대한한의학원전학회지
2019년 02월 32권 1호

pISSN 1229-8328
J Korean Med Classics 2019:32(1) 061-074
https://doi.org/10.14369/jkmc.2019.32.1.061

JKMC
Journal of Korean Medical Classics

한의학 고문헌 데이터 분석을 위한 단어 임베딩 기법 비교: 자연어처리 방법을 적용하여

한국한의학연구원 연구원
오준호*

Comparison between Word Embedding Techniques in Traditional Korean Medicine for Data Analysis: Implementation of a Natural Language Processing Method

Oh Junho*
Researcher at Korea Institute of Oriental Medicine

Objectives : The purpose of this study is to help select an appropriate word embedding method when analyzing East Asian traditional medicine texts as data.

Methods : Based on prescription data that imply traditional methods in traditional East Asian medicine, we have examined 4 count-based word embedding and 2 prediction-based word embedding methods. In order to intuitively compare these word embedding methods, we proposed a "prescription generating game" and compared its results with those from the application of the 6 methods.

Results : When the adjacent vectors are extracted, the count-based word embedding method derives the main herbs that are frequently used in conjunction with each other. On the other hand, in the prediction-based word embedding method, the synonyms of the herbs were derived.

Conclusions : Counting based word embedding methods seems to be more effective than prediction-based word embedding methods in analyzing the use of domesticated herbs. Among count-based word embedding methods, the TF-vector method tends to exaggerate the frequency effect, and hence the TF-IDF vector or co-word vector may be a more reasonable choice. Also, the t-score vector may be recommended in search for unusual information that could not be found in frequency. On the other hand, prediction-based embedding seems to be effective when deriving the bases of similar meanings in context.

Key words : Word embedding, East Asian traditional medicine, Korean Medicine, data analysis, natural language processing

* Corresponding Author : Oh Junho.
Korea Institute of Oriental Medicine, 1672 Yuseong-daero, Yuseong-gu, Daejeon, 34054
Tel +82-42-868-9317, E-mail : junho@kiom.re.kr
Received(January 18, 2019), Revised(February 11, 2019), Accepted(February 11, 2019)

I. 서론

데이터 분석은 데이터를 정리하거나 변형한 뒤 적절한 방식으로 모델링하여 그 속에서 유용한 정보를 발견하거나, 원하는 결론을 도출하거나, 인간의 결정을 돕는 과정이다.[1] 그간 데이터 분석은 미리 정해진 형태에 따라 체계적으로 잘 정리된 정형 데이터(structured data)를 대상으로 이루어져 왔으나, 최근에는 음성・영상・텍스트 같이 체계적으로 정리되지 않은 비정형 데이터(unstructured data)를 대상으로 한 분석도 활발하게 시도되고 있다. 이 가운데 텍스트는 대표적인 비정형 데이터로서 양적으로 가장 풍부한 데이터이다. 오늘날 인터넷 상에서 교류되고 있는 대부분의 정보들이 문자로 기록되고 있기 때문이다.

텍스트는 인간이 지식을 기록해 왔던 가장 오래되고 확고한 방식이라는 점에서 그 의미가 현대에 국한되지 않는다. 텍스트 데이터는 현재에도 끊임없이 생산되고 있지만, 이미 선조들에 의해 만들어진 텍스트 또한 적지 않다. 고서(古書)나 고문서(古文書) 등의 고문헌이 바로 그것이다. 이러한 텍스트들이 대부분 기록물로서 보존되는데 그치고 있으나 오늘날에도 여전히 사용되고 있는 분야가 있다. 바로 의학과 관련된 동아시아 전통지식이다. 동아시아 전통의학은 최소 2천년 넘게 지속되어 온 인류의 유산으로, 고인(古人)이 남긴 수많은 텍스트로 이루어져 있다. 이러한 고문헌들은 오랫동안 새롭게 의학을 배우는 이들에게 길잡이가 되었을 뿐만 아니라 새로운 문헌이 탄생하는 토양이 되어 주었다. 그리고 지금도 전통의학을 기반으로 한 치료의 근거가 되고 있다.

동아시아 전통의학 고문헌은 상당히 방대하다. 정확히 집계할 수는 없으나, 중화민국 성립 이전의 주요 중의학(中醫學) 저작을 전산화한 ≪중화의전(中華醫典)≫에는 1천여 종의 고문헌 4억여 자(字)가 수록되어 있다.[2] 또 한국에서 집필되었거나 한국에서 많이 읽힌 동아시아 전통의학 문헌을 전산화한 ≪한의학고전DB≫에는 77종의 의서 846만여 자(字)가 온라인에 공개되어 있다.[3] 물론 이들 문헌을 빅데이터(big data)라고 하기는 어려움이 있다. 하지만 의학에 대한 지식이 집적된 양질의 스몰데이터(small data)라고 하기에는 부족함이 없다. 따라서 텍스트로 이루어진 이들 전통의학 문헌을 컴퓨터의 힘을 빌려 분석해 낼 수 있다면 기존의 연구나 학습을 도울 수 있을 것이다.

그간 전통의학의 치료 지식을 요약하거나 시각화하기 위해 본초와 방제[4] 혹은 침구 경혈 조합[5]을 데이터로 삼아 분석하려는 시도들이 있었다. 또 빈도를 기반으로 한 텍스트 분석 방법이 학계의 지식과 관련이 있는지 검토하거나[6] 정성적인 연구 방법을 보완할 수 있는지 탐색해 보는 연구도[7] 있었다. 하지만 이러한 연구가 보다 풍성해지기 위해서는 전

1) Wikipedia. Data analysis. [cited on Jan 12, 2019]. Avaiable from: https://en.wikipedia.org/wiki/Data_analysis
2) 中华医典. 中国中医药学会、湖南电子音像出版社、嘉鸿科技开发有限公司. 2003.
3) 한국한의학연구원. 한의학고전DB. [cited on Jan 12, 2019]. Avaiable from: https://mediclassics.kr
4) 송영섭 외 3명. 데이터 마이닝을 이용한 대변과 약물간의 연관성 분석 -방약합편을 중심으로-. 대한한의진단학회. 2012. 16(2). pp. 33-45.
김기욱, 김태열, 이병욱. 본초 목록을 이용한 방제의 본초 구성 자동 추출 방법. 대한한의학원전학회지. 2014. 27(3). pp.155-166.
오준호. HF-IFF: TF-IDF를 응용한 병증-본초 연관성(relevancy) 측정과 본초 특성의 시각화 -청강의감 방제를 대상으로-. 대한본초학회. 2015. 30(3). pp.63-68.
방민우, 김기욱, 이병욱. 구성을 이용한 방제의 효능 추론 및 분류 방법에 관한 연구. 대한한의학방제학회지. 2017. 25(1). pp.29-38.
5) 오준호. 계층적 군집분석(hierarchical clustering)을 통한 침구자생경(鍼灸資生經) 경혈 선택 요인 분석. 대전대학교 한의학연구소 논문집. 2014. 23(1). pp.115-124.
박인수 외 5명 텍스트마이닝을 통한 사암침법 오수혈 사용 패턴 분석. 경락경혈학회지. 2015. 32. pp.66-74.
오준호. 고의서에 나타난 경혈과 병증의 연관성 측정 및 시각화 - 침구자생경 분석 예를 중심으로 -. 경락경혈학회지. 2016. 33. pp.18-32.
6) 오준호. 의학 사상의 유사성은 계량 분석 될 수 있는가. 대한한의학원전학회지. 2018. 31(2). pp.71-82.
7) 김기욱, 김태열, 이병욱. 본초 비율의 순위를 이용한 문헌의 특징 분석 방법 -태평혜민화제국방(太平惠民和劑局方), 난실비장(蘭室秘藏), 소문선명론방(素問宣明論方)을 중심으로 -. 대한한의학원전학회지. 2014. 27(4). pp.73-84.
오월환 외 3명 방제구성을 이용한 유하간(劉河間) 및 이동원(李東垣)의 저작과 『화제국방(和劑局方)』의 특성 비교. 대한한의학원전학회지. 2015. 28(1). pp.55-69.

