대한한의학원전학회지
2022년 2월 35권 1호

pISSN 1229-8328
J Korean Med Classics 2022:35(1) 093-102
https://doi.org/10.14369/jkmc.2022.35.1.093

JKMC
Journal of Korean Medical Classics

# 기계 학습을 이용한 한의학 용어 유의어 사전 구축 방안

한국한의학연구원 책임연구원
**오준호***

# A Strategy for Constructing the Thesaurus of Traditional East Asian Medicine (TEAM) Terms With Machine Learning

**Oh Junho***
Senior Researcher. Korea Institute of Oriental Medicine.

**Objectives** : We propose a method for constructing a thesaurus of Traditional East Asian Medicine terminology using machine learning.

**Methods** : We presented a method of combining the 'Automatic Step' which uses machine learning and the 'Manual Step' which is the operator's review process. By applying this method to the sample data, we constructed a simple thesaurus and examined the results.

**Results** : Out of the 17,874 sample data, a thesaurus was constructed targeting 749 terminologies. 200 candidate groups were derived in the automatic step, from which 79 synonym groups were derived in the manual step.

**Conclusions** : The proposed method in this study will likely save resources required in constructing a thesaurus.

**Key words** : synonym, thesaurus, machine learning, Korean Medicine, Traditional East Asian Medicine (TEAM).

* Corresponding Author : Oh Junho.
Senior Researcher. Korea Institute of Oriental Medicine. 1672 Yuseong-daero, Yuseong-gu, Daejeon, 34054.
Tel +82-42-868-9317, E-mail: junho@kiom.re.kr
저자들은 본 논문의 내용과 관련하여 그 어떠한 이해상충도 없습니다.
Received(January 28, 2022), Revised(February 7, 2022), Accepted(February 7, 2022)

# I. 서론

유의어(Synonym)는 특정 단어와 의미가 같거나 비슷한 단어를 가리킨다. 유의어 사전(Thesaurus)은 단어에 대한 유의어를 모아 놓은 사전이다. 유의어 사전은 유의어 간의 의미 비교를 통하여 단어의 뜻을 깊이 이해할 수 있도록 도와준다. 따라서 정확한 의미를 전달하기 위해 적확한 단어를 선택해야 하는 경우, 또는 글을 퇴고하며 같은 단어의 반복을 피하고 다양한 어휘를 사용하고자 하는 경우 등 여러 가지 목적으로 사용된다.

유의어 사전은 언어학적 측면에서뿐만 아니라 디지털 자료를 더 잘 활용하기 위해서도 필요하다. 컴퓨터가 발달하면서 많은 정보가 전자 텍스트(digital text)로 변형되었고, 사용자는 검색 키워드를 통해 방대한 자료 속에서 원하는 정보를 쉽고 빠르게 검색할 수 있게 되었다. 그러나 유의어를 통해 키워드를 확장하여 검색하지 않으면 원하는 검색 결과가 누락되기 쉽다. 이런 이유로 오늘날 많은 검색 엔진들은 내부적으로 유의어 사전을 통해 사용자가 입력한 검색어뿐만 아니라 그 유의어까지 검색해 주고 있다.

또한 디지털 데이터가 급격히 증가하면서 여러 분야에서 기계 학습(machine learning)을 이용하여 인간 생활을 편리하게 하려는 노력들이 경주되고 있다. 인간의 자연 언어로 만들어진 디지털 데이터를 처리하고 분석하기 위해서는 자연어처리(natural language processing) 또는 텍스트 분석(text analysis) 기법이 필요하다. 이 기법에는 공통적으로 주어진 디지털 텍스트를 최소 의미 단위인 토큰(token)으로 구분하는 작업이 선행되어야 한다. 이 때 유의어 사전을 바탕으로 의미가 유사하거나 동일한 토큰을 묶어주면 분석 결과가 더욱 선명해진다. 이는 데이터의 양이 적을수록 더욱 크게 영향을 받는다. 이처럼 유의어 사전은 디지털 텍스트의 검색과 분석을 위해서도 필요하다.

한의학 분야에서도 디지털 텍스트가 증가하여 원하는 정보를 더 쉽게 찾아낼 수 있게 되었다. 이러한 변화에 따라 새로운 지견을 얻기 위해 한의학 텍스트 분석에 대한 수요도 높아지고 있다. 그러나 동아시아 전통 사회에서 잉태된 한의학은 현대와 다른 사상적 기반과 언어적 토양 위에서 형성되었다. 따라서 한의학 분야의 디지털 텍스트를 더욱 잘 활용하기 위해서는 한의학 용어만을 위한 유의어 사전의 구축이 불가피하다. 이러한 문제의식 아래 한의학 용어 관리[1], 한의학 시소러스 구축[2]에 대한 연구가 이루어졌고, 모든 용어를 다양한 관계에 따라 연결하는 온톨로지 연구[3]까지 수행된 바 있다.

그런데도 현재까지 데이터를 검색하거나 분석할 때 손쉽게 활용할 수 있는 공개된 한의학 용어집, 특히 유의어 사전은 존재하지 않는다. 설사 기성 유의어 사전이 존재한다고 하더라도 자신의 데이터에 적용하려면 최적화에 적지 않은 노력을 투입해야 한다. 한자(漢字)라는 특성 때문에 발생하는 인코딩(encoding) 문제, 다중코드자 문제, 이체자 문제, 입력 오류의 문제 등 데이터가 형성되는 단계마다 여러 가지 문제들이 존재하는데 기성 용어집은 이러한 문제들을 모두 고려할 수 없기 때문이다.

유의어 사전은 구축 자체도 어렵다. 유의어 사전 구축을 위해서는 데이터에서 용어를 추출하고, 추출된 용어를 작업자가 관계 짓는 작업이 수반된다. 그러나 작업자에 따라 유의어 관계의 판정이 달라질 수 있고, 데이터가 늘어날수록 수작업의 효율도 극히 낮아진다. 그 필요성에도 불구하고 유의어 사전이 쉽게 구축되지 못하는 이유이다.

