대한한의학원전학회지
2019년 08월 32권 3호

pISSN 1229-8328
J Korean Med Classics 2019:32(3) 047-057
https://doi.org/10.14369/jkmc.2019.32.3.047

JKMC
Journal of Korean Medical Classics

# 한의학 고문헌 텍스트 분석을 위한 비지도학습 기반 단어 추출 방법 비교

한국한의학연구원 연구원
**오준호***

# Comparison of Word Extraction Methods Based on Unsupervised Learning for Analyzing East Asian Traditional Medicine Texts

Oh Junho*
Researcher at Korea Institute of Oriental Medicine

**Objectives** : We aim to assist in choosing an appropriate method for word extraction when analyzing East Asian Traditional Medical texts based on unsupervised learning.

**Methods** : In order to assign ranks to substrings, we conducted a test using one method(BE:Branching Entropy) for exterior boundary value, three methods(CS:cohesion score, TS:t-score, SL:simple-ll) for interior boundary value, and six methods(BExSL, BExTS, BExCS, CSxTS, CSxSL, TSxSL) from combining them.

**Results** : When Miss Rate(MR) was used as the criterion, the error was minimal when the TS and SL were used together, while the error was maximum when CS was used alone. When number of segmented texts was applied as weight value, the results were the best in the case of SL, and the worst in the case of BE alone.

**Conclusions** : Unsupervised-Learning-Based Word Extraction is a method that can be used to analyze texts without a prepared set of vocabulary data. When using this method, SL or the combination of SL and TS could be considered primarily.

**Key words** : Text segmentation, Word extraction, tokenization, East Asian Traditional Medicine, Korean medicine

* 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(July 22, 2019), Revised(August 5, 2019), Accepted(August 5, 2019)

# Ⅰ. 서론

한의학 고문헌은 전통사회에서 만들어진 동아시아 전통의학 문헌을 가리킨다. 이 문헌들은 의술을 배우고 연구하는 목적으로 오랜 시간동안 전승되며 읽혀 왔다. 그 과정에서 일부 문헌들은 세월의 깊이를 이기지 못하고 사라지기도 하였고, 어떤 문헌들은 뜻있는 의가(醫家)에 의해 새로 저술되기도 하였다. 이렇게 오늘날까지 남겨진 한의학 고문헌은 한의학 지식의 보고(寶庫)라고 할 수 있다.

현대에 이르러 출판 및 연구의 편의를 위해 이들 문헌 속 내용이 디지털 형태로 다시 만들어지고 있다. 문헌의 내용을 디지털 텍스트로 구축하면 방대한 분량을 작은 저장 공간에 담을 수 있을 뿐만 아니라 검색과 열람에도 용이하다. 동아시아 전통의학에서 가장 많은 서적을 보유하고 있는 중국은 이미 상당한 양의 디지털 텍스트를 구비하고 있다. 〈中华医典〉의 경우, 역대 의서 1,156부를 수록하고 있는데, 문자로 4억여자에 달하는 방대한 규모이다.[1] 한국은 현재 한국한의학연구원에서 854만여자 규모 79종의 의서를 온라인 웹서비스 〈한의학고전DB〉를 통해 공개하고 있다.[2] 중국의 성과에 비하면 크게 미치지 못하는 분량이지만 자료의 양이 꾸준히 늘어나고 있다.

그간 이러한 전자 텍스트들은 출판·열람·검색의 용도로 이용되어 왔으나, 텍스트와 같은 비정형 데이터의 분석 기법이 발전하면서 데이터 분석을 위한 자원으로 새롭게 각광받고 있다. 그러나 이를 위해서는 많은 준비가 요구된다. 텍스트를 분석하기 위해서는 분석 대상이 될 데이터가 디지털 형태로 마련되어야 하며, 목적에 맞는 적절한 분석 방법과 모델이 개발되어야 한다. 이미 분석 방법과 모델에 대한 연구는 정보학, 컴퓨터공학, 수학 등 다양한 학문 영역에서 연구 되고 있다. 그러나 분석 대상이 되는 데이터의 성격에 따라 어떠한 방법을 어떠한 방식으로 적용해야 하는지는 여전히 분석자가 판단해야 할 몫이다. 특히 텍스트 데이터는 어떤 언어로 작성되었는지에 따라 데이터를 가공하는 방식이 달라지기 때문에 이에 대한 검토가 더욱 중요하다.

한의학 고문헌은 동아시아 전통사회에서 식자층이 사용하던 한문(漢文)으로 기술되어 있다. 고대 한문은 문법이나 단어 형태가 현대 중국어와 상이하므로 현대 중국어에 적용되고 있는 자연어처리 기술을 그대로 적용시킬 수 없다. 그뿐만 아니라 한의학 고문헌은 한의학이라는 특수한 주제를 다루고 있는 기술 서적이다. 치료 대상이 되는 병증과 치료 방법이 되는 방제, 본초, 경혈 등의 용어들이 내용의 대부분을 이루고 있기 때문에 이에 대한 고려도 필요하다. 이에 본 연구는 한의학 고문헌 텍스트를 분석할 때 제기되는 단어 추출에 대한 문제를 다루어 보고자 한다.

# Ⅱ. 본론

## 1. 배경 : 단어 추출의 문제

자연어로 이루어진 비정형의 텍스트 데이터를 분석하기 위해서는 기본적으로 분석의 최소 단위가 도출되어야 한다. 컴퓨터에게 텍스트 자료는 단순한 문자(character)의 나열에 지나지 않기 때문이다. 이러한 분석의 최소 단위를 토큰(Token)이라고 한다. 텍스트에서 토큰을 추출하는 것, 즉 텍스트를 토큰의 집합으로 환원시켜 주는 것을 토큰 추출(Tokenization)이라고 한다.