통의학 분야에 적합한 분석 방법을 모색하는 일이 선행되어야 한다.

텍스트를 데이터로서 분석하려고 할 때에는 무엇보다도 먼저 컴퓨터가 이해할 수 있는 형태로 텍스트를 변환해 주어야 한다. '단어 임베딩(word embedding)'은 이렇게 텍스트를 구성하는 단어들을 수치로 나타내는 방법을 가리킨다. 낱낱의 단어들은 단어 임베딩을 통해 컴퓨터가 연산할 수 있는 수치적인 형태, 즉 실수(實數)로 이루어진 고차원 벡터로 맵핑된다.[8] 이 단어 임베딩은 자연어처리(NLP: natural language processing)[9] 분야의 주요 연구 주제로서 활발하게 연구되고 있다.

본 연구에서는 기존에 자연어처리 분야에서 사용되어 온 단어 임베딩 방식을 동아시아 전통의학 용어에 어떻게 적용하면 좋을지 검토해 보기 위해 수행되었다. 이를 위해 동아시아 전통의학 지식을 대표한다고 할 수 있는 방제를 데이터로 삼아 주요 단어 임베딩 방법을 적용해 보고 그 결과를 비교하였다. 첫 번째 절에서는 연구에 사용한 대상 데이터와 단어 임베딩 방법을 개괄하였고, 두 번째 절과 세 번째 절에서는 단어 임베딩의 2가지 방법인 카운트 기반 방법(Counting-based word embedding)과 예측 기반 방법(Prediction-based word embedding)을 수행하고 설명하였다. 네 번째 절에서는 "방제 생성 게임" 통해 그 결과를 비교해 보았다.

Ⅱ. 본론

1. 대상 데이터 및 연구 방법

본 연구에서는 '한국전통지식포털'에 실려 있는 '전통의학처방' 데이터를 사용하였다. 이 데이터는 한국 특허청에서 전통의학 지식 중 방제 정보를 종합적이고 체계적으로 정리하기 위해 구축한 것으로, 방제의 출전 및 구성 약재, 효능 및 주치 병증, 용법 및 금기사항 등에 대한 정보 등을 담고 있다. 이 방제들은 ≪동의보감(東醫寶鑑)≫·≪방약합편(方藥合編)≫ 등 고문헌, ≪청강의감((晴崗醫鑑)≫·≪동의사상신편(東醫四象新編)≫ 등 근대 저작물, ≪방제학≫·≪동의방제와 처방해설≫ 등 현대 저작물을 포괄하고 있다.[10] 본 연구에서는 온라인에 게재된 이 데이터를 웹 스크래핑(Web Scraping) 방법으로 수집한 뒤에 구성 약재 부분만 추출하여 대상 데이터로 사용하였다.[11]

데이터 선택의 이유는 다음과 같다. 첫째, 동아시아 전통의학에서 '방제'는 치료 지식을 가장 잘 내포하고 있는 자료로서 대표성을 가진다. 둘째, 한국전통지식포털의 전통의학처방 데이터는 방제를 구성하는 본초에 대한 용어가 상당부분 통일되어 있어 복잡한 전처리 작업 없이 사용할 수 있다. 셋째, 데이터 분석을 위해서는 많은 양의 데이터가 필요한데, 해당 방제 데이터는 약 2만건 정도의 규모[12]로 단어 임베딩 방법을 검토하기에 무리가 없다고 판단되었다.

대상 데이터의 구성은 다음과 같다. 대상 데이터는 19,162종의 방제로 이루어져 있으며, 이 방제에는 중복을 포함하여 본초가 모두 140,197번 나타났다. 따라서 1개의 방제는 평균 7.316개의 본초로 구성되었다고 할 수 있다. 절반의 방제가 6개 이하의 본초로 이루어져 있으며, 본초 수를 기준으로 가장 작은 크기의 방제는 본초 1개, 가장 큰 크기의 방제는 본초 59개로 이루어져있다.(Fig. 1. 참조)

본초를 기준으로 보면, 대상 데이터에는 모두 1,841종의 본초가 사용되었다. 1개의 본초는 평균 76.152개의 방제에 사용된 셈이다. 그러나 전체의 절반에 해당하는 본초가 6회 이하로 사용되어 사용

8) Wikipedia. Word embedding. [cited on Jan 12, 2019]. Avaiable from: https://en.wikipedia.org/wiki/Word_embedding

9) 인간과 컴퓨터가 상호 작용하기 위한 방법을 연구하는 학문 분야로서, 자연어로 이루어진 많은 양의 데이터를 처리하고 분석하는 방법을 모색하고 있다.
Wikipedia. Natural language processing. [cited on Jan 12, 2019]. Avaiable from: https://en.wikipedia.org/wiki/Natural_language_processing

10) 특허청. 한국전통지식포털. [cited on Jan 12, 2019]. Avaiable from: http://www.koreantk.com

11) 데이터 수집일 : 2016년 1월 5일

12) 동아시아 전통의학 전체 방제의 규모는 약 9만종 정도로 추정된다.
Peng W. Dictionary of Chinese medicine prescription. 1st ed. Beijing. People's Medical Publishing House. 2005. pp.3-4.

빈도에 있어서 쏠림이 심하게 나타났다.(Fig. 2. 참조) 가장 높은 빈도로 사용된 감초는 7,372회 사용되었고, 두 번째로 많이 사용된 당귀는 감초의 절반에 해당하는 3,834회 사용되었다. 가장 많이 사용된 본초 20종은 다음과 같다.(괄호 안은 전체 사용 빈도)

감초(7,372), 당귀(3,834), 인삼(3,810), 백출(3,137), 진피(2,989), 천궁(2,701), 반하(2,340), 황금(2,052), 백복령(1,990), 방풍(1,943), 백작약(1,823), 황련(1,733), 목향(1,712), 건강(1,500), 황기(1,500), 길경(1,456), 숙지황(1,456), 시호(1,403), 창출(1,390), 대황(1,379)

대상 데이터에는 매우 많은 본초가 등장하는데, 절반 이상은 6회 이하 사용되었다. 단어가 많을수록 연산의 횟수가 증가하기 때문에 본 연구에서는 7회 이상 등장하는 본초 857종을 대상으로 단어 임베딩을 시행하였다.

Fig. 1. Histogram According to Prescription Size

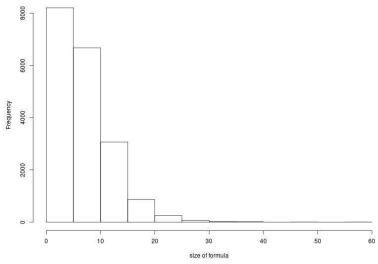
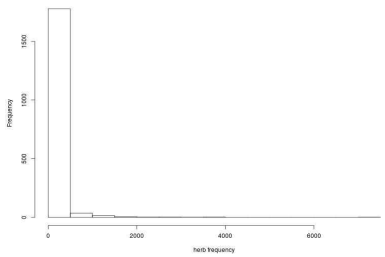


Fig. 2. Histogram According to Herb Frequency



단어 임베딩은 분석 대상에 있는 모든 단어들을 실수(實數)로 이루어진 고차원 벡터로 맵핑하는 것을 뜻한다. 단어 임베딩은 언어학의 분산 가설(distributional hypothesis)을 전제로 한다. 주변에 등장하는 단어가 유사할수록 의미가 서로 비슷해진다는 전제이다.[13] 단어 임베딩에는 크게 2가지 방법이 존재한다. 카운트 기반 단어 임베딩 방법(counting-based word embedding)과 예측 기반 단어 임베딩 방법(prediction-based word embedding)이 그것이다.[14] 카운트 기반 단어 임베딩은 전통적으로 가장 폭넓게 사용되어 온 것으로서 단어의 출현 빈도를 이용한 방법이고, 예측 기반 단어 임베딩은 비교적 최근에 제안된 것으로서 신경망 모델(neural network model)을 적용하여 단어 사이의 관계를 학습시키는 방법이다.