이에 이 글에서는 주어진 데이터에서 비지도학습(unsupervised learning)을 통해 유의어 사전을 만드는 방안을 제안하고자 한다. 이 글의 목적은 방법의 우수성을 밝히는 데 있지 않고 스몰 데이터(small data) 단위에서 유의어 사전을 구축할 수 있는 실현

1) 이병욱, 심범상, 엄동명. 한의학 용어관리 시스템을 결합한 고전원문 제공 서비스에 관한 연구. 대한한의학원전학회지. 2009. 22(4). pp.167-176.
차승준, 외 6인. 한의학 용어 수집 및 관리 시스템 구축. 대한예방한의학회지. 2010. 14(1). pp.59-76.
김혜은 외 4인. 한의학 증상용어의 형태소 분석을 위한 자연어 표기 분석. 대한예방한의학회지. 2013. 17(2). pp.179-187.

2) 백유상. 한의학정보 검색엔진 개발을 위한 시소러스 연구. 대한한의학원전학회지. 2006. 19(1). pp.155-167.

3) 장현철 외 18명. 온톨로지 기반 한의학 지능형 정보체계 연구. 대전. 한국한의학연구원. 2013.

가능하면서 효율적인 방법을 제안하는 데 있다. 따라서 제안하고자 하는 전략을 설명하고 이를 예시 데이터에 적용하여 그 결과를 살펴볼 것이다.

# Ⅱ. 본론

## 1. 유의어 사전 구축 전략

이 글에서 제안하는 유의어 사전의 구축 전략은 컴퓨터를 이용한 자동화와 사람의 수작업이 합쳐진 하이브리드 방식이다. 먼저 컴퓨터를 이용해 데이터 자체를 학습하여 유의어 사전의 초안을 만든 다음, 이를 얼개 삼아 작업자가 검토하고 첨삭하는 방식이다. 전자를 '자동 단계(Automatic Step)', 후자를 '수동 단계(Manual Step)'라고 부르도록 하겠다. 자동 단계는 주어진 데이터를 바탕으로 비지도 학습을 수행하여 유사한 용어끼리 짝지어주는 단계이다. 수동 단계는 이렇게 짝지어진 용어를 작업자가 검토하여 오류를 바로잡고 수정하는 단계이다.

자동 단계만으로 유의어 사전이 구축된다면 좋겠지만 학습 데이터가 충분하지 않은 스몰 데이터의 경우, 자동 단계의 정확도가 그다지 높지 않기 때문에 현 단계에서 작업자의 검토 과정을 완전히 배제하기는 어렵다. 다만 자동 단계에서 만들어진 유의어 사전의 초안이 정확할수록 수동 단계의 처리가 단순해지므로 이 단계에서 더 좋은 결과를 도출하기 위한 고민이 필요하다.

수동 단계는 작업자가 자동 단계 결과를 검토하는 단계로서 이 글에서 특별히 제시할만한 진보된 방법은 없다. 따라서 자동 단계에 관해 설명을 집중하고, 수동 단계에 대해서는 간략히 서술하고자 한다. 설명을 위해 이 글에서 제시한 전략에 따라 예시 데이터를 대상으로 유의어 사전을 구축하고 그 결과도 함께 살펴볼 것이다(이하 '예시 분석').

## 2. 자동 단계(Automatic Step)

자동 단계는 기계 학습 가운데 비지도 학습을 이용하며 구체적으로 다음의 단계에 따라 수행된다.

데이터 준비
- 학습할 데이터 확보 (디지털 형태의 데이터)
- 용어(word)로 사용될 토큰(token) 추출

용어 임베딩(embedding)
- 용어(토큰)를 고차원 벡터로 임베딩
- 임베딩 벡터를 통해 용어 사이의 의미 거리 측정

군집 구성(clustering)
- 측정된 거리를 바탕으로 용어를 군집(cluster)으로 묶음
- 군집화 결과를 유의어 후보로 출력

작업의 반복
- 용어 임베딩 및 군집 구성을 반복
- 반복된 결과에서 공통적으로 나타나는 결과를 최종 결과로 수용

### 2.1. 데이터 준비

본 연구에서는 웹 스크래핑 기법(web scraping) 기법을 이용하여 <한국전통지식포탈>에 공개된 '전통의학처방' 데이터를 수집하고[4], 이 가운데 구성 약재 부분만을 추출하여 기계 학습에 이용하였다(이하 '예시 데이터'). 이 데이터는 한국 특허청이 한국 전통 의학 지식을 소개하면서 여러 의서에 실려 있는 처방 정보를 디지털 형태로 정리하여 공개한 데이터이다. 본 연구에서 이 데이터를 선택한 이유는 연구와 설명의 편의를 위한 것으로, 웹을 통해 누구나 접근할 수 있고 데이터의 규모가 지나치게 작지 않으며 별도의 토큰 추출이 필요하지 않다는 장점 때문이다.

이렇게 수집된 예시 데이터는 모두 20,120건이었다. 유의어 사전 구축 결과를 쉽게 나타내기 위해 처방을 이루고 있는 본초 구성에 주목하였다. 이에 다시 데이터 파싱(parsing) 과정을 통해 예시 데이터에 존재하는 본초 구성만을 추출하였다. 이 가운데 1종의 본초로만 이루어진 처방은 본초 사이의 관계를 파악할 수 없기 때문에 제외하였다. 이렇게 17,874건의 처방에 대한 본초 구성 데이터를 준비

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

할 수 있었다.