텍스트 분석에서는 일반적으로 해당 언어의 단어를 토큰으로 삼는다. 인간이 단어를 기반으로 언어를 구사하기 때문이다. 따라서 토큰 추출은 주어진 텍스트에서 분석에 필요한 단어를 추출한다는 점에서 단어 추출(Word Extraction)이라고 할 수 있다. 또한 중국어나 일본어처럼 띄어쓰기가 존재하지 않거나, 어떠한 이유로 띄어쓰기 정보가 삭제된 텍스트의 경우에는 텍스트를 적절하게 나누어주는 것만으로 토큰이 드러나게 된다. 따라서 이러한 경우에는 텍스트를 적절하게 구분해 준다는 의미에서 단어 구분(Word Segmentation)이라고 지칭하기도 한다.[3]

1) 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
2) 한국한의학연구원. 한의학고전DB. [cited on July 17, 2019]. Avaiable from: https://mediclassics.kr

영미권 언어의 경우, 띄어쓰기 및 구두점(punctuation marks)으로 텍스트를 분절하는 것만으로 이 문제를 어느 정도 해소할 수 있다. 한국어는 어미의 변화로 의미의 변화를 표현하기 때문에 이를 분석할 수 있는 형태소분석기가 필요하다.[4] 현대 중국어의 경우에는 텍스트 내에서 의미에 따라 단어를 구분해 주어야 하며, 이를 위한 도구가 개발되어 있다.[5] 그러나 한의학 고문헌의 경우에는 고대 한문으로 이루어져 있기 때문에 이들 방법들을 그대로 적용할 수 없다.

전통 사회의 한문(漢文)은 표의문자로서 하나의 글자로도 의미 전달이 가능하다. 따라서 하나의 글자 자체가 단어의 역할을 한다. 하지만 글자와 글자가 모여 더 구체적인 의미를 나타내는 경우가 많으며, 대부분의 한의학 용어들은 이렇게 구성되어 있다. 이런 경우에는 몇 가지 글자가 합쳐져 하나의 의미를 나타낸다.

텍스트에서 분석의 최소 단위가 되는 토큰, 즉 단어를 추출하는 가장 확실한 방법은 '용어집(dictionary or vocabulary)'을 미리 구축하였다가 주어진 텍스트에 적용시키는 방법이다. 예를 들어 "治勞役太甚或飮食失節身熱而煩自汗倦怠"라는 텍스트가 주어졌다고 하자. 용어집에 '勞役', '飮食', '飮食失節', '身熱', '煩', '自汗', '倦怠'과 같은 단어가 이미 등록되어 있다고 한다면, 주어진 텍스트에서 이러한 단어들이 포함되어 있는지를 확인하여 ['勞役', '飮食失節', '身熱', '煩', '自汗', '倦怠']와 같이 단어를 추출해 낼 수 있다. 이때 용어집에 없는 용어는 분석에서 배제되므로 용어집의 완성도가 분석 결과에 크게 영향을 미친다.

그러나 용어집을 만드는 일은 상당히 방대한 작업이다. 이미 존재하는 일반 사전의 표제어를 사용하는 방안이 있을 수 있으나, 좋은 성능을 기대하기는 힘들다. 자연어 텍스트는 일반 사전에 등재되어 있지 않은 많은 표현을 포함하고 있기 때문이다. 예를 들어 薑附는 乾薑과 附子를 줄인 표현으로, 일반 사전에는 표제어로 등록되어 있지 않다. 또한 蘿菖은 萊菔, 蘿蔔, 萊卜, 蘿卜, 萊菔, 蘿菔, 萊菖, 蘿菖, 萊蔔, 蘿蔔 등 다양한 표기가 존재한다. 다양한 표기 역시 일반 용어집에는 표제어로 등록되어 있지 않다. 따라서 분석을 위해서는 이런 자연어 표기를 포괄할 목적으로 만들어진 용어집이 요구된다.

만약 용어집이 없다면 분석을 시도할 수 없는 것일까. 용어집이 없는 경우에도 어느 정도의 분석은 가능하다. 가장 간단한 방법은 글자 수를 기준으로 텍스트를 기계적으로 분절하는 방법이 있다. 앞의 예의 문장을 예로 들면, 한 글자를 토큰으로 하여 "治 勞 役 太 甚 或 飮 食 失 節 身 熱 而 煩 自 汗 倦 怠"로 만들거나, 두 글자를 기준으로 "治勞 勞役 役太 太甚 甚或 或飮 飮食 食失 失節 節身 身熱 熱而 而煩 煩自 自汗 汗倦 倦怠"로 토큰을 추출할 수 있다. 이러한 방법은 별다른 준비 없이 토큰을 추출할 수 있으나 '役太', '甚或', '節身' 등 원하지 않는 토큰이 다수 나타날 가능성이 크기 때문에 분석 결과가 부정확해 질 수 있다.

이에 대한 대안으로, 텍스트 자체의 특성을 통해 '단어로 추정'되는 문자열을 도출해 내는 방법이 있다. 주어진 텍스트에서 도출 가능한 문자열을 모두 도출한 다음, 각 문자열이 단어일 '점수(score)'를 계산한다(이하 '추출점수'로 지칭함). 이렇게 하면 이 추출점수를 기준으로 문자열에 순위를 부여할 수 있다. 주어진 텍스트에서 순위가 높은 문자열부터 뽑아내면 단어를 추출하는 것과 유사한 결과를 얻을 수 있다.(<Fig. 1> 참조)

이를 위해서는 텍스트의 특성을 반영하는 대량의 텍스트와, 문자열의 추출점수를 계산할 방법(measure)이 필요하다. 이를 기계학습의 관점에서 보면, 미리 단어라고 규정된 '참값(ground truth)'이 없으므로 '비지도학습(unsupervised learning)'에 해당한다. 이 방법은 문자 개수를 기준으로 단순히 구분하였을 때보다 오류를 줄일 수 있다.