카운트 기반 방법을 통해, 문서에 나타나는 단어의 빈도를 벡터로 나타낸 TF 벡터, 여기에 가중치를 부여한 TF-IDF 벡터를 도출할 수 있다. 이 벡터들은 용어가 문서에서 얼마나 자주 등장하는지를 의미한다. 그러므로 함께 등장하지 않는 단어에 대해서는 용어 사이의 관계를 추정할 수 없다는 단점을 가지고 있다. 이를 개선하기 위해 단어와 단어 사이의 관계를 정량화하는 방법인 $co\text{-}word$ 벡터 및 이에 가중치를 부여한 $t\text{-}score$ 벡터를 사용할 수 있다. 본고에서는 양자를 구분하기 위해 TF 벡터와 TF-IDF 벡터를 1차 벡터(1st order vector), $co\text{-}word$ 벡터와 $t\text{-}score$ 벡터를 2차 벡터(2nd order vector)라고 지칭하였다.

한편, 예측 기반 방법에는 word2vec과 GloVe 등의 방법이 있다. word2vec은 인공신경망을 이용한 단어 임베딩 방법으로, 주어진 '중심 단어'로부터 주변에 등장하는 '주변 단어'를 예측하는 SG(skip-gram) 방식과, 주어진 '주변 단어'로부터 '중심 단어'를 예측하는 CBOW(Continuous Bag-Of-Words) 방식으로 구분된다.[15] word2vec

13) M Sahlgren. The distributional hypothesis. Italian Journal of Linguistics. 2008. 20. pp.33-53.
14) 박대서, 김화종. TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합벡터 제안. 한국정보기술학회논문지. 2018. 16(2). pp.1-16.
15) Mikolov, Tomas, Kai Chen, Gregory S. Corrado and

은 주어진 텍스트를 통해 각각의 경우에 단어가 등장할 조건부 확률을 최대로 하도록 학습 과정을 계속 반복하여 나가는 방법이다. GloVe 역시 신경망 모델을 통해 단어 벡터를 학습 시키는 과정을 거치지만[16], 단어 벡터를 단어가 동시에 등장할 확률에 맞춰 간다는 점에서 차이가 있다.[17]

본 연구에서는 대상 데이터에 카운트 기반 단어 임베딩 방법 4가지와 예측 기반 단어 임베딩 방법 2가지를 적용하고 그 결과를 비교하였다. 자연어처리에서 이들 기법들은 문서와 단어를 기준으로 설정되어 있다. 본 연구에서는 방제를 대상 데이터로 다루고 있기 때문에 자연어처리에서 말하는 '문서'와 '단어'를 각각 '방제'과 '본초'에 적용시켰다. 데이터 연산은 python version 3.6을 사용하였고 카운트 기반 방법의 연산에는 scikit-learn library를, 예측 기반 방법 연산에는 gensim 및 glove_python library를 활용하였다.

2. 카운트 기반 방법(counting-based word embedding)

카운트 기반 방법을 통해 1차 벡터(1st order vector)와 2차 벡터(2nd order vector)로 단어 임베딩 벡터를 도출하였다. 이해를 돕기 위해 아래 예시 데이터를 통해 단어 벡터를 설명하고자 한다.((Table 1 참조)

Table 1. Example Data Set

방제1: 인삼 백출 복령 감초
방제2: 진피 인삼 백출 복령 감초
방제3: 반하 진피 복령 감초
방제4: 반하 진피 인삼 백출 복령 감초

1) 1차 벡터 (1st order vector)

카운트 기반의 1차 벡터는 해당 단어가 문서에 몇 번 등장하는지를 벡터로 표현하는 TF(TF: term frequency) 벡터가 가장 기본이 된다. 보통 단어는 하나의 문서에 등장하지 않거나 1회 이상 등장하기 때문에 등장 횟수를 기준으로 0 또는 자연수의 값을 가진다. 그러나 본 연구에서 다루는 방제 데이터에는 본초가 등장하지 않거나 1회 등장하는 경우만 있기 때문에 0 또는 1의 값을 가지게 된다.[18] TF 벡터는 문서마다 해당 용어가 몇 번 등장하는지를 기준으로 하기 때문에 벡터의 길이가 문서의 수와 같아지게 된다.

예시 데이터에서 TF 벡터를 도출해 보면, 인삼은 [1,1,0,0], 반하는 [0,0,1,1]로 벡터 공간 상에 임베딩 된다(Table. 2 참조). 인삼은 첫 번째와 두 번째 방제에 등장하여 각각 벡터의 첫 번째와 두 번째에 1의 값을 가지게 되었다. 예시 데이터의 경우 이렇게 방제의 수가 4개 이므로 TF 벡터의 길이는 4가 된다.

TF 벡터는 단어의 출현 여부를 보여준다는 점에서 벡터의 의미를 직관적으로 이해할 수 있다. 그러나 매우 자주 쓰이는 단어가 있는가 하면 상당히 드물게 사용되는 단어도 존재한다. TF 벡터에는 이러한 차이가 반영되어 있지 않다. 자주 사용되는 단어의 경우 TF 벡터의 값이 전반적으로 높게 나타날 개연성이 있다.

TF-IDF 벡터는 이러한 단점을 보완하기 위해 TF 벡터에 역문헌빈도(IDF: Inverse Document Frequency)라는 값을 가중치로 부여한 벡터이다

Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781. 2013. [cited on Jan 12, 2019]. Avaiable from: https://arxiv.org/abs/1301.3781

16) Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014. [cited on Jan 12, 2019]. Avaiable from: https://www.aclweb.org/anthology/D14-1162

17) 황현선, 이창기, 장현기, 강동호. 단어 간의 상대적 위치 정보를 이용한 단어 임베딩. 정보과학회논문지. 2018. 45(9). pp.943-949.

18) 만약 방제의 용량을 모두 통일하여 데이터로 만들 수 있다면, 방제의 용량을 횟수 대신 활용할 수도 있을 것이다.

(〔수식 1〕 참조). IDF는 특정 단어가 포함된 문서 개수의 역수 값을 의미한다(〔수식 2〕 참조). 어떤 단어가 여러 가지 문서에 고르게 포함되어 있을수록 IDF 값은 작아지고, 특정 문서에 적게 나타날수록 IDF의 값은 커지게 된다. 대상 데이터에 적용해 보면, 많은 방제에 등장하는 감초와 같은 본초는 작은 IDF의 값을 가지게 된다. 이렇게 특정 방제에 등장하는 본초에 가중치를 부여하여 빈도가 높은 단어의 영향력을 억제하는 효과를 얻을 수 있다.(Table 3 참조)

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t) \quad \text{〔수식 1〕}$$

$$idf(t,d) = \log\frac{1+n_d}{1+df(d,t)} + 1 \quad \text{〔수식 2〕}$$

Table 2. TF Vectors for Example Data

TF	방제1	방제2	방제3	방제4
인삼	1	1	0	1
백출	1	1	0	1
복령	1	1	1	1
감초	1	1	1	1
반하	0	0	1	1
진피	0	1	1	1

Table 3. TF-IDF Vectors for Example Data

TFIDF	방제1	방제2	방제3	방제4
인삼	1.097	1.097	0.000	1.097
백출	1.097	1.097	0.000	1.097
복령	1.000	1.000	1.000	1.000
감초	1.000	1.000	1.000	1.000
반하	0.000	0.000	1.222	1.222
진피	0.000	1.097	1.097	1.097

이러한 방법으로 한국전통지식포털 방제 데이터에서 각 본초에 대해 TF 벡터와 TF-IDF 벡터를 도출하였다. 두 벡터의 길이는 모두 방제의 개수와 같은 19,162열을 가지고 있다.

2) 2차 벡터 (2nd order vector)

단어와 문서 사이의 관계를 통해 도출된 것이 1차 벡터라면, 단어와 단어 사이의 관계를 통해 도출된 것이 2차 벡터이다. 1차 벡터는 동일한 문서에 등장하지 않는 단어의 관계를 적절하게 설명할 수 없다는 한계를 가지고 있다.[19] 2차 벡터는 이점을 어느 정도 보완할 수 있다. 가장 대표적인 2차 벡터로는 공기어($co\text{-}word$) 벡터가 있다. 이것은 단어와 단어가 함께 등장한 빈도를 정리한 것이다.