다음 과정은 주어진 데이터에서 분석의 최소 단위가 되어 줄 토큰을 추출하는 일이다. 예시 데이터의 경우 데이터 구축 단계에서 이미 본초와 본초를 구분해 놓았기 때문에 토큰 추출 작업은 따로 필요 없었다.[5] 예시 데이터에서 토큰, 즉 본초는 모두 1,826종이었다(이하 예시 분석 설명에서 '용어', '토큰', 그리고 '본초'는 같은 의미를 가진다). 사용 빈도가 낮은 토큰은 용어 임베딩이 어려우므로 제외하고, 등장 횟수가 10회 이상인 본초 749종만을 용어 임베딩 과정에 사용하였다. 사용 빈도가 낮은 토큰은 모두 1,077종으로 적지 않았다. 이들 토큰을 적절하게 처리할 수 없다는 점이 이 글에서 제시한 전략의 한 가지 맹점이다.

### 2.2. 용어 임베딩(embedding)

다음으로 컴퓨터를 통해 계산을 이어가기 위해 추출한 토큰을 숫자로 표현하는 과정, 즉 용어 임베딩이 뒤따른다. 용어 임베딩은 용어의 의미를 고차원 공간 위의 벡터로 표상화시키는 방법으로서 여러 가지 방법이 존재하므로 목적에 따라 합당한 방법을 선택하여야 한다. 이 글에서는 의미가 유사한 용어 사이의 관계를 파악할 목적이므로 word2vec을 사용하였다.[6]

word2vec은 중심 단어와 주변 단어의 관계를 통해 단어를 고차원 벡터에 임베딩 시키는 기법이다. '중심 단어'로부터 주변에 등장하는 '주변 단어'를 예측하는 방식(Skip-gram)이나, 반대로 '주변 단어'로부터 '중심 단어'를 예측하는 방식(Continuous Bag of Words Model ; CBOW)이 가능하다. 본 연구에서 사용한 후자의 방법은 주변 단어가 주어졌을 때 중심 단어가 나타날 조건부 확률을 계산하고 문서를 따라가면서 조건부 확률이 최대가 되도록 단어 임베딩 벡터를 업데이트시켜 나간다. 이렇게 하면 유사한 문맥에 등장하는 단어들이 인접하여 나타나게 된다.[7]

용어 임베딩에서 벡터의 크기는 중요한 초매개변수(hyper-parameter)이다. 그러나 여기에 대한 확고한 기준은 없는 형편이다. 100~300차원 내외에서 연구자가 상황에 맞게 선택하는 것이 일반적이다. 예시 분석에서는 용어 종류의 50%에 해당하는 374를 임베딩 벡터의 차원 수로 선정하였다. 이는 되도록 차원을 늘려 용어 사이의 차이를 드러내기 위함이다. 이렇게 word2vec 기법을 사용해 용어 749종에 대하여 374차원의 벡터를 도출할 수 있었다.[8]

임베딩 벡터는 고차원 공간의 좌표를 의미하므로 이를 통해 벡터 사이의 거리를 측정할 수 있다. 이렇게 측정된 거리는 해당 용어의 의미 차이라고 해석할 수 있다. 본 연구에서는 예시 데이터 본초 749종에서 임베딩 벡터를 도출한 뒤 이를 한 쌍씩 묶어 서로에 대한 코사인 거리(cosine distance)를 측정하여 각 용어 사이의 의미 거리를 측정하였다. 코사인 거리는 최소값이 0, 최대값이 1로 나타나며, 본 예시에서는 0에 가까울수록 용어의 의미가 유사하다고 해석할 수 있다. 이렇게 측정된 용어 사이의 코사인 거리 결과를 히스토그램으로 나타내면 <그림 1>과 같다.

용어의 임베딩과 코사인 거리 측정으로부터 중간 결과를 검토해 볼 수 있다. 예시 데이터에서 '육계(肉桂)', '진피(陳皮)', '치자(梔子)', '봉출(蓬朮)'과 가장 인접한 용어(인접어) 6가지를 2차원 평면에 표시해 보면 <그림 2>와 같다. 육계의 경우, 가장 가까이 위치한 용어는 계심(桂心), 관계(官桂), 계피(桂皮), 계지(桂枝) 순이었다. 4가지 용어 모두 한의학 지식에 부합하는 결과가 도출된 것을 확인할 수 있다.

5) 데이터가 자연어로 구성된 텍스트라면 토큰 추출 자체가 커다란 난관일 수 있다. 자연어에서의 토큰 추출은 하나의 연구 주제로서 도전적인 작업이다. 이는 이 글의 주제에서 벗어난 내용이므로 다루지 않는다. 한의학 텍스트에서 토큰 추출에 대해서는 다음의 연구가 있다. 오준호. 한의학 고문헌 텍스트 분석을 위한 비지도학습 기반 단어 추출 방법 비교. 대한한의학원전학회지. 2019. 32(3). pp.47-57.

6) 오준호. 한의학 고문헌 데이터 분석을 위한 단어 임베딩 기법 비교 : 자연어처리 방법을 적용하여. 대한한의학원전학회지. 2019. 32(1). pp.61-74.

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

8) word2vec 임베딩에는 프로그래밍 언어 R(ver 4.0.3)의 word2vec(ver 0.3.3) library를 사용하였다.