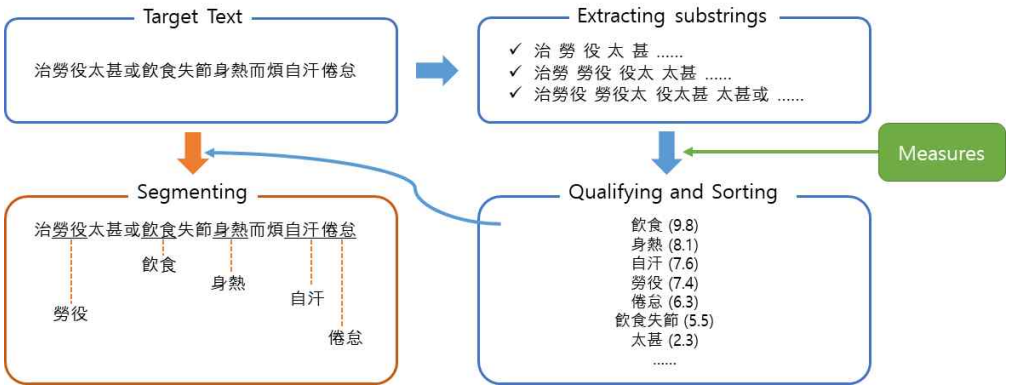
현재 한의학 고문헌 텍스트를 분석하는 연구가

3) 본고에서는 단어 추출이라는 용어를 주로 사용하되, 더 구체적인 의미로 표현해야 하는 경우에 토큰 추출(Tokenization)이나 단어 구분 혹은 분절(Word Segmentation)이라는 용어를 사용하기로 한다.

4) 한국어 형태소 분석기로는 Kkma, Komoran, Hannanum, Twitter, Mecab 등이 있다.

5) 중국어 단어 구분기로는 Stanford Word Segmenter, jieba 등이 있다.

<Fig. 1> Concept Diagram of word extraction based on unsupervised learning




다양하게 시도되고 있으나 정작 이에 필수적인 용어집은 준비되지 못한 실정이다. 이런 상황에서 한의학 고문헌 텍스트를 분석할 때 용어집 없이 단어를 추출하는 비지도학습 기반의 토큰 추출 방법이 대안이 될 있다. 이에 본 연구에서는 비지도학습 기반의 단어 추출 방법에 적용할 수 있는 몇 가지 방법을 수행하고 그 결과를 비교하였다.

## 2. 텍스트 데이터 준비

본 연구를 수행하기 위해서는 크게 3가지 데이터가 요구된다. 문자열의 특성을 파악하기 위한 학습용 데이터(training data), 실제 단어 추출을 수행할 실험용 데이터(test data), 그리고 그 결과를 검토해 볼 검증용 데이터(validation data)가 그것이다.

학습용 데이터는 문자열의 특성을 도출해야 하므로 크기가 크고 해당 분야의 특성을 잘 담고 있는 데이터일 필요가 있다. 이에 본 연구에서는 학습용 데이터로 한국한의학연구원에서 구축하고 있는 한의학 고문헌 텍스트를 사용하였다.[6] 이 데이터는 한자(漢字) 기준으로 17,842,632자 규모이다.

실험용 데이터로는 데이터의 완성도가 좋고 인지도가 높은 《동의보감》 원문을 사용하였다. 한자 기준 870,540자 크기의 데이터이다. 검증용 데이터는 성능평가를 위한 데이터이다. 따라서 데이터에 텍스트 분절에 대한 참값(ground truth)이 포함되어 있어야 한다. 하지만 현재 객관적으로 인정받은 텍스트 분절 혹은 단어 추출에 대한 데이터는 존재하지 않는다. 본 연구에서는 차선책으로 《동의보감》의 표점 정보를 이용하였다. 본 연구에 사용한 《동의보감》 텍스트의 표점은 219,475건이며, 표점과 표점 사이에 평균 3.966자가 포함되어 있다.

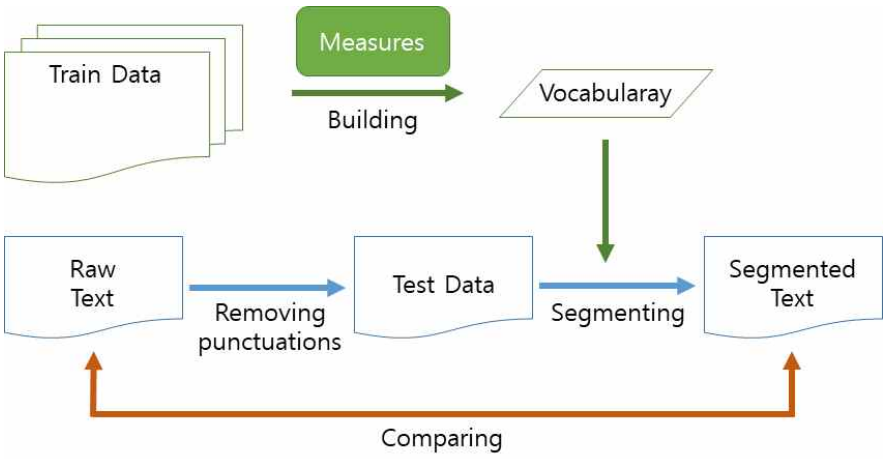
표점은 한 문장 중간 혹은 한 문장이 끝난 후의 휴지를 표시하기 위한 기호이다[7]. 표점으로 나누어진 부분은 텍스트 분절에서 반드시 나뉘어져야 하는 곳으로, 표점은 텍스트 분절의 최소 기준이라고 할 수 있다. 다만 표점으로 나뉘지 않았다고 해서 분절해서는 안 될 지점이라고 할 수는 없다. 다시 말해 표점으로 나뉘지 않은 부분은 분절의 참 거짓을 판단할 수 없다. 이 부분은 본 연구의 평가 방법이 가지는 한계이다.

《동의보감》 텍스트의 표점 정보를 이용하면 다음과 같이 텍스트 분절 성능을 평가할 수 있다. 여러 방법으로 텍스트를 분절(ⓐ)하고, 이 분절 부분을 《동의보감》 텍스트에서 표점으로 분절된 부분(ⓑ)과 비교한다. 이때 ⓑ의 개수는 고정되어 있으므로, ⓐ가 ⓑ에 위배되는 개수가 적을수록 오류가 적은 방법이라고 할 수 있다. 한 가지 더 고려할 사항은 텍스트가 작게 나누어질수록 표점과 겹치지 않게 된다는 점이다. 극단적으로 한 글자로 텍스트를 분절하면, 표점에 위배되는 경우가 전혀 나타나지 않게

6) 학습용 데이터는 본 연구를 위한 목적으로 한국한의학연구원으로부터 제공 받아 사용되었다.