예시 데이터를 $co\text{-}word$ matrix로 만들어 보았다(Table 4 참조). 인삼과 백출이 함께 등장하는 방제의 개수가 3개이므로 인삼과 백출이 교차하는 곳이 3의 값을 가진다. 이렇게 모든 본초와 본초 사이에 함께 나타나는 횟수를 계산하여 벡터를 구성할 수 있다. 여기서 각 행과 열은 동일한 값을 가지므로, 행렬 구분의 의미가 없으나 편의상 행을 각 본초의 벡터 값으로 볼 수 있다. 즉, 인삼은 [3,3,3,3,1,2], 반하는 [1,1,2,2,2,2]로 공간상의 임베딩 값을 가진다. 1차 벡터가 방제 속에 등장하는 본초의 빈도를 의미했다면, 2차 벡터는 두 본초가 얼마나 자주 함께 등장하였는지를 의미하게 된다.

$co\text{-}word$ matrix는 단순히 함께 등장하는 빈도를 계측한 것으로, 출현 빈도가 높을수록 더 높은 값을 가질 개연성이 있다. 따라서 이를 보정하기 위해 가중치를 부여할 수 있다. 가중치는 co-occurrence score로 사용할 수 있는 다양한 척도(measure)가 있지만[20], 여기에서는 계산이 직관적이고 쉬운 $t\text{-}score$를 사용하였다.[21] 본 연구에서는 분모가 0이 되는 것을 막기 위해 분모에 $add\text{-}one$ smoothing을 시행한 공식을 사용하였다(〔수식 3〕 참조).

19) 김우주, 김동희, 장희원. Word2vec을 활용한 문서의 의미 확장 검색방법. 한국콘텐츠학회논문지. 2016. 16(10). pp.687-692.
20) Stefan Bordag. A Comparison of Co-occurrence and Similarity Measures as Simulations of Context. Computational Linguistics and Intelligent Text Processing. Alexander Gelbukh. Computational Linguistics and Intelligent Text Processing. Berlin. Springer. 2008. pp52-63.
21) 강범모. 언어, 컴퓨터, 코퍼스 언어학(개정판). 서울. 고려대학교출판부. 2011. pp.122-123.

$$t\text{-}score(a,b) = \frac{o_{a,b} - e_{a,b}}{\sqrt{o_{a.b} + 1}} \qquad [수식\ 3]$$

여기서 $o_{a,b}$는 본초a와 본초b가 함께 출현한 관찰값이고, $e_{a,b}$는 본초a와 본초b가 함께 출현할 기댓값이다. 예를 들어 반하는 4개의 방제 가운데 2번 등장하였고, 진피는 3번 등장하였다. 그리고 이들은 실제로 3개의 방제에서 함께 나타났다. 따라서 관찰값$o_{a,b}$는 3이며, 기대값$e_{a,b}$는 $4 \times \frac{2}{4} \times \frac{3}{4}$이 되어 1.5가 된다. 두 본초는 1.5개의 방제에서 함께 나타나리라고 기대할 수 있지만, 실재로는 3번 나타났으므로 서로 더 잘 어울려 출현한다고 할 수 있다. 이렇게 각 본초 조합에 대한 $t\text{-}score$를 모두 도출하여 matrix로 만들면 Table. 5와 같다. 여기서 각 본초의 벡터는 각 행의 값으로, 인삼의 경우 [0.275, 0.275, 0.000, 0.000, -0.250, -0.104], 반하의 경우 [-0.250, -0.104, -0.414, 0.000, 0.207, 0.275]가 된다.

Table 4. *co-word* vectors for Example Data

co-word	인삼	백출	복령	감초	반하	진피
인삼	3	3	3	3	1	2
백출	3	3	3	3	1	2
복령	3	3	4	4	2	2
감초	3	3	4	4	2	3
반하	1	1	2	2	2	2
진피	2	2	2	3	2	3

Table 5. *t-score* vectors for Example Data

t-score	인삼	백출	복령	감초	반하	진피
인삼	0.275	0.275	0.000	0.000	-0.250	-0.104
백출	0.275	0.275	0.000	0.000	-0.250	-0.104
복령	0.000	0.000	0.000	0.000	0.000	-0.414
감초	0.000	0.000	0.000	0.000	0.000	0.000
반하	-0.250	-0.250	0.000	0.000	0.414	0.207
진피	-0.104	-0.104	-0.414	0.000	0.207	0.275

이러한 방법으로 한국전통지식포털 방제 데이터에서 $co\text{-}word$ 매트릭스 및 $t\text{-}score$ 매트릭스를 도출하였다. 각 본초 벡터의 길이는 분석 대상 본초 개수와 같은 857열을 가지고 있다.[22]

3. 예측 기반 방법 (Prediction-based word embedding)

예측 기반 방법은 주어진 문서 집합인 코퍼스를 신경망(neural network)을 이용해 학습하여 주어진 예측을 가장 잘 수행하는 방식으로 단어들의 가중치를 업데이트 시켜 나가는 방법이다. 학습 결과 만들어진 가중치 값이 단어의 임베딩 벡터가 된다. 본 연구에서는 가장 폭넓게 탐구되고 있는 word2vec과 GloVe 2가지 방법을 대상 데이터에 적용하였다.

1) word2vec

word2vec의 SG(Skip-gram) 방식[23]은 '중심 단어'로부터 주변에 등장하는 '주변 단어'를 예측하는 방식이다. 즉, 중심 단어가 주어졌을 때 주변 단어의 특정 조합이 나타날 조건부 확률을 계산한다. 문서를 따라가면서 조건부 확률이 최대가 되는 방향으로 단어 임베딩 벡터를 업데이트 시켜 나간다. 이렇게 되면 유사한 문맥에 등장하는 단어들이 인접하여 나타나게 된다.[24]

본 연구에서는 대상 데이터를 word2vec 방식으로 학습시켰다. 단어 벡터는 100차원으로 임베딩하였으며, 신경망 반복 횟수(epoch)는 30회로 하였다.[25]

22) co-word matrix나 t-score matrix 등 2차 벡터를 도출하기 위해 만들어진 매트릭스(matrix)는 그대로 네트워크를 형성하게 되므로, 네트워크 분석을 위해 사용될 수 있다.

23) 최근에는 성능상의 이유로 SG(Skip-gram) 방식이 주로 사용되고 있기 때문에 본고에서도 이를 사용하였다. 설명 역시 이를 기준으로 하였다.

24) 강형석, 양장훈. 한국어 단어 임베딩 모델의 평가에 적합한 유추 검사 세트. 디지털콘텐츠학회논문지. 2018. 19(10). pp.1999-2008.

25) 임베딩 차원은 사용한 library의 기본값인 100으로 설정하였으며, 신경망 반복 횟수(epoch)는 긴 학습 시간을 고려하여 30회로 설정하였다. 그 이외의 매개변수(parameter) 값들은 library의 기본값을 따랐다.

2) GloVe

GloVe 역시 신경망 모델을 적용하고 있는 단어 임베딩 방식이다. word2vec과는 달리 전체 학습데이터에서 등장하는 단어들의 통계 정보를 활용하는 모델이다. 최초에 *co-word* 벡터를 기반으로 학습 데이터에서 각 단어들이 동시에 나타날 확률을 계산한다. 그런 뒤에 학습을 거쳐 벡터 사이의 유사도가 단어들이 동시에 나타날 최초의 확률과 유사해지는 방향으로 학습을 진행시킨다.[26]

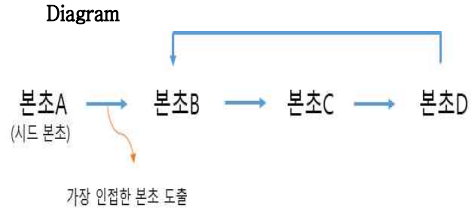
본 연구에서는 대상 데이터를 GloVe 방식으로 학습시켰다. 단어 벡터는 100차원으로 임베딩 하였으며, 신경망 반복 횟수(epoch)는 30회로 하였다.