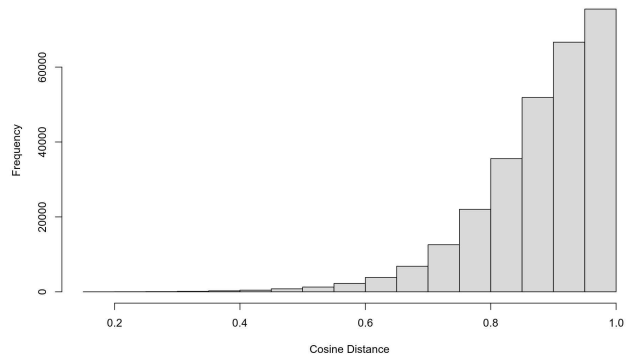


그림 1 예시 분석에서 용어 사이의 코사인 거리 히스토그램

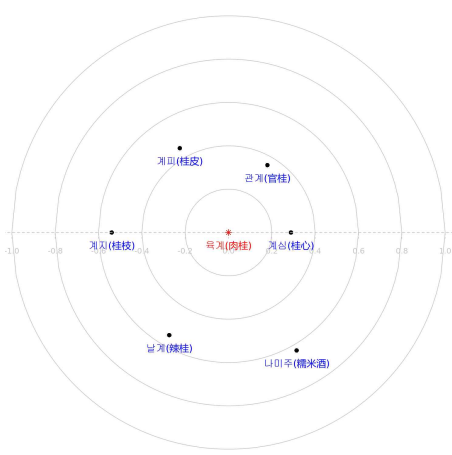


그림 2A '육계(肉桂)'의 인접어

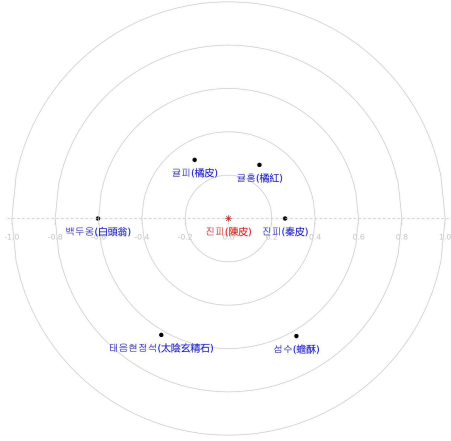


그림 2B '진피(陳皮)'의 인접어

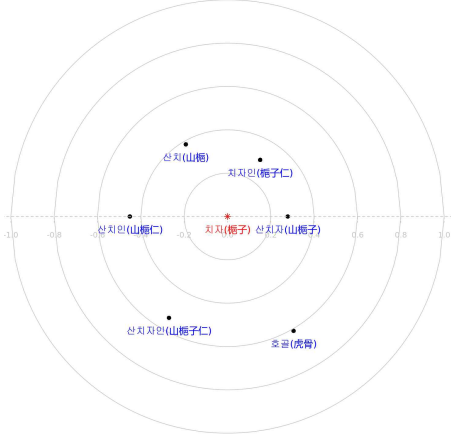


그림 2C '치자(梔子)'의 인접어

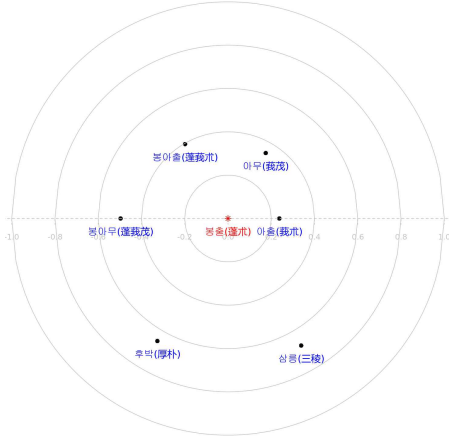


그림 2D '봉출(蓬朮)'의 인접어

### 2.3. 군집 구성(clustering)

모든 용어(또는 토큰) 사이의 거리가 측정되면 거리를 기준으로 가까이 있는 용어들끼리 짝을 지어줄 수 있다. 이렇게 짝지어진 군집은 비슷한 의미로 사용된 용어끼리 모인 것이라고 기대할 수 있다.

비지도 학습을 통해 데이터를 무리 짓는 방법을 군집 분석(Clustering analysis)이라고 하며, 대표적으로 K-평균 군집화(K-means Clustering)와 계층적 군집화(Hierarchical Clustering) 방법이 있다. 이 글에서는 거리가 가까운 용어들끼리 묶어나가기 위해 후자의 방법을 사용하였다. 계층적 군집화 방법은 거리가 가까운 군집끼리 합치면서 군집을 만들어 가는 방식을 말한다. 예시 데이터에서 계층적 군집화 수행 결과는 <그림 3>과 같다.

계층적 군집화에서는 무엇을 기준으로 군집(cluster)을 구분할 것인가, 즉 군집을 나눌 초매개변수를 어떤 값으로 할 것인가를 정해야 한다. 예시 분석에서는 코사인 거리를 기준으로 하였으므로 0에서 1 사이에서 초매개변수를 설정할 수 있다. 이에 대한 분명한 기준은 존재하지 않는다. 예시 분석의 경우, 1에 가깝게 설정하면 군집의 개수가 적어지고 0에 가깝게 설정할수록 군집의 개수가 많아진다. 지나치게 1에 가깝게 설정하면 의미 차이가 큰 용어가 같은 군집에 모이게 되고, 지나치게 0에 가깝게 설정하면 의미가 유사해도 다른 군집으로 나뉠 수 있다. 현재 이에 대한 분명한 기준을 제시할 수 없으므로 실제로 분석을 수행할 때는 기준을 바꾸고 결과를 확인해 가면서 원하는 결과가 도출되는지 검토하는 과정을 거쳐야 할 것이다.

예시 분석의 경우, 앞의 코사인 거리 정보를 이용하여 전체의 1%에 해당하는 0.541을 기준으로 삼았다. 이렇게 거리가 가까운 순서를 기준으로 상위 1% 이상에 해당하는 용어쌍을 군집으로 묶었다. 이처럼 계층적 군집화를 수행한 결과 431종의 군집을 도출할 수 있었다. 그러나 이 가운데 231종은 군집 구성 용어가 1가지로 유의어에 대한 정보를 담고 있지 않았다. 따라서 나머지 200종의 군집을 최종 결과로 도출할 수 있었다.