7) 黃永年(김언종, 김수경 옮김). 고적정리개론. 한국고전번역원. 2013. p.209.

<Fig. 2> Concept Diagram for comparison of word extraction methods




된다. 따라서 분절 수에 비해 오류가 적은 방법인지 탐색해 보아야 한다.

본고에서는 먼저 여러 가지 점수 측정 방법으로 《동의보감》 텍스트를 분절한 다음, 《동의보감》 표점과 위배되는 부분의 수(M: Miss)를 측정하였다. 이를 《동의보감》 전체 표점으로 나누어 전체 표점에서 표점과 위배된 비율(MR: Miss Rate)을 계산하였다. 과도하게 분절된 경우에는 이를 보정해 주어야 하므로, 분절된 수량을 가중치로(S: Segments) 곱하여 최종 결과를 수치로 표현하였다.

## 3. 점수 측정 방법(measurements)

비지도학습 기반 단어 추출 방법으로 몇 가지 방법이 제안되어 있다. 이 방법들은 크게 외부경계값(exterior boundary value)을 이용한 방법과 내부경계값(interior boundary value)을 이용한 방법으로 나눌 수 있다.[8] 외부경계값은 단어와 인접한 문자가 다양하게 변화된다는 특성을 이용한 방법으로 Branching Entropy 등이 있다. 내부경계값은 단어를 이루는 문자 집합이 그렇지 않은 경우에 비해 더 공고하게 연결되어 있다는 점에 착안한 방법으로, 연속되는 문자에 점수를 부여하여 서로 비교하는 방법을 취한다. 점수를 부여하는 방법에 따라 cohesion score, t-score, simple-ll 등으로 나눌 수 있다. 본 연구에서는 외부경계값을 이용한 방법 1가지와 내부경계값을 이용한 방법 3가지, 그리고 이들을 2가지씩 조합한 방법 6가지로 테스트를 진행하였다.

보통 하나의 단어에서 앞과 뒤의 문자는 매우 다양하게 나타나기 마련이다. 텍스트 상에 “甘草各…”, “甘草之…”, “甘草大…” 등이 출현한다면, “甘草”라는 문자열이 하나의 단어가 될 가능성이 크다고 추측할 수 있다. 텍스트 상에서 “甘草”의 앞이나 뒤에 오는 문자들이 다양하게 출현하고 있기 때문이다. 이러한 특성을 이용해 문자열 밖에 존재하는 문자를 통해 단어의 경계를 정할 수 있다.

Branching Entropy(BE)는 이러한 외부경계값을 이용한 방법이다. BE는 단어의 경계에서 그 다음 문자에 대한 불확실성(Entropy)이 높아진다고 보고 이를 측정한다.[9] 측정 공식은 다음과 같다.

$$BE(X|X_n = x_n) = -\sum_{x \in \chi} P(x|x_n)\log P(x|x_n)$$

“甘草”, “食甘”, “甘則”라는 문자열 가운데, “甘

8) 김현중, 조성준, 강필성. KR-WordRank : WordRank를 개선한 비지도학습 기반 한국어 단어 추출 방법. 대한산업공학회지. 2014. 40(1). pp.18-33.

9) Zhihui Jin, Kumiko Tanaka-Ishii. Unsupervised Segmentation of Chinese Text by Use of Branching Entropy. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. 2006. pp.428-435.

草"가 단어가 된다는 사실을 어떻게 알 수 있을까. 그것은 "食" 다음에 "甘"이 오거나, "甘" 다음에 "則"이 나타나는 경우 보다 "甘"에 이어 "草"라는 문자가 오는 경우가 압도적으로 많기 때문이다. 다시 말해 '甘'과 '草'라는 문자가 매우 자주 즐겨 만난다는 사실을 통해 "甘草"가 단어일 가능성이 크다는 것을 알 수 있다. 따라서 문자와 문자가 얼마나 자주 즐겨 만나는지를 측정할 수 있다면, 단어가 될 가능성이 높은 문자열을 추출해 낼 수 있을 것이다. 이런 방법을 이용한 것이 내부경계값이다.

cohesion score(CS)는 문자열이 나타날 조건부 확률을 곱해 나가는 방법으로, 해당 문자열이 나타날 확률에 문자열의 길이로 가중치를 부여한 방법이다. 측정 방법은 다음과 같다[10].

$$CS(c_1,...,c_n) = \sqrt[n]{\prod_{i=1}^{n-1} P(c_1,...,c_{i+1}|c_1,...,c_i)}$$

t-score(TS)과 simple-ll(SL)는 문자가 함께 나타나는 정도에 대한 측정값이다. 이 방법들은 이론적인 기댓값과 실제 관찰값의 차이를 통해 측정값을 도출한다. 문자열이 있을 때, 텍스트 안에서 대상 문자열이 관찰된 실제 관찰값(observed value)을 O, 문자의 빈도로부터 문자열이 나타날 기댓값(Expected value)를 E라고 하였을 때, 다음과 같이 계산된다.[11]

$$TS = \frac{O-E}{\sqrt{O}}$$

$$SL = 2\left(O \cdot \log\frac{O}{E} - (O-E)\right)$$

각각의 측정 방법은 독립적으로 사용할 수도 있지만, 서로 조합하여 사용할 수도 있다. 동일한 문자열에 대해 다른 방법으로 점수를 도출한 다음, 결과를 서로 곱하는 방식이다. 앞에서 설명한 BE, CS, TS, SL을 이러한 방식으로 조합해 보면, BE와 CS의 곱(BExCS), BE와 TS의 곱(BExTS), BE와 SL의 곱(BExSL), CS와 TS의 곱(CSxTS), CS와 SL의 곱(CSxSL), TS와 SL의 곱(TSxSL) 등 6가지 방법이 추가로 만들어진다.

