveeva I, Levow G, Farahat A, and Royer C (2007). Term representation with generalized latent semantic analysis, *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*. Available from: https://doi.org/10.1075/cilt.292.08.

Mikolov T, Le QV, and Sutskever I (2013). *Exploiting Similarities among Languages for Machine Translation*.

Naptali W, Tsuchiya M, and Nakagawa S (2009). Word co-occurrence matrix and context dependent class in lsa based language model for speech recognition, *International Journal of Computers*, **1**.

Park S, Byun J, Baek S, Cho Y, and Oh A (2018). Subword-level Word Vector Representations for Korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, **1**, 2429–2438.

Sahlgren M (2008). The distributional hypothesis, *Italian Journal of Disability Studies*, **20**, 33–53.

Schönemann PH (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, **31**, 1–10.

Yoon P (2013). *Korean semantic lecture*, Youkrack.

Received January 4, 2021; Revised February 13, 2021; Accepted February 15, 2021

단어 임베딩 기법을 이용한 한글의 의미 변화 파악

선현석[a], 이영섭[b], 임창원[1,a]

[a]중앙대학교 응용통계학과; [b]동국대학교 통계학과

요 약

최근 들어 많은 사람들이 자신의 관심사를 SNS에 게시하거나 인터넷과 컴퓨터의 기술 발달로 디지털 형태의 문서 저장이 가능하게 됨으로써 생성되는 텍스트 자료의 양이 폭발적으로 증가하게 되었다. 이에 따라 수많은 문서 자료로부터 가치 있는 정보를 창출하기 위한 기술의 요구 또한 증가하고 있다. 본 연구에서는 대통령 연설 기록문과 신문기사 공공데이터를 활용하여 한글 단어들이 시간에 따라 어떻게 의미가 변화되어 가는지를 통계적 기법을 통해 발굴하였다. 이를 이용하여 한글의 통시적 변화 연구에 활용할 수 있는 방안을 제시한다. 기존 언어학자나 원어민의 직관에 의해 연구되던 한글의 이론적 언어 현상 연구에서 벗어나 누구나 사용할 수 있는 공공문서를 통해 수치화된 값을 도출하고 단어의 의미변화 현상을 설명하고자 한다.

주요용어: 의미변화, 워드투벡터, 프로크러스티즈 정렬, 말뭉치 언어학

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(NRF-2017M3C4A7083281).

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: clim@cau.ac.kr