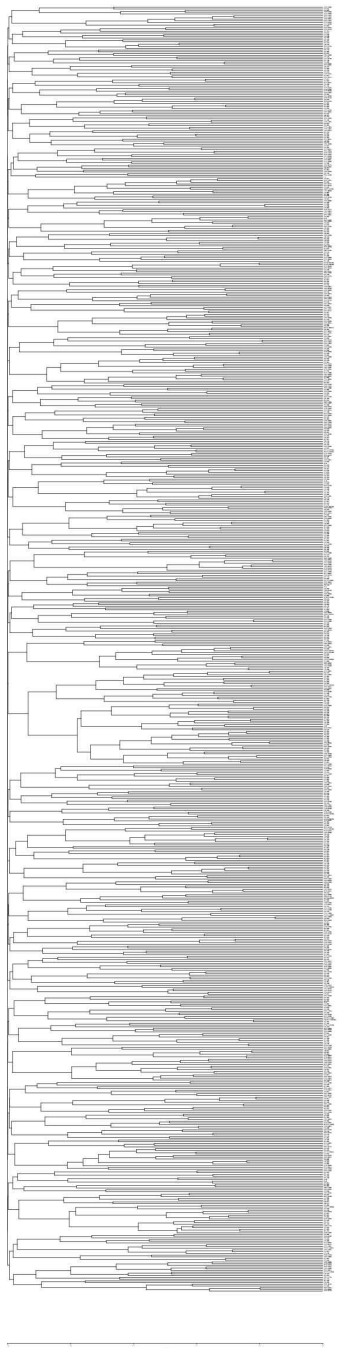

그림 3 예시 데이터 군집 결과

### 2.4. 작업의 반복

작업자의 판단이 중요한 작업에서는 독립된 몇 명의 작업자가 따로 작업을 수행하고 결과를 비교하여 결과의 정확도를 높이곤 한다. 이와 유사하게 자동 단계에서도 같은 과정을 독립적으로 반복하고 그 결과를 비교함으로써 결과를 더 정교하게 만들 수 있다. word2vec 임베딩은 임의의 초기 벡터로부터 학습을 시작하기 때문에 시행할 때마다 임베딩 결과가 조금씩 달라진다. 만약 여러 차례 수행한 결과에서 동일하게 도출된 용어쌍이 있다면 이는 좀 더 신뢰할 수 있는 결과일 것이다. 따라서 본 연구에서는 앞서 설명한 용어 임베딩과 계층적 군집 구성 단계를 몇 세트 반복하는 단계를 추가하였다.

예시 데이터를 대상으로 용어 임베딩과 군집 구성을 모두 9세트 반복하고, 이 가운데 과반수인 5회 이상 군집으로 묶인 용어쌍을 최종 결과로 수용하였다. 9세트를 반복하여 모두 1,852종의 군집을 얻을 수 있었는데 이 가운데 중복을 제외한 고유한 결과는 628종이었다. 이 고유한 군집 가운데 5회 이상 등장한 군집은 모두 147종이었다. 이를 최종 결과로 수용하였다.

최종 결과로 도출된 군집에 포함된 용어 수는 모두 340종이었다. 분석 대상 전체 용어 수가 749종이었으므로 최종 군집에 포함되지 않은 용어는 모두 409종이었다.

## 3. 수동 단계(Manual Step)

수동 단계는 자동 단계에서 획득한 결과를 작업자가 검토하고 수정하는 단계이다. 작업자의 정성적인 판단을 통해 수행되는 과정이기 때문에 구체적인 방법에 대해서 논하기는 어렵다. 대신 예시 분석에서 도출된 결과를 검토해 보는 것으로 설명을 대신하고자 한다.

자동 단계에서 도출된 용어 조합을 수동 단계로 분류하여 <표 1>와 같은 결과에 도달할 수 있었다. 자동 단계의 결과를 수동 단계에서 〔유형A〕, 〔유형B〕, 〔유형C〕로 나눌 수 있었는데, 〔유형A〕는 기대한 대로 유의어 관계로 짝지어진 조합, 〔유형B〕는 유의어가 아닌 용어가 일부 섞여 있는 조합, 〔유형C〕는 유의어가 관찰되지 않는 조합이다.

〔유형A1〕은 간단한 검토를 거쳐 유의어 사전에 그대로 포함시킬 수 있는 용어 조합이다. 자동 단계에서 이 유형이 많이 도출될수록 수동 단계의 작업은 용이해진다. 작업자가 최종적으로 판단해야 할 점이 있다면 조합 사이에 병합이 필요한지 여부이다. 예를 들어 【견우(牽牛), 견우자(牽牛子)】, 【백견우(白牽牛), 백축(白丑)】, 【흑견우(黑牽牛), 흑견우자(黑牽牛子), 흑축(黑丑)】 등의 집합이 있을 때 의미가 같다고 보아 이들을 합쳐 【견우(牽牛), 견우자(牽牛子), 백견우(白牽牛), 백축(白丑), 흑견우(黑牽牛), 흑견우자(黑牽牛子), 흑축(黑丑)】 집합을 생성할 것인지, 아니면 의미 차이를 인정하여 그대로 둘 것인지의 판단이 필요하다.

흥미로운 것은 〔유형A2〕이다. 자동 단계의 또 다른 이점을 보여주는 예로서, 이를 통해 데이터 자체의 오류를 파악할 수 있다. 예시 분석 결과를 보면, 데이터가 만들어질 때 '마황(麻黃)'을 '마황(馬蟥)'으로, '망초(芒硝)'를 '망소(芒消)'로, '백지(白芷)'를 '백지(柏脂)'로, '호초(胡椒)'를 '호초(好醋)'로 일부 잘못 입력하였을 가능성이 크다는 사실을 알 수 있다. 특히 '마황(馬蟥)'이나 '망소(芒消)'는 육안으로도 오류를 포착할 수 있으나 '백지(柏脂)'나 '호초(好醋)'의 경우에는 일반 본초로 인식되기 때문에 자동 단계 없이 사람의 수작업만으로는 오류를 포착하기 어렵다.

〔유형B〕는 유의어가 아닌 용어가 일부 섞여 있는 경우이다. 이러한 유형이 수작업 단계에서 중점적으로 검토해야 할 대상이다. 자동 단계에서 초매개변수를 조정하는 등 이 유형을 줄여나가기 위한 노력이 필요하다.