## 4. 결과 비교

본 연구에서는 앞서 소개한 4가지 방법과 이들을 조합한 6가지 방법, 총 10가지 방법을 사용하여 《동의보감》 텍스트를 분절하였다. 각각의 방법으로 분절된 《동의보감》 텍스트와 표점으로 분절된 기존의 텍스트를 비교하여 얼마나 오류가 있는지 확인해 보았다.

먼저 《동의보감》에서 추출한 문자열을 위의 10가지 방법에 따라 추출점수를 매겨 순위를 결정하였다. 각각의 방법에서 추출점수가 가장 높은 상위 10종의 문자열을 제시해 보면 다음과 같다.(<Table. 1> 참조, 내림차순)

10) Hyunjoong Kim. LOVITxDATA SCIENCE. [cited on July 17, 2019]. Avaiable from: https://lovit.github.io/nlp/2018/04/09/cohesion_ltokenizer

11) 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. Springer. 2008. pp 52-63.

**<Table. 1> Top 10 substring list**

| method | substring list | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| BE | 不能 | 而不 | 寒熱 | 一兩 | 不得 | 不可 | 二兩 | 者爲 | 之氣 | 三分 |
| CS | 霹靂 | 蜈蚣 | 琥珀 | 螵蛸 | 醍醐 | 膀胱 | 躑躅 | 崑崙 | 吸咀 | 葶藶 |
| TS | 一兩 | 甘草 | 各一 | 一錢 | 爲末 | 半兩 | 每服 | 三分 | 當歸 | 二錢 |
| SL | 一兩 | 甘草 | 各一錢 | 當歸 | 爲末 | 茯苓 | 麥門冬 | 各一 | 一錢 | 半兩 |
| BExSL | 一兩 | 甘草 | 各一錢 | 當歸 | 爲末 | 半兩 | 茯苓 | 各一兩 | 小便 | 麥門冬 |
| BExTS | 一兩 | 半兩 | 不可 | 不能 | 甘草 | 三分 | 二兩 | 爲末 | 當歸 | 一錢 |
| BExCS | 膀胱 | 桔梗 | 檳榔 | 當歸 | 茯苓 | 厚朴 | 柴胡 | 羌活 | 琥珀 | 薄荷 |
| CSxTS | 茯苓 | 甘草 | 當歸 | 一兩 | 柴胡 | 茱萸 | 檳榔 | 桔梗 | 膀胱 | 爲末 |
| CSxSL | 甘草 | 茯苓 | 當歸 | 一兩 | 麥門冬 | 柴胡 | 茱萸 | 檳榔 | 吳茱萸 | 各一錢 |
| TSxSL | 一兩 | 甘草 | 爲末 | 當歸 | 各一 | 一錢 | 各一錢 | 半兩 | 茯苓 | 每服 |

문자열에 순위가 부여되면 주어진 텍스트에서 순위가 높은 문자열부터 단어로 추출하게 된다. 추출된 결과와 본래 텍스트의 표점 부분을 비교하였을 때 표점 부분에서 텍스트가 분절되지 않은 부분, 즉 오류가 일어난 경우를 관찰하였다. 본 연구에 사용한 《동의보감》 텍스트의 표점은 219,475건인데, 방법에 따라 대략 2만 7천건에서 2만 9천건 정도의 오류를 나타냈다. 이를 정리하면 아래와 같다. (<Table. 2A, 2B> 참조)

여기서 위배 비율(MR:miss rate)은 텍스트 분절 결과와 표점이 서로 위배된 비율을 의미한다. 위배 비율이 낮을수록 오류가 적다고 해석할 수 있다. 이를 보면, TSxSL, BExTS, SL, BExSL, BExCS, CSxSL, TS, CSxTS, BE, CS 순으로 높아졌다. TS와 SL를 조합하여 곱한 값을 사용하였을 때 오류가 가장 적었고, CS를 단독으로 사용하였을 때 오류가 가장 컸다.

분절된 단어의 수를 가중치로 반영하면 그 결과는 조금 달랐다. 분절이 많아질수록 오류는 줄어들게 된다. 따라서 분절 개수를 가중치로 주어 이를 보정하였다. (MR)과 (S)를 곱한 결과값은 SL, CSxSL, BExSL, TSxSL, BExTS, BExCS, CSxTS, TS, CS, BE의 순으로 높아졌다. 이를 기준으로 하면, SL을 사용하였을 때 결과가 가장 좋았고, BE를 단독으로 사용하였을 때 결과가 가장 좋지 않았다.

**<Table. 2A> Miss rates when using single measure**

| method | miss (M) | miss rate (MR) | sensitivity | segments (S) | (MR) x (S) |
|---|---|---|---|---|---|
| BE | 29,545 | 0.134617 | 0.865383 | 456039 | 61390.4648 |
| CS | 29,588 | 0.134813 | 0.865187 | 444169 | 59879.5871 |
| TS | 28,440 | 0.129582 | 0.870418 | 459481 | 59540.4472 |
| SL | 28,147 | 0.128247 | 0.871753 | 443893 | 56927.9247 |

**<Table. 2B> Miss rates when using combination of two measures**

| method | miss (M) | miss rate (MR) | sensitivity | segments (S) | (MR) x (S) |
|---|---|---|---|---|---|
| BExSL | 28,178 | 0.128388 | 0.871612 | 446284 | 57297.5991 |
| BExTS | 27,870 | 0.126985 | 0.873015 | 455145 | 57796.5196 |
| BExCS | 28,262 | 0.128771 | 0.871229 | 452185 | 58228.2833 |
| CSxTS | 28,633 | 0.130461 | 0.869539 | 450031 | 58711.6420 |
| CSxSL | 28,352 | 0.129181 | 0.870819 | 443430 | 57282.7309 |
| TSxSL | 27,775 | 0.126552 | 0.873448 | 455425 | 57634.9442 |