4. 결과 비교

이처럼 방제에 등장하는 본초들을 자연어처리에서 사용하는 단어 임베딩 방식을 적용하여 고차원 벡터로 임베딩 할 수 있었다. 하지만 이 결과를 객관적인 방법으로 비교하기는 현실적인 어려움이 있다. 현재 한의학 분야에는 WordNet[27] 등과 같은 표준 데이터가 존재하지 않기 때문이다. 차선책으로 전문가 집단이 습득하고 있는 지식과 얼마나 가까운지 검토해 볼 수 있지만, 단어 임베딩 결과는 고차원 벡터로서 직접 육안으로 살펴 서로 비교할 수 있는 형태로 되어 있지 않다.

따라서 본고에서는 불완전하나마 하나의 게임을 제안하고자 한다. 그것은 바로 "방제 생성 게임"이다. 이 게임의 내용은 이렇다. 먼저 시드(seed)가 될 최초의 본초A를 입력 받는다. 그런 뒤에 이 A와 가장 인접한 위치에 있는 본초B를 찾는다. 이때 각각의 본초들은 고차원 벡터로 이루어져 있기 때문에 벡터 사이의 위상 관계를 통해 가장 인접한 본초를 찾을 수 있다.[28] 이러한 원리를 이용하여 본초A와 가장 인접한 다음 본초B를 도출한다. 이런 방식으로 이미 도출된 본초가 다시 나타날 때까지 이 과정을 반복한다. 예를 들어 최초의 본초A를 입력 받아 가장 인접한 관계의 본초B를 찾고, 다시 본초B를 받아 가장 인접한 관계의 본초C를 찾는다. 이렇게 본초C, 본초D, 본초E 등을 차례로 찾을 수 있다. 만약 본초D와 가장 인접한 관계의 본초가 A, B, C 중에 있다면 이 과정을 중단한다. 이렇게 하면 본초A·B·C·D로 이루어진 가상의 방제를 얻을 수 있다. 그리고 관찰자는 이 가상의 방제가 최초의 본초A와 어떤 관계를 가지는지 직관적으로 검토할 수 있다.

Fig. 3. Prescription Generating Game Concept Diagram



방제 생성 게임 결과는 최초의 시드 본초에 따라 좌우된다. 본 연구에서는 사용 빈도가 높은 본초를 중심으로 결과를 확인해 보고자 대상 데이터에서 사용 빈도가 가장 높은 약재 20종을 시드 본초로 선정하여 결과를 도출하였다(Table 6 참조).

이 결과에서 다음과 같은 점을 관찰할 수 있었다. 우선, 카운트 기반 단어 임베딩 결과와 예측 기반 단어 임베딩 결과에서 뚜렷한 차이를 발견할 수 있었다. 카운트 기반 단어 임베딩에서는 함께 자주 사용되는 본초들이 게임의 결과로 도출되는 경향을 보였다. [당귀, 천궁], [인삼, 백출], [황금, 시호], [방풍, 강활], [목향, 빈랑], [건강, 부자], [대황, 망초] 등이 그러한 예이다. 한의학에서 약효를 서로 돕거나 부작용을 억제하기 위해 2가지 이상의 약재들이

26) 황현선, 이창기, 장현기, 강동호. 단어 간의 상대적 위치 정보를 이용한 단어 임베딩. 정보과학회논문지. 2018. 45(9). pp.943-949.

27) 영어 의미 어휘 목록이다. WordNet 2.0은 152,059개의 단어, 115,424쌍의 동의어 집합(synset), 203,145쌍의 단어-의미 쌍(Word-Sense Pairs)으로 구성되어 있다. wordnet. [cited on Jan 12, 2019]. Avaiable from:https://wordnet.princeton.edu/documentation/20-wnstats7wn

28) '인접한 본초'는 고차원 공간상에서의 인접함을 의미하는 것으로, 구체적으로는 코사인 유사도 값이 가장 작은 본초를 뜻한다.

Table 6. Result of Prescription Generating Game

시드 본초 (빈도)	단어 임베딩 방식	방제 생성 결과	시드 본초 (빈도)	단어 임베딩 방식	방제 생성 결과
감초 (7372)	TF 벡터	[감초, 인삼]	백작약 (1823)	TF 벡터	[백작약, 당귀, 천궁]
	TF-IDF 벡터	[감초, 인삼, 백출]		TF-IDF 벡터	[백작약, 당귀, 천궁]
	co-word 벡터	[감초, 인삼, 백출]		co-word 벡터	[백작약, 당귀, 천궁]
	t-score 벡터	[감초, 길경, 전호]		t-score 벡터	[백작약, 황기, 숙지황, 두중, 우슬]
	word2vec 벡터	[감초, 자감초]		word2vec 벡터	[백작약, 백작]
	GloVe 벡터	[감초, 건갈, 갈근]		GloVe 벡터	[백작약, 백복령]
당귀 (3834)	TF 벡터	[당귀, 천궁]	황련 (1733)	TF 벡터	[황련, 황금, 시호]
	TF-IDF 벡터	[당귀, 천궁]		TF-IDF 벡터	[황련, 황금, 시호]
	co-word 벡터	[당귀, 천궁]		co-word 벡터	[황련, 황금, 시호, 감초, 인삼, 백출]
	t-score 벡터	[당귀, 천궁]		t-score 벡터	[황련, 황금, 시호]
	word2vec 벡터	[당귀, 당귀신]		word2vec 벡터	[황련, 천황련]
	GloVe 벡터	[당귀, 관계, 계피]		GloVe 벡터	[황련, 나미주]
인삼 (3810)	TF 벡터	[인삼, 감초]	목향 (1712)	TF 벡터	[목향, 빈랑]
	TF-IDF 벡터	[인삼, 백출]		TF-IDF 벡터	[목향, 빈랑]
	co-word 벡터	[인삼, 백출]		co-word 벡터	[목향, 빈랑]
	t-score 벡터	[인삼, 백출]		t-score 벡터	[목향, 정향, 육두구]
	word2vec 벡터	[인삼, 소초, 영양각, 우황, 금박]		word2vec 벡터	[목향, 토목향, 청귤피, 신석, 생백반]
	GloVe 벡터	[인삼, 자감초, 작약]		GloVe 벡터	[목향, 백두구, 맥아]
백출 (3137)	TF 벡터	[백출, 인삼, 감초]	건강 (1500)	TF 벡터	[건강, 부자]
	TF-IDF 벡터	[백출, 인삼]		TF-IDF 벡터	[건강, 부자]
	co-word 벡터	[백출, 인삼]		co-word 벡터	[건강, 백출, 인삼]
	t-score 벡터	[백출, 인삼]		t-score 벡터	[건강, 오수유]
	word2vec 벡터	[백출, 익지인, 익지]		word2vec 벡터	[건강, 백강]
	GloVe 벡터	[백출, 복령]		GloVe 벡터	[건강, 관계, 계피]
진피 (2989)	TF 벡터	[진피, 감초, 인삼]	황기 (1500)	TF 벡터	[황기, 인삼, 감초]
	TF-IDF 벡터	[진피, 감초, 인삼, 백출]		TF-IDF 벡터	[황기, 인삼, 백출]
	co-word 벡터	[진피, 후박]		co-word 벡터	[황기, 인삼, 백출]
	t-score 벡터	[진피, 후박, 곽향, 백두구, 축사]		t-score 벡터	[황기, 숙지황, 두중, 우슬]
	word2vec 벡터	[진피, 귤홍]		word2vec 벡터	[황기, 두중, 속단, 구척]
	GloVe 벡터	[진피, 청피, 축사]		GloVe 벡터	[황기, 백작, 당귀신]
천궁 (2701)	TF 벡터	[천궁, 당귀]	길경 (1456)	TF 벡터	[길경, 감초, 인삼]
	TF-IDF 벡터	[천궁, 당귀]		TF-IDF 벡터	[길경, 지각]
	co-word 벡터	[천궁, 당귀]		co-word 벡터	[길경, 감초, 인삼, 백출]
	t-score 벡터	[천궁, 당귀]		t-score 벡터	[길경, 전호]
	word2vec 벡터	[천궁, 궁궁, 대두황권, 백렴, 백급, 부용엽]		word2vec 벡터	[길경, 마두령, 자완용, 관동화]
	GloVe 벡터	[천궁, 형개, 형개수]		GloVe 벡터	[길경, 건갈, 갈근]
반하 (2340)	TF 벡터	[반하, 진피, 감초, 인삼]	숙지황 (1456)	TF 벡터	[숙지황, 당귀, 천궁]
	TF-IDF 벡터	[반하, 진피, 감초, 인삼, 백출]		TF-IDF 벡터	[숙지황, 당귀, 천궁]
	co-word 벡터	[반하, 진피, 후박]		co-word 벡터	[숙지황, 당귀, 천궁]
	t-score 벡터	[반하, 후박, 곽향, 백두구, 축사]		t-score 벡터	[숙지황, 두중, 우슬]
	word2vec 벡터	[반하, 반하국]		word2vec 벡터	[숙지황, 건지황]
	GloVe 벡터	[반하, 남성]		GloVe 벡터	[숙지황, 쇄양, 속단, 숙건지황, 석곡]
황금 (2052)	TF 벡터	[황금, 시호]	시호 (1403)	TF 벡터	[시호, 황금]
	TF-IDF 벡터	[황금, 시호]		TF-IDF 벡터	[시호, 황금]
	co-word 벡터	[황금, 시호, 감초, 인삼, 백출]		co-word 벡터	[시호, 감초, 인삼, 백출]
	t-score 벡터	[황금, 시호]		t-score 벡터	[시호, 황금]
	word2vec 벡터	[황금, 편금, 화피, 우방자, 악실]		word2vec 벡터	[시호, 북시호, 자소엽, 자소]
	GloVe 벡터	[황금, 치자]		GloVe 벡터	[시호, 승마]
백복령 (1990)	TF 벡터	[백복령, 인삼, 감초]	창출 (1390)	TF 벡터	[창출, 진피, 감초, 인삼]
	TF-IDF 벡터	[백복령, 백출, 인삼]		TF-IDF 벡터	[창출, 진피, 감초, 인삼, 백출]
	co-word 벡터	[백복령, 인삼, 백출]		co-word 벡터	[창출, 진피, 후박]
	t-score 벡터	[백복령, 인삼, 백출]		t-score 벡터	[창출, 후박, 곽향, 백두구, 축사]
	word2vec 벡터	[백복령, 복령]		word2vec 벡터	[창출, 백출, 익지인, 익지]
	GloVe 벡터	[백복령, 백작약]		GloVe 벡터	[창출, 진피, 청피, 축사]
방풍 (1943)	TF 벡터	[방풍, 강활]	대황 (1379)	TF 벡터	[대황, 망초]
	TF-IDF 벡터	[방풍, 강활]		TF-IDF 벡터	[대황, 망초]
	co-word 벡터	[방풍, 강활, 독활]		co-word 벡터	[대황, 망초]
	t-score 벡터	[방풍, 강활, 독활]		t-score 벡터	[대황, 망초]
	word2vec 벡터	[방풍, 감국화, 감국]		word2vec 벡터	[대황, 대청, 두시, 향시]
	GloVe 벡터	[방풍, 방기, 독활]		GloVe 벡터	[대황, 망초]