〔유형C〕는 작업자가 기대한 유의어 관계가 아닌 결과들이다. word2vec의 임베딩은 주변 용어를 통해 해당 용어의 의미를 추측하는 방법으로, 주변 용어의 분포가 유사하다면 임베딩 결과가 비슷해지고 용어 사이의 거리도 짧아진다. 따라서 여기에 섞여 있는 용어들은 작업자가 기대한 유의어는 아닐 수 있으나 기계 학습의 관점에서 보았을 때는 의미가 유사하다고 본 용어 조합이다. 〔유형C1〕의 경

**표 1 예시 데이터의 군집 결과**

| 유형A1 (유의어 관계) : 79종 | | | |
|---|---|---|---|
| 가자(訶子) , 가자육(訶子肉) , 가자피(訶子皮) | 담죽엽(淡竹葉) , 죽엽(竹葉) | 사인(砂仁) , 축사(縮砂) | 원삼(元蔘) , 현삼(玄蔘) |
| 갈근(葛根) , 건갈(乾葛) | 당귀(當歸) , 당귀신(當歸身) | 산사(山査) , 산사육(山查肉) | 육종용(肉蓯蓉) , 종용(蓯蓉) |
| 감국(甘菊) , 감국화(甘菊花) , 국화(菊花) | 대조(大棗) , 조(棗) | 산수유(山茱萸) , 산수유육(山茱萸肉) | 익지(益智) , 익지인(益智仁) |
| 감조(甘草) , 자감초(炙甘草) | 마인(麻仁) , 마자인(麻子仁) | 상백피(桑白皮) , 청상피(青桑皮) | 인진(茵蔯) , 인진호(茵蔯蒿) |
| 강잠(僵蠶) , 백강잠(白殭蠶) | 명반(明礬) , 백반(白礬) | 생건지황(生乾地黃) , 생지황(生地黃) | 자완(紫菀) , 자원(紫苑) |
| 건강(乾薑) , 백강(白薑) | 목단(牧丹) , 목단피(牧丹皮) | 서각(犀角) , 서각설(犀角屑) | 자조(紫草) , 자조용(紫草茸) |
| 견우(牽牛) , 견우자(牽牛子) | 박하(薄荷) , 박하엽(薄荷葉) | 서점자(鼠粘子) , 악실(惡實) , 우방자(牛蒡子) | 적작(赤芍) , 적작약(赤芍藥) |
| 경삼릉(京三棱) , 형삼릉(荊三棱) | 반하국(半夏麴) , 반하국(半夏麴) | 석련육(石蓮肉) , 연육(蓮肉) , 연자(蓮子) | 주사(朱砂) , 진사(辰砂) |
| 계심(桂心) , 관계(官桂) , 날계(辣桂) , 육계(肉桂) | 방분(蚌粉) , 합분(蛤粉) | 선각(蟬殼) , 선태(蟬蛻) , 선퇴(蟬退) | 죽여(竹茹) , 청죽여(青竹茹) |
| 계지(桂枝) , 계피(桂皮) | 백견우(白牽牛) , 백축(白丑) | 소엽(蘇葉) , 자소(紫蘇) , 자소엽(紫蘇葉) | 지각(枳殼) , 지실(枳實) |
| 고량강(高良薑) , 양강(良薑) | 백두구(白豆蔻) , 백두구인(白頭蔻仁) | 소자(蘇子) , 자소자(紫蘇子) | 천근(茜根) , 측백엽(側柏葉) |
| 과루근(瓜蔞根) , 괄루근(栝蔞根) | 백복령(白茯苓) , 복령(茯苓) , 적복령(赤茯苓) | 소회향(小茴香) , 회향(茴香) | 천동(天冬) , 천문동(天門冬) |
| 과루인(瓜蔞仁) , 괄루인(栝蔞仁) | 백복신(白茯神) , 복신(茯神) | 속각(粟殼) , 앵속각(罌粟殼) | 조과(草果) , 조과인(草果仁) |
| 괴각(槐角) , 괴화(槐花) | 백석지(白石脂) , 적석지(赤石脂) | 신곡(神麴) , 신국(神麴) | 종(蔥) , 총백(蔥白) |
| 귤피(橘皮) , 귤홍(橘紅) , 진피(秦皮) , 진피(陳 | 백작(白芍) , 백작약(白芍藥) , 작약(芍藥) | 신이(辛夷) , 신이화(辛夷花) | 파극(巴戟) , 파극천(巴戟天) |
| 금불초(金沸草) , 선복화(旋覆花) | 백편두(白扁豆) , 편두(扁豆) | 아교(阿膠) , 아교주(阿膠珠) | 파두(巴豆) , 파두상(巴豆霜) |
| 남성(南星) , 천남성(天南星) | 백하수오(白何首烏) , 적하수오(赤何首烏) | 연호색(延胡索) , 현호색(玄胡索) | 향여(香茹) , 향유(香薷) |
| 녹각교(鹿角膠) , 녹각상(鹿角霜) | 봉아무(蓬莪茂) , 봉아출(蓬莪朮) , 봉출(蓬朮) , | 오미(五味) , 오미자(五味子) | 형개(荊芥) , 형개수(荊芥穗) |
| 단향(檀香) , 백단(白檀) | 부자(附子) , 포부자(炮附子) | 용담초(龍膽草) , 초용담(草龍膽) | 흑견우(黑牽牛) , 흑견우자(黑牽牛子) , 흑축(黑丑) |
| 담성(膽星) , 우담남성(牛膽南星) | 빙편(氷片) , 편뇌(片腦) | 우슬(牛膝) , 천우슬(川牛膝) | |
| **유형A2 (초기 데이터 오류) : 4종** | | | |
| 마황(馬嘶) , 마황(麻黃) | 망소(芒消) , 망초(芒硝) | 백지(柏脂) , 백지(白芷) | 호초(好醋) , 호초(胡椒) |
| **유형B (일부 유의어 관계) : 11종** | | | |
| 가려늑(訶黎勒) , 백단향(白檀香) , 소합유(蘇合油) , 소합향(蘇合香) , 안식향(安息香) , 오서각(烏犀角) , 청목향(青木香) | 구맥(瞿麥) , 구맥수(瞿麥穗) , 저제(猪蹄) | 대청(大青) , 두시(豆豉) , 향시(香豉) | 자연동(自然銅) , 호경골(虎脛骨) , 호골(虎骨) |
| 감인(芡仁) , 검실(芡實) , 금앵자(金櫻子) | 구판(龜板) , 귀판(龜板) , 쇄양(鎖陽) | 대풍자(大風子) , 호마(胡麻) , 호마자(胡麻子) | 회산약(懷山藥) , 회생지황(懷生地黃) , 회숙지황(懷熟地黃) |
| 곤포(昆布) , 해대(海帶) , 귤핵(橘核) | 규자(葵子) , 동규자(冬葵子) , 편축(萹蓄) , 해금사(海金沙) | 복분자(覆盆子) , 저실(楮實) , 저실자(楮實子) | |
| **유형C1 (호환 가능 관계) : 13종** | | | |
| 감수(甘遂) , 대극(大戟) , 원화(芫花) | 구척(狗脊) , 속단(續斷) | 양경골(羊脛骨) , 호동루(胡桐淚) | 청용석(青礞石) , 풍화초(風化硝) |
| 갱미(粳米) , 소맥(小麥) | 금박(金箔) , 우황(牛黃) , 천축황(天竺黃) | 애엽(艾葉) , 익모초(益母草) | |
| 건칠(乾漆) , 맹충(蝱蟲) , 수질(水蛭) | 백면(白麵) , 속미(粟米) | 우유(牛乳) , 인유(人乳) | |
| 관동화(款冬花) , 자완용(紫菀茸) | 소석(消石) , 태음현정석(太陰玄精石) | 육리인(郁李仁) , 화마인(火麻仁) | |
| **유형C2 (무관한 관계) : 40종** | | | |
| 감송(甘松) , 삼내(三柰) , 영릉향(零陵香) | 귤엽(橘葉) , 염(鹽) | 물석자(沒石子) , 석류피(石榴皮) | 수은(水銀) , 조석(硝石) |
| 감조절(甘草節) , 조각자(皂角刺) | 금모구척(金毛狗脊) , 지룡(地龍) | 박소(朴消) , 백활석(白滑石) | 식수유(食茱萸) , 해조(海藻) |
| 강진향(降眞香) , 자금피(紫金皮) , 황백지(香白 | 난발(亂髮) , 노봉방(露蜂房) , 도지(桃枝) , 상지 | 발회(髮灰) , 봉방(蜂房) | 아위(阿魏) , 요사(硇砂) |
| 건모과(乾木瓜) , 정피(丁皮) | 녹반(綠礬) , 위피(蝟皮) | 백선피(白鮮皮) , 토복령(土茯苓) | 은시호(銀柴胡) , 호황련(胡黃連) |
| 건산약(乾山藥) , 추석(秋石) | 누로(漏蘆) , 박소(朴硝) | 백조상(百草霜) , 편자강황(片子薑黃) | 이어(鯉魚) , 저요자(猪腰子) |
| 경묵(京墨) , 택란(澤蘭) | 담두시(淡豆豉) , 해백(蝦白) | 복령피(茯苓皮) , 생강피(生薑皮) | 저아조각(猪牙皂角) , 홍내소(紅內消) |
| 괴지(槐枝) , 백엽(柏葉) , 송지(松脂) , 윤모(蜳 | 마두령(馬兜鈴) , 상엽(桑葉) | 비해(萆薢) , 석남엽(石南葉) | 조협(皂莢) , 축조(蜀椒) |
| 권백(卷柏) , 백미(白薇) | 마발(馬勃) , 절패모(浙貝母) , 판람근(板藍根) | 산두근(山豆根) , 아소(牙消) | 토과근(土瓜根) , 황련(黃連) |
| 귀구(鬼臼) , 반석(礬石) , 자황(雌黃) | 마황근(麻黃根) , 황백(黃蘗) | 석웅황(石雄黃) , 영사(霜砂) | 편금(片芩) , 화피(樺皮) |
| 귀전우(鬼箭羽) , 도노(桃奴) | 목별자(木鼈子) , 백급(白芨) | 석위(石葦) , 왕불류행(王不留行) | 호도육(胡桃肉) , 홍조(紅棗) |