위배 비율(MR)을 기준으로 오류가 가장 적었던 TSxSL와 가장 많았던 CS, 가중치를 적용한 뒤에 성능이 가장 우수했던 SL과 가장 저조했던 BE의 텍스트 분절 결과 가운데 특정 부분을 비교하면 다음과 같다. ( <Table. 3> 참조. 단어 구분은 공백문자, 【】 안은 표점과 위배된 부분. 밑줄은 단어 추출 오류 일부. )

**<Table. 3> Example of word extraction result**

| 구분 | 원문 |
|---|---|
| 원문 | 補中益氣湯 |
| TSxSL | 補中益氣 湯 |
| CS | 補中益氣湯 |
| SL | 補中益氣湯 |
| BE | 補中 益氣 湯 |
| 원문 | 治勞役太甚。或飮食失節。身熱而煩。自汗倦怠。黃芪一錢半。人參。白朮。甘草各一錢。當歸身。陳皮各五分。升麻。柴胡各三分。右剉。作一貼。水煎服。東垣。 |
| TSxSL | 治 勞役 太甚 或 飮食 失節 身熱 而煩 自汗 倦怠 黃芪 一錢 半 人參 白朮 甘草 各一 錢 當歸 身 陳皮 各 五分 升麻 柴胡 各 三分 【右剉作一貼】 水 煎服 東垣 |
| CS | 治 勞役 太甚 或 飮食 失節 身熱 而煩 自汗 倦怠 黃芪 一錢 半 人參 白朮 甘草 各一錢 當歸 身 陳皮 各 五分 升麻 柴胡 各 三分 右 【剉作一貼】 水 煎服 東垣 |
| SL | 治 勞役 太甚 或 飮食 失節 身熱 而煩 自汗 倦怠 黃芪 一錢 半 人參 白朮 甘草 各一錢 當歸 身 陳皮 各 五分 升麻 柴胡 各 三分 【右剉作一貼】 水 煎服 東垣 |
| BE | 治 勞役 太甚 或 飮食 失節 身熱 而煩 自汗 倦怠 黃芪 一錢 半 人參 白朮 甘草 各一錢 當歸 身 陳皮 各 五分 升麻 柴胡 各 三分 右剉 作 一貼 水 煎服 東垣 |
| 원문 | 一方。黃芪一錢半。人參。白朮。陳皮。當歸。甘草各一錢。升麻。柴胡各五分。加黃柏三分。以滋腎水。紅花二分。入心養血。醫鑑。 |
| TSxSL | 一方 黃芪 一錢 半 人參 白朮 陳皮 當歸 甘草 各一 錢 升麻 柴胡 各 五分 加 黃柏 三分 以 滋腎 水 紅花 二分 入心 養血 醫鑑 |
| CS | 一方 黃芪 一錢 半 人參 白朮 陳皮 當歸 甘草 各一錢 升麻 柴胡 各 五分 加 黃柏 三分 以 滋腎 水 紅花 二分 入心 養血 醫鑑 |
| SL | 一方 黃芪 一錢 半 人參 白朮 陳皮 當歸 甘草 各一錢 升麻 柴胡 各 五分 加 黃柏 三分 以 滋腎 水 紅花 二分 入心 養血 醫鑑 |
| BE | 一方 黃芪 一錢 半 人參 白朮 陳皮 當歸 甘草 各一錢 升麻 柴胡 各 五分 加 黃柏 三分 以滋 腎 水 紅花 二分 入心 養血 醫鑑 |
| 원문 | 夫脾胃一虛。肺氣先絶。故用黃芪。以益皮毛而閉腠理。不令自汗。上喘氣短。損其元氣。用人參以補之。心火承脾。用炙甘草之甘溫。以瀉火熱而補胃中元氣。若脾胃急痛。腹中急縮者。宜多用之。此三味。除濕熱煩熱之聖藥也。白朮苦甘溫。除胃中熱。利腰臍間血。升麻。柴胡苦平味之薄者。升胃中之淸氣。又引黃芪。甘草甘溫之氣味上升。能補衛氣之散解而實其表。又緩帶脈之縮急。當歸以和血脈。橘紅以理胸中之氣。助陽氣上升以散滯氣。此立方本旨也。 |
| TSxSL | 夫 脾胃 一虛 肺氣 先絶 故用 黃芪 以益 皮毛 而閉 腠理 不令 自汗 上喘 氣短 損其 元氣 用 人參 以 補之 心火 承脾 用炙 甘草 之 甘溫 以 瀉火 熱而 補 胃中 元氣 若 脾胃 急痛 腹中 急縮 【者宜】 多 用之 此 三味 除 濕熱 煩熱 之聖藥 也 白朮 苦 甘溫 除 胃中 【熱利】 腰臍 間血 升麻 柴胡 苦 平味 之薄者 升 胃中 之 淸氣 又引 黃芪 甘草 甘溫 之 氣味 上升 能補 衛氣 之散解 而實 其表 又緩帶 脈之 縮急 當歸 以和 血脈 橘紅 以理 胸中 之氣 助 陽氣 上升 以散 滯氣 此立 方本 旨也 |
| CS | 夫 脾胃 一虛 肺氣 先絶 故用 黃芪 以益 皮毛 而閉 腠理 不令 自汗 上喘 氣短 損其 元氣 用 人參 以 補之 心火 承脾 用炙 甘草 之 甘溫 以 瀉火 熱而 補 胃中 元氣 若 脾胃 急痛 腹中 急縮 |