함께 즐겨 사용되는데, 이를 약대(藥對)라고 한다. 카운트 기반 방법에서 도출된 단어 임베딩 결과는 "함께 즐겨 사용되는 본초"를 인접한 본초로 도출함으로써 이러한 약대가 두드러지게 표현되었다.

카운트 기반 단어 임베딩 방법 간에는 비교적 유사한 결과가 도출되었다. 당귀, 인삼, 천궁, 대황을 시드 본초로 했을 때는 4가지 방식 모두 결과가 완전히 동일했다. 다만 TF 벡터의 단어 임베딩 결과에서 최빈도 본초인 감초로 수렴되는 경우가 상대적으로 많이 나타났다. 진피, 반하, 창출 등을 시드 본초로 하였을 경우 TF 벡터 방식으로 생성된 방제는 모두 감초로 수렴되는 결과를 보였다. TF 벡터의 경우 빈도가 높은 본초의 영향이 높게 평가되었다고 할 수 있다.

이러한 경향은 TF 벡터 방식이 가장 심했고, TF-IDF 벡터 방식이나 $co\text{-}word$ 벡터 방식에서도 어느 정도 나타났다. TF-IDF 벡터는 자주 나타나는 본초에 대해 가중치를 적게 주었기 때문에 이러한 효과가 상대적으로 억제될 수 있었고, $co\text{-}word$ 벡터의 경우에는 단어들 사이의 공기어 빈도를 측정하였기 때문에 빈도가 높은 본초가 여전히 과잉 대표되는 경향을 보인 것으로 풀이된다.(Table 7 참조)

$t\text{-}score$ 벡터의 경우에는 $co\text{-}word$ 벡터가 가지는 이러한 단점을 보완하여 기댓값과 관찰값의 차이를 사용하였기 때문에 빈도가 높은 본초로 결과가 수렴되는 경우가 가장 적게 나타났다. 그러나 그로 인해 특이한 결과가 도출되기도 하였다. 다른 방식에서는 백작약과 가장 인접한 본초로 당귀를, 건강과 가장 인접한 본초로 부자를, 황기와 가장 인접한 본초로 인삼을 도출한 데 반해, $t\text{-}score$ 벡터 방식에서는 각각 황기, 오수유, 숙지황을 도출하였다. 이러한 차이는 높은 빈도의 효과가 억제되어 나타난 결과로 풀이된다.

이러한 빈도에 대한 편향은 카운트 기반 단어 임베딩 방식에서 불가피한 것이라고 할 수 있다. 카운트 기반 방식 자체가 단어의 빈도를 기반으로 하고 있기 때문이다.

Table 7. Cases of Prescription Generation Result Converted to Licorice

	TF 벡터	TF-IDF 벡터	$co\text{-}word$ 벡터	$t\text{-}score$ 벡터
진피	[진피, **감초**, 인삼]	[진피, **감초**, 인삼, 백출]	[진피, 후박]	[진피, 후박, 곽향, 백두구, 축사]
반하	[반하, 진피, **감초**, 인삼]	[반하, 진피, **감초**, 인삼, 백출]	[반하, 진피, 후박]	[반하, 후박, 곽향, 백두구, 축사]
창출	[창출, 진피, **감초**, 인삼]	[창출, 진피, **감초**, 인삼, 백출]	[창출, 진피, 후박]	[창출, 후박, 곽향, 백두구, 축사]
황련	[황련, 황금, 시호]	[황련, 황금, 시호]	[황련, 황금, 시호, **감초**, 인삼, 백출]	[황련, 황금, 시호]
길경	[길경, **감초**, 인삼]	[길경, 지각]	[길경, **감초**, 인삼, 백출]	[길경, 전호]
황기	[황기, 인삼, **감초**]	[황기, 인삼, 백출]	[황기, 인삼, 백출]	[황기, 숙지황, 두충, 우슬]

카운트 기반 방법에서 함께 자주 사용되는 본초들이 도출되는 경향을 보인 데 반해 예측 기반 단어 임베딩은 함께 사용되는 본초가 아니라 유의어 관계에 있는 본초를 '인접한' 본초로 도출해 주었다. 특히 word2vec 벡터에서는 [감초, 자감초], [당귀, 당귀신], [진피, 귤홍], [반하, 반하국], [백복령, 복령], [백작약, 백작], [건강, 백강], [숙지황, 건지황] 등의 결과가 나타났다. 본초는 사용 부위나 수치법에 따라 이름을 달리하는 경우가 많다. 동일한 방제라고 하더라도 의서에 따라 다른 방식으로 표기되는 경우도 있다. word2vec 벡터를 이용한 단어 임베딩 방식은 이러한 본초들을 인접한 본초로 찾아주었다.

GloVe 벡터의 경우에도 감초나 길경을 시드 본초로 하였을 때 '건갈-갈근'을, 당귀나 건강을 시드 본초로 하였을 때 '관계-계피'를, 천궁을 시드 본초로 하였을 때는 '형개-형개수'를 연속하여 도출해 냈다. 이러한 결과는 GloVe 벡터 방식 역시 word2vec만큼 두드러지지는 않으나 어느 정도 유의어 관계에 있는 본초를 찾아 주었음을 알 수 있다. 또한 [백출, 복령], [시호, 승마] 등 다른 단어 임베딩 방식에서 도출되지 않았던 주요한 약대가 도출되는 모습도 확인할 수 있었다. 이러한 결과는 $t\text{-}score$ 벡터와 같이 빈도에 대한 효과가 크게 억제되었기 때문으로 풀이된다.