우 일반적인 한의학 지식을 바탕으로 그 유사성을 이해할 수 있다. 【감수(甘遂), 대극(大戟), 원화(芫花)】, 【건칠(乾漆), 맹충(蝱蟲), 수질(水蛭)】 등의 집합은 교과서적으로 약성(藥性)이 유사한 약재가 짝지어진 경우이다. 또한 【갱미(粳米), 소맥(小麥)】, 【금박(金箔), 우황(牛黃), 천축황(天竺黃)】, 【우유(牛乳), 인유(人乳)】 등도 약재의 성상과 기원을 통해 결과로 도출된 이유를 어느 정도 납득할 수 있는 경우이다. 이들은 작업자가 기대한 유의어 관계는 아니지만, 용어가 가지는 의미 측면에서 어느 정도 통찰을 준다.

이상의 결과를 통해 수동 단계에서 고려되어야 할 몇 가지 사항을 정리할 수 있다. 첫째, 데이터 자체의 오류이다. ［유형A2］와 같은 경우로 작업자는 데이터 자체의 오류 가능성을 상정하고 검토에 임해야 한다. 둘째, 표기 방법의 문제이다. 예시 결과에서 【신곡(神麴), 신국(神麴)】(한자漢字의 이체자 문제), 【구판(龜板), 귀판(龜板)】(디지털 텍스트의 다중코드자 문제), 【원삼(元蔘), 현삼(玄蔘)】(동아시아 피휘避諱의 문제) 등이 그것이다. 작업자는 같은 의미의 텍스트가 달라질 수 있는 패턴을 어느 정도 숙지하고 있어야 한다.