| | |
|---|---|
| | 【者宜】 多 用之 此 三味 除 濕熱 煩熱 之 聖藥也 白朮 苦 甘溫 除胃中熱 利 腰臍 間血 升麻 柴胡 苦 平味 之 薄者 升 胃中 之 淸氣 又引 黃芪 甘草 甘溫 之 氣味 上升 能補 衛氣 之 散解 而實 其表 又緩帶 脈之 縮急 當歸 以和 血脈 橘紅 以理 胸中 之氣 助 陽氣 上升 以散 滯氣 此 立方 本 旨也<br>凡 脾胃 不足 之證 須用 升麻 柴胡 引 脾胃 中 淸氣 行於陽 |
| SL | 夫 脾胃 一虛 肺氣 先絶 故用 黃芪 以益 皮毛 而閉 腠理 不令 自汗 上喘 氣短 損其 元氣 用 人參 以 補之 心火 承脾 用炙 甘草 之 甘溫 以 瀉火 熱而 補 胃中 元氣 若 脾胃 急痛 腹中 急縮 【者宜】 多 用之 此 三味 除 濕熱 煩熱 之聖藥 也 白朮 苦 甘溫 除 胃中 【熱利】 腰臍 間血 升麻 柴胡 苦 平味 之薄者 升 胃中 之 淸氣 又引 黃芪 甘草 甘溫 之 氣味 上升 能補 衛氣 之 散解 而實 其表 又緩 帶脈 之縮急 當歸 以和 血脈 橘紅 以理 胸中 之氣 助 陽氣 上升 以散 滯氣 此 立方 本 旨也 |
| BE | 夫 脾胃 一虛 肺氣 先絶 故用 黃芪 以益 皮毛 而閉 腠理 不令 自汗 上喘 氣短 損其 元氣 用 人參 以 補之 心火 承脾 用炙 甘草 之 甘溫 以瀉 火 熱而 補 胃中 元氣 若 脾胃 急痛 腹中 急縮 【者宜】 多 用之 此 三味 除 濕熱 煩熱 之聖 藥也 白朮 苦 甘溫 除 胃中 【熱利】 腰臍 間血 升麻 柴胡 苦平 味之 薄者 升 胃中 之淸 【氣又】 引 黃芪 甘草 甘溫 之氣 味 上升 能補 衛 氣之 散解 而實 其表 又緩帶 脈之 縮急 當歸 以和 血脈 橘紅 以理 胸中 之氣 助 陽氣 上升 以散 滯氣 此 立方 本 旨也 |

## 5. 고찰

이상의 결과에서 검토해야 할 점은, 용어집이 구비되지 않은 경우 텍스트 분석에서 비지도학습 기반의 단어 추출 방법을 적용할 수 있는가, 있다면 어떤 방법을 사용해야 하는가, 마지막으로 이 방법에서 극복되어야 할 점은 무엇인가 하는 점이다.

먼저 한의학 고문헌 텍스트 분석에서 비지도 학습 기반의 단어 추출 방법을 적용할 수 있을까. <Table. 3>의 결과 예시를 보면, 측정 방법 사이에 결과의 차이가 있고 얼마간의 오류가 있음에도 불구하고 주요한 한의학 용어들을 구분해주고 있다는 점을 확인할 수 있다. 측정 방법에 이론적인 차이가 있으나, 거시적으로 본다면 이 방법들은 모두 데이터 안에서 반복적으로 나타나는 문자열에 높은 점수를 부여하기 때문이다. 전문 용어라는 것은 반복적으로 사용되는 문자의 조합인 경우가 많다. 따라서 이들이 상위에 랭크되어 단어로 추출되는 결과로 이어진다. <Table. 1>에서도 이를 확인할 수 있다. 그러므로 용어집이 구비되지 않은 상태에서 텍스트 분석을 수행해야 할 때, 어느 정도의 오류가 용인되거나 후처리를 할 수 있는 상황이라면 본 연구에서 수행한 비지도학습 기반의 단어 추출 방법을 적용할 수 있을 것으로 보인다.

다만 이 방법들이 분명한 한계를 가지고 있다는 사실 역시 고려되어야 한다. 위배 비율(MR)을 기준으로 보았을 때, 값이 가장 낮은 TSxSL 방법을 사용한 경우에도 이를 민감도(Sensitivity)로 표현하면 0.873에 그치고 있다. 따라서 정밀한 분석을 필요로 하는 경우에는 이 방법을 적용하기 전에 충분한 검토가 필요하다.

그렇다면 어떤 방법을 사용해야 할까. 표점 위배 비율(MR)에서 나타나듯 각각의 방법들 사이에 현저한 차이가 있다고 보기 어렵다. 그러나 상대적으로 비교하였을 때, 위배 비율(MR) 면에서 TS와 SL을 조합한 방법이 가장 우수했고, 가중치를 추가한 결과에서는 SL이 가장 좋은 성능을 나타냈다. 그러므로 SL 혹은 TSxSL을 가장 먼저 고려해 볼 수 있다.

마지막으로 본 연구에서 수행한 비지도학습 기반의 단어 추출 방법이 가지는 몇 가지 문제점에 대해 검토해 보기로 하자. 첫 번째는 문자열의 길이가 짧을수록 더 높은 점수를 부여받기 쉽다는 점이다. <Table 3>을 보면, 모든 방법에서 “當歸身”을 “當歸 身”으로 분절하였다는 점을 확인할 수 있다. 이는 “當歸身”과 “當歸”를 비교하였을 때, 후자가 단어로서 더 높은 점수를 획득하였기 때문이다. 이러한 문제는 “吳茱萸”를 “吳 茱萸”로 분절하거나 “炙甘草”를 “炙 甘草”로 분절하는 예와 같이 긴 문자열

과 짧은 문자열이 공존할 때 짧은 문자열의 우선순위가 더 높게 나타나기 쉬우므로 발생하는 현상이다. 일반적인 경우 4글자보다는 3글자가, 3글자보다는 2글자의 출현 빈도가 더 높을 수밖에 없기 때문이다. 만약 "當歸身"이 "當歸"보다 더 높은 점수를 부여받도록 한다면, 더 희소하게 등장하는 문자열에 더 높은 점수가 부여되는 모순이 발생하게 된다. 이러한 한계로 인해 텍스트 분절은 2글자 단위로 이루어지는 경향이 발생한다. <Table 3>의 결과에서 그러한 경향을 확인할 수 있다.