이상의 결과를 정리해 보겠다. 단어 임베딩 방법은 저마다 인접한 단어의 의미를 서로 다른 방식으로

표현해 주었다. 카운드 기반 방법에서는 단어 임베딩 결과로 나타난 벡터에서 함께 자주 사용되는 단어(본초)일수록 공간상에 벡터 값이 서로 가깝게 나타났다. 그에 반해 예측 기반 방법에서는 단어 임베딩 결과로 생성된 벡터에서 유의어 관계에 있는 단어(본초)일수록 서로 가깝게 나타났다. 또 특별한 가중치 없이 빈도 자체를 사용한 TF 벡터에서는 출현 빈도가 높은 단어가 많은 영향력을 보인데 반해, 나머지 방법들에서는 이러한 점이 보정되어 나타났다.

Ⅲ. 결론

동아시아 전통의학에서는 고문헌이라고 하는 텍스트로 이루어진 방대한 양의 데이터가 존재한다. 그간 이러한 텍스트는 의학을 익히는 이들이 반복해서 읽고 깨달아야 할 대상으로만 여겨져 왔다. 그러나 오늘날 인터넷을 통해 대량의 텍스트 데이터가 생겨나면서 비정형 데이터인 텍스트를 분석하여 인간에게 유용하게 사용하고자 하는 시도들이 늘어나고 있다. 따라서 충분한 양의 데이터가 준비된다면 전통의학의 문헌들도 이러한 방식을 적용하여 암묵적이었던 지식들을 명시적으로 표현하거나, 잘 드러나지 않았던 정보들을 찾아내는 데 활용할 수 있을 것이다.

텍스트를 분석하기 위해서는 텍스트를 컴퓨터가 이해할 수 있는 형태로 변환시켜주는 단어 임베딩이 선행되어야 한다. 텍스트의 단어들은 단어 임베딩 과정을 거쳐 실수(實數)로 이루어진 고차원 벡터로 매칭되게 된다. 본 연구에서는 자연어처리 분야에서 사용되고 있는 단어 임베딩 방법을 동아시아 전통의학 데이터에 적용해 보고 그 결과를 비교해 보았다. 향후 전통의학 텍스트를 분석할 때 분석 목적에 맞는 적절한 방법을 선택하는 데 도움을 주기 위해서이다.

본 연구에서는 동아시아 전통의학의 치법을 함축하고 있는 방제 데이터를 대상으로 하여, 카운트 기반 단어 임베딩 방식 4가지와 예측 기반 단어 임베딩 방식 2가지를 적용하여 각각의 단어 임베딩 결과를 도출하였다. 자연어처리 기법에서 문서와 단어는 대상 데이터에서 방제와 본초로 각각 병치시켜 적용하였다.

단어 임베딩 결과를 직관적으로 비교하기 위해 "방제 생성 게임"을 제안하였고, 그 결과를 가지고 단어 임베딩 결과를 비교해 보았다. 인접한 벡터를 추출하였을 때, 카운트 기반 단어 임베딩 방식은 자주 함께 사용되는 '인삼-백출', '당귀-천궁' 등 약대(藥對) 관계에 있는 본초들이 도출되었다. 이에 반해, 예측 기반 단어 임베딩 방식에서는 '감초-자감초', '복령-백복령' 등 유의어 관계의 본초들이 도출되었다.

한편, 단어 임베딩은 기본적으로 단어의 빈도를 출발점으로 삼기 때문에 빈도가 높은 단어가 두드러지게 나타날 수 있다. 본 연구 결과에서도 이러한 점을 확인할 수 있었다. TF 벡터의 경우 빈도가 가장 높은 감초의 영향이 높게 평가되는 경향을 보였고, TF-IDF 벡터 방식이나 $co\text{-}word$ 벡터 방식에서도 이러한 효과가 어느 정도 관찰되었다. $t\text{-}score$의 경우는 오히려 빈도의 영향을 낮게 평가하여 낯선 결과들이 도출되기도 하였다.

어떤 임베딩 방식이 우월하다고 단정하기보다는 분석하는 목적에 따라 단어 임베딩 방식을 선택해야 할 것으로 보인다. 예를 들어, 처방 데이터에서 특정 병증에 사용된 유의미한 약대(藥對)를 도출하는 등 함께 사용된 본초가 중요한 의미를 지니는 경우에는 카운트 기반 단어 임베딩 방식이 효과적일 것으로 보인다. 이 가운데 TF 벡터 방식[29]은 빈도의 효과가 과장되는 경향이 있으므로 TF-IDF 벡터[30]

29) 용어 임베딩에 TF 벡터를 사용한 예로는 다음의 연구가 있다.
김안나 외 5인. 불면 처방 활용 본초의 네트워크 분석. 대한한의학원전학회지. 2018. 31(4). pp.68-78.
오준호. 계층적 군집분석(hierarchical clustering)을 통한 침구자생경(鍼灸資生經) 경혈 선택 요인 분석. 대전대학교 한의학연구소 논문집. 2014. 23(1). pp.115-124.
두 본초에 해당하는 TF 벡터의 내적(inner product)을 구하면 두 본초가 함께 사용된 빈도가 도출된다. 따라서 본초 조합의 빈도를 관찰한 전자의 연구는 본초 각각을 TF 벡터로 상정한 것과 동일한 의미를 가진다. 후자는 TF 벡터를 그대로 사용하지 않고 벡터의 길이를 고르게 하기 위해 상대빈도를 사용하였다.

30) 용어 임베딩에 TF-IDF 벡터를 사용한 예로는 다음의 연

나 *co-word* 벡터가 좀 더 합리적인 선택이 될 수 있다.

반면, 이미 알려진 조합 외에 특수한 약재 조합을 탐색하거나, 동일한 처방에서 즐겨 사용되지 않았던 본초 사이의 관계를 검토해야 할 때에는 빈도의 영향이 적은 *t-score* 벡터와 같은 방법이 권장될 수 있을 것이다. 또한 동의어나 유의어 관계에 있는 본초를 추출해 내거나 주변 본초와의 맥락 속에서 약효가 유사한 본초를 추정하기 위한 목적일 경우에는 예측 기반 임베딩 방식이 효과적일 것으로 보인다.

본 연구는 한의학 고문헌에서 방제 데이터만을 사용하였고, 빈도가 높은 본초에 대한 결과만을 비교하였다. 이를 통해 단어 임베딩이 가지는 특성을 어느 정도 비교할 수 있었으나 이를 한의학 고문헌 전체로 일반화하기에는 부족함이 있다. 이에 대해서는 후속 연구를 기대한다.

감사의 말씀

본 연구는 한국한의학연구원 주요사업 “한의 고문헌 지식 분석 시스템 개발(KSN1812200)”의 지원을 받아 수행되었습니다.