## Ⅲ. 결론

지금까지 기계 학습을 이용하여 한의학 용어에 대한 유의어 사전 구축 방법을 제안하고 이 방법으로 예시 데이터를 분석하여 간단한 유의어 사전을 구축해 보았다. 결론을 대신하여 본 연구에서 제안한 방법이 가지는 특징과 한계를 짚어보고자 한다.

유의어 사전 구축 작업이 어려운 가장 큰 이유는 자연어로서 용어의 의미가 분절되지 않고 연속적이기 때문이다. 예를 들어 【감초(甘草), 자감초(炙甘草)】, 【감초(甘草), 생감초(生甘草)】의 쌍이 존재한다면, 논리적으로 【자감초(炙甘草), 생감초(生甘草)】의 쌍이 성립해야 한다. 하지만 앞의 2가지가 쉽게 수긍되더라도 자감초와 생감초가 유의어 관계인가에 대해서는 의견이 나뉠 수 있다. 유의어는 용어 사이에 의미의 범위가 겹치는 관계이므로 이런 용어들을 모으다 보면 어떤 순간에는 이처럼 의미의 범주 사이에 모순이 나타나기도 한다. 이런 경우 작업자의 판단이 끊임없이 요구되므로 기존의 수작업 방법에서는 작업자마다 결과가 달라지기 쉽고 동일한 작업자라고 해도 시간에 따라 작업 결과가 달라질 수 있다.

본 연구에서 제안한 자동 단계는 이러한 부분에 상당한 도움을 줄 수 있다. 기계 학습 결과가 정답이라고 단정할 수는 없지만, 사람의 작업 결과보다는 일관된 결과를 도출해 줄 수 있기 때문이다. 자동 단계의 결과는 작업자들 사이에 공유할 수 있는 공통된 작업 기준이 될 수 있다. 그뿐만 아니라 본 연구에서와같이 자동 단계 이후에 수동 단계를 진행하는 경우에는 일관된 중간 결과를 바탕으로 작업하는 효과를 볼 수 있으므로 수작업에 드는 시간과 노력을 크게 단축할 수 있다.

하지만 자동 단계에는 출현 빈도가 낮은 용어를 학습시킬 수 없다는 문제뿐만 아니라 작업자가 임의로 지정할 수밖에 없는 초매개변수가 적절하지 않거나 용어 자체에 내재된 의미가 명확하게 분리되지 않는 등 다양한 이유로 누락된 군집이 발생할 수 있다. 자동 단계에서 일단 누락되면 이를 다시 찾기는 쉽지 않다. 본 예시 분석에서도 【녹두(綠豆), 녹두(菉豆)】, 【향부(香附), 향부미(香附米), 향부자(香附子)】, 【산치인(山梔仁), 산치자(山梔子), 치자(梔子), 치자인(梔子仁)】 등이 반복 작업에서 과반 이상 관찰되지 않아서 최종 결과에서는 빠진 것을 확인할 수 있었다. 이는 자동 단계의 초매개변수를 조정하거나[9] 보조적인 방법을 추가하여[10] 극복해야 할 과제이다.

## 감사의 글

본 연구는 한국한의학연구원 주요사업 “AI 한의사 개발을 위한 임상 빅데이터 수집 및 서비스 플랫폼 구축(KSN2021110)”의 지원을 받아 수행되었습니다.

## References

1. 강형석, 양장훈. 한국어 단어 임베딩 모델의 평가에 적합한 유추 검사 세트. 디지털콘텐츠학회논문지. 2018. 19(10). pp.1999-2008.
2. 김혜은 외 4인. 한의학 증상용어의 형태소 분석을 위한 자연어 표기 분석. 대한예방한의학회지. 2013. 17(2). pp.179-187.
3. 백유상. 한의학정보 검색엔진 개발을 위한 시소러스 연구. 대한한의학원전학회지. 2006. 19(1). pp.155-167.
4. 오준호. 한의학 고문헌 텍스트 분석을 위한 비지도학습 기반 단어 추출 방법 비교. 대한한의학원전학회지. 2019. 32(3). pp.47-57.
5. 오준호. 한의학 고문헌 데이터 분석을 위한 단어 임베딩 기법 비교 : 자연어처리 방법을 적용하여. 대한한의학원전학회지. 2019. 32(1).

9) 예시 분석의 경우, 군집 구성에서 군집을 나눌 초매개변수를 현행 기준보다 1에 가깝게 조정하는 방법, 혹은 반복 작업에서 최종 결과로 수용하는 기준을 현행 5회에서 3회나 4회로 줄이는 방법 등을 들 수 있다.

10) 예를 들어 자동 단계와 수동 단계를 모두 수행하고 난 뒤에 최종 결과를 제외한 나머지 용어만으로 다시 재차 자동 단계의 군집 구성에서부터 수동 단계까지의 작업을 반복하는 방법을 들 수 있다. 이렇게 반복해 나간다면 처음 자동 단계에서 누락된 조합이 추가로 도출될 수 있을 것이다.

pp.61-74.
6. 이병욱, 심범상, 엄동명. 한의학 용어관리 시스템을 결합한 고전원문 제공 서비스에 관한 연구. 대한한의학원전학회지. 2009. 22(4). pp.167-176.
7. 차승준 외 6인. 한의학 용어 수집 및 관리 시스템 구축. 대한예방한의학회지. 2010. 14(1). pp.59-76.
8. 장현철 외 18명. 온톨로지 기반 한의학 지능형 정보체계 연구. 대전. 한국한의학연구원. 2013.
9. 특허청. 한국전통지식포털. [cited on Jan 12, 2019]. Avaiable from: http://www.koreantk.com