두 번째는 문맥에 따라 단어 여부를 탄력적으로 검토하지 못한다는 점이다. "補中益氣湯"이라는 문자열의 경우, "補中益氣湯"(CS), "補中_益氣_湯"(BE), "補中益氣_湯"(TSxSL) 등으로 다양하게 나타났다. 위의 예의 경우에는 "補中益氣湯"을 하나의 단어로 추출하는 것이 옳다고 여겨진다. 그러나 문맥에 따라 "補中"과 "益氣" 역시 단어가 될 수 있다. 이런 경우를 상정하면 "補中_益氣_湯"이나 "補中益氣_湯"과 같은 결과도 무조건 틀렸다고 볼 수 없다. 처방 이름을 의미할 때는 "補中益氣湯"으로 추출되고, 효능을 의미할 때는 "補中_益氣"나 "補中益氣"로 추출되는 것이 이상적일 것이다. 그러나 문자열에 추출 점수를 부여하고 추출점수가 높은 문자열부터 추출하여 텍스트를 분절하는 현재의 방법으로는 문맥을 고려할 수 없으므로 이러한 결과를 기대할 수 없다.

이러한 한계는 단어 추출 방식을 변경함으로써 어느 정도 극복할 수 있다. 텍스트를 분절할 때, 추출점수가 높은 문자열부터 단어로 추출하는 것이 아니라 길이가 긴 문자열부터 단어로 추출하는 방식을 이용하는 것이다. 용어집에 ["補中益氣湯", "補中益氣", "補中", "益氣"]가 모두 있다면, 위의 예시에서는 "補中益氣湯"이 추출되고, 본문 가운데 효능 설명에서는 "補中"이나 "益氣"가 추출될 것이다. 이는 "當歸身"과 "當歸", "吳茱萸"와 "茱萸"의 문제도 해결해 줄 수 있다. 그러나 이러한 경우에는 일정 점수를 기준으로 단어로 인정할 문자열과 그렇지 않은 문자열을 구분해 주어야 한다는 과제가 남겨진다. 이 역시 간단한 문제가 아니며, 이 기준을 잘못 정하면 상당한 오류를 야기할 수 있다. 이에 대해서는 추가적인 고찰이 필요하다.

## Ⅲ. 결론

텍스트 형태의 비정형 데이터를 분석하기 위해서는 텍스트에서 분석의 최소 단위가 되는 토큰을 추출해 주어야 한다. 토큰은 일반적으로 해당 언어의 단어를 기준으로 한다. 그러므로 분석에 앞서 텍스트에서 단어를 추출하기 위한 방법이 마련되어야 한다. 단어 추출을 위해 가장 확실한 방법은 용어집을 사용하는 것이지만, 용어집 구축에는 많은 자원이 소요된다. 이런 경우 비지도학습 기반의 용어 추출 방법을 고려할 수 있다.

본 연구에서는 한의학 고문헌에 비지도학습 기반의 몇 가지 용어 추출 방법을 적용하고 그 성능을 비교해 보았다. 이를 위해 학습용 데이터로 한국한의학연구원에서 구축한 17,842,632자 규모의 한의학 고문헌 텍스트를 사용하였고, 실험용 데이터로는 870,540자 규모의 《동의보감》 텍스트 데이터를 사용하였다. 결과 검증을 위해서는 219,475건의 《동의보감》 표점 정보를 이용하였다. 본 연구에 순위 부여에 필요한 점수를 계산하기 위해 외부경계값을 이용한 방법 1가지(BE:Branching Entropy)와 내부경계값을 이용한 방법 3가지(CS:cohesion score, TS:t-score, SL:simple-ll), 그리고 이들을 조합한 방법 6가지 방법(BExSL, BExTS, BExCS, CSxTS, CSxSL, TSxSL)으로 테스트를 진행하였다.

그 결과 표점이 서로 위배된 비율(MR)을 기준으로 한 경우, TS와 SL를 조합하여 곱한 값을 사용하였을 때 오류가 가장 적었고, 분절된 텍스트의 수를 가중치로 적용한 기준을 사용한 경우, SL을 사용하였을 때 결과가 가장 좋았다.

본 연구는 평가용 데이터를 확보할 수 없는 상태에서 텍스트가 분절된 지점이 《동의보감》 표점 부분과 위배되는지만 검토하였다. 그러므로 분절되지 않아야 할 곳에서 분절되었는지에 대해서는 적절하게 평가할 수 없다는 한계를 지닌다. 그러나 한의학 고문헌 텍스트 분석을 위해 비지도학습 기반 단어의 추출 방법을 고찰하였다는 점에서 의의를 찾을 수 있다.

## 감사의 말씀

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

## References

1. Huang Yongnian. Introduction of Ancient books Arrangement. Institute for the Translation of Korean Classics. 2018. 2013. p.209.
2. Hyun-joong Kim, Sungzoon Cho, Pilsung Kang. KR-WordRank : An Unsupervised Korean Word Extraction Method Based on WordRank. Journal of the Korean Institute of Industrial Engineers. 2014. 40(1). pp.18-33.
3. 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. Springer. 2008. pp 52-63.
4. Zhihui Jin, Kumiko Tanaka-Ishii. Unsupervised Segmentation of Chinese Text by Use of Branching Entropy. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. 2006. pp.428-435.
5. 김현중, 조성준, 강필성. KR-WordRank : WordRank를 개선한 비지도학습 기반 한국어 단어 추출 방법. 대한산업공학회지. 2014. 40(1). pp.18-33.
6. 黃永年(김언종, 김수경 옮김). 고적정리개론. 한국고전번역원. 2018. 2013. p.209.
7. Chinese Medical Database. Beijing. Hunan Electronic Audio and Video Publishing House. 2003.
8. Hyunjoong Kim. LOVITxDATA SCIENCE. [cited on July 17, 2019]. Avaiable from: https://lovit.github.io/nlp/2018/04/09/cohesion_ltokenizer
9. Korea Institute of Oriental Medicine. Mediclassics. [cited on Jan 12, 2019]. Avaiable from: https://mediclassics.kr
10. 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
11. 한국한의학연구원. 한의학고전DB. [cited on July 17, 2019]. Avaiable from: https://mediclassics.kr