References

1. Kang BM. Language, computer, corpus linguistics (revised edition). Seoul. Korea University Press. 2011. pp.122-123.
강범모. 언어, 컴퓨터, 코퍼스 언어학(개정판). 서울. 고려대학교출판부. 2011. pp.122-123.
2. Peng W. Dictionary of Chinese medicine prescription (1st edition). Beijing. People's Medical Publishing House. 2005. pp.3-4.
3. Stefan Bordag. A Comparison of Co-occurrence and Similarity Measures as Simulations of Context. Alexander Gelbukh ed.. Computational Linguistics and Intelligent Text Processing. New York. Springer. 2008. pp.52-63.
4. Bae HJ et al. Investigation of the Possibility of Research on Medical Classics Applying Text Mining. The Journal of Korean Medical Classics. 2018. 31(4). pp.27-46.
배효진 외 4인. 텍스트마이닝을 활용한 한의학 원전 연구의 가능성 모색. 대한한의학원전학회지. 2018. 31(4). pp.27-46.
5. Bang MW, Kim KW, Lee BW. A Study on the Inference and Classification Method of the Effectiveness Using the Herb Composition. Herbal formula science. 2017. 25(1). pp.29-38.
방민우, 김기욱, 이병욱. 구성을 이용한 방제의 효능 추론 및 분류 방법에 관한 연구. 대한한의학방제학회지. 2017. 25(1). pp.29-38.
6. Hwang HS et al. Word Embedding using Relative Position Information between Words. Journal of KIISE. 2018. 45(9). pp.943-949.
황현선 외 3인. 단어 간의 상대적 위치정보를 이용한 단어 임베딩. 정보과학회논문지. 2018. 45(9). pp.943-949.
7. Kang HS, Yang JH. The Analogy Test Set Suitable to Evaluate Word Embedding Models for Korean. Journal of Digital Contents Society. 2018. 19(10). pp.1999-2008.
강형석, 양장훈. 한국어 단어 임베딩 모델의 평가에 적합한 유추 검사 세트. 디지털콘텐츠학회논문지. 2018. 19(10). pp.1999-2008.

구가 있다.
배효진 외 4명. 텍스트마이닝을 활용한 한의학 원전 연구의 가능성 모색. 대한한의학원전학회지. 2018. 31(4). pp.27-46.
오준호. HF-IFF: TF-IDF를 응용한 병증-본초 연관성(relevancy) 측정과 본초 특성의 시각화 -청강의감 방제를 대상으로-. 대한본초학회. 2015. 30(3). pp.63-68.

8. Kim AN et al. Network Analysis on Herbal Combinations in Korean Medicine for Insomnia. The Journal Of Korean Medical Classics. 2018. 31(4). pp.68-78.
김안나 외 5인. 불면 처방 활용 본초의 네트워크 분석. 대한한의학원전학회지. 2018. 31(4). pp.68-78.
9. Kim KW, Kim TY, Lee BW. Automatic Extraction Method of Compositional Herb Using Herb List. The Journal of Korean Medical Classics. 2014. 27(3). pp.155-166.
김기욱, 김태열, 이병욱. 본초 목록을 이용한 방제의 본초 구성 자동 추출 방법. 대한한의학원전학회지. 2014. 27(3). pp.155-166.
10. Kim KW, Kim TY, Lee BW. Analysis of Prescriptions from Taepyeonghyeminhwajegukbang, Somunsunmyungronbang and Nansilbijang based on Herb weight ratio grade. The Journal of Korean Medical Classics. 2014. 27(4). pp.73-84.
김기욱, 김태열, 이병욱. 본초 비율의 순위를 이용한 문헌의 특징 분석 방법 -태평혜민화제국방(太平惠民和劑局方), 난실비장(蘭室秘藏), 소문선명론방(素問宣明論方)을 중심으로-. 대한한의학원전학회지. 2014. 27(4). pp.73-84.
11. Kim WJ, Kim DH, Jang HW. Semantic Extention Search for Documents Using the Word2vec. Journal of the Korea Contents Association. 2016. 16(10). pp.687-692.
김우주, 김동희, 장희원. Word2vec을 활용한 문서의 의미 확장 검색방법. 한국콘텐츠학회논문지. 2016. 16(10). pp.687-692.
12. M Sahlgren. The distributional hypothesis. Italian Journal of Linguistics. 2008. 20. pp.33-53.
13. Oh JH. Deduction of Acupoints Selecting Elements on Zhenjiuzishengjing using hierarchical clustering. Journal of DaeJeon University KM institute. 2014. 23(1). pp.115-124.
오준호. 계층적 군집분석(hierarchical clustering)을 통한 침구자생경(鍼灸資生經) 경혈 선택 요인 분석. 대전대학교 한의학연구소 논문집. 2014. 23(1). pp.115-124.
14. Oh JH. HF-IFF: Applying TF-IDF to Measure Symptom-Medicinal Herb Relevancy and Visualize Medicinal Herb Characteristics -Studying Formulations in Cheongkangeuigam-. The Korea Association of Herbology. 2015. 30(3). pp.63-68.
오준호. HF-IFF: TF-IDF를 응용한 병증-본초 연관성(relevancy) 측정과 본초 특성의 시각화 -청강의감 방제를 대상으로-. 대한본초학회. 2015. 30(3). pp.63-68.
15. Oh JH. Measure of the Associations of Accupoints and Pathologies Documented in the Classical Acupuncture Literature. Korean Journal of Acupuncture. 2016. 33. pp.18-32.
오준호. 고의서에 나타난 경혈과 병증의 연관성 측정 및 시각화 - 침구자생경 분석 예를 중심으로 -. 경락경혈학회지. 2016. 33. pp.18-32.
16. Oh, JH. Can Similarities in Medical thought be Quantified. The Journal of Korean Medical Classics. 2018. 31(2). pp.71-82.
오준호. 의학 사상의 유사성은 계량 분석 될 수 있는가. 대한한의학원전학회지. 2018. 31(2). pp.71-82.
17. Park DS, Kim HJ. A Proposal of Join Vector for Semantic Factor Reflection in TF-IDF Based Keyword Extraction. The Journal of Korean Institute of Information

Technology. 2018. 16(2). pp.1-16.
박대서, 김화종. TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합벡터 제안. 한국정보기술학회논문지. 2018. 16(2). pp.1-16.

18. Park IS et al. Characterization of Five Shu Acupoint Pattern in Saam Acupuncture Using Text Mininig. Korean J Acupunct. 2015. 32. pp.66-74.
박인수 외 5인. 텍스트마이닝을 통한 사암침법 오수혈 사용 패턴 분석. 경락경혈학회지. 2015. 32. pp.66-74.
19. Song YS et al. A study of relationship between excrement and materia medica in Bangyakhappyeon based on the data mining analysis. 2012. 16(2). Journal of Korean Institute of Oriental Medical Diagnostics. 2012. 16(2). pp.33-45.
송영섭 외 3인. 데이터 마이닝을 이용한 대변과 약물간의 연관성 분석 -방약합편을 중심으로-. 대한한의진단학회. 2012. 16(2). pp.33-45.
20. Wu YH et al. Feature Comparison by Prescription Configuration Analysis among Liuhejian`s and Lidongyuan`s Books and 『Hejijufang』. The Journal of Korean Medical Classics. 2015. 28(1). pp.55-69.
오월환 외 3인. 방제구성을 이용한 유하간(劉河間) 및 이동원(李東垣)의 저작과 『화제국방(和劑局方)』의 특성 비교. 대한한의학원전학회지. 2015. 28(1). pp.55-69.
21. Chinese Medical Database. Beijing. Hunan Electronic Audio and Video Publishing House. 2003.
中华医典. 北京. 湖南电子音像出版社. 2003.
22. Korea Institute of Oriental Medicine. Mediclassics. [cited on Jan 12, 2019]. Avaiable from: https://mediclassics.kr
한국한의학연구원. 한의학고전DB. [cited on Jan 12, 2019]. Avaiable from: https://mediclassics.kr
23. Korea Intellectual Property Office. Korean Traditonal Knowledge Portal. [cited on Jan 12, 2019]. Avaiable from: http://www.koreantk.com
특허청. 한국전통지식포털. [cited on Jan 12, 2019]. Avaiable from: http://www.koreantk.com
24. Mikolov, T et al. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781. 2013. [cited on Jan 12, 2019]. Avaiable from: https://arxiv.org/abs/1301.3781
25. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014. [cited on Jan 12, 2019]. Avaiable from: https://www.aclweb.org/anthology/D14-1162
26. Wikipedia. Data analysis. [cited on Jan 12, 2019]. Avaiable from: https://en.wikipedia.org/wiki/Data_analysis
27. Wikipedia. Natural language processing. [cited on Jan 12, 2019]. Avaiable from: https://en.wikipedia.org/wiki/Natural_language_processing
28. Wikipedia. Word embedding. [cited on Jan 12, 2019]. Avaiable from: https://en.wikipedia.org/wiki/Word_embedding
29. Wordnet. [cited on Jan 12, 2019]. Avaiable from: https://wordnet.princeton.edu/documentation/20-wnstats7wn