대한한의학원전학회지
2020년 5월 33권 2호

pISSN 1229-8328 eISSN 2713-850X
J Korean Med Classics 2020:33(2) 051-059
https://doi.org/10.14369/jkmc.2020.33.2.051

JKMC
Journal of Korean Medical Classics

한의학 고문헌 텍스트에서의 저자 판별
- 기능어의 역할을 중심으로 -

한국한의학연구원 책임연구원
오준호*

A Comparative Study of Feature Extraction Methods for Authorship Attribution in the Text of Traditional East Asian Medicine with a Focus on Function Words

Oh Junho*
Senior Researcher at Korea Institute of Oriental Medicine

Objectives : We would like to study what is the most appropriate "feature" to effectively perform authorship attribution of the text of Traditional East Asian Medicine

Methods : The authorship attribution performance of the Support Vector Machine (SVM) was compared by cross validation, depending on whether the function words or content words, single word or collocations, and IDF weights were applied or not, using 'Variorum of the Nanjing' as an experimental Corpus.

Results : When using the combination of 'function words/uni-bigram/TF', the performance was best with accuracy of 0.732, and the combination of 'content words/unigram/TFIDF' showed the lowest accuracy of 0.351.

Conclusions : This shows the following facts from the authorship attribution of the text of East Asian traditional medicine. First, function words play an important role in comparison to content words. Second, collocations was relatively important in content words, but single words have more important meanings in function words. Third, unlike general text analysis, IDF weighting resulted in worse performance.

Key words : authorship attribution, Function words, Korean Medical Classics, East Asian traditional medicine. Variorum of the Nanjing.

* 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(April 17, 2020), Revised(May 11, 2020), Accepted(May 11, 2020)

Copyright © The Society of Korean Medical Classics. All rights reserved.
This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ⅰ. 서론

저자 판별(authorship attribution)[1]은 언어로 이루어진 데이터에서 작자의 특징을 유추하는 작업이다.[2] 초기의 저자 판별은 저자가 불분명한 텍스트의 저자를 밝히는 문제에서 시작되었다. 과거에는 '자세히 읽기(close reading)'를 통해 저자를 판별하는 전통적인 방법이 주로 사용되었으나 현재는 전산언어학 혹은 정보학을 이용하여 문제를 해결하는 계량적인 방법이 연구되고 있다.[3]

저자 판별은 "개인이 의식적 혹은 무의식적으로 사용하는 언어 습관이 텍스트에 고유하게 나타난다"[4]는 기본 전제 위에 성립한다. 즉, "모든 화자는 개인어라고 할 수 있는 자기 자신만의 고유한 언어적 형태를 가지고 있다"[5]는 것이다. 이를 계량적으로 판별하기 위해서는 텍스트의 저자가 다른 저자와 구분되는 "안정적이며 드러나는 특징"을 가지며, 이를 계량할 수 있어야 한다.[6] 이러한 특징을 '문체(style)'라고 할 수 있다.

한의학 고문헌 텍스트는 동아시아 전통사회에서 한문(漢文)으로 작성된 글로, 대부분 의서(醫書) 형태로 존재한다. 한의학 고문헌 텍스트 역시 몇 가지 문제에서 문체의 차이를 분석해야 하는 경우가 존재한다. 먼저, 책의 저자를 밝히는 문제이다. 책에 저자가 표시되어 있지 않거나 표시되어 있다고 하더라도 본명을 밝히지 않은 경우는 저자가 누구인지 확정하기 어렵다. 『보유신편(保幼新編)』은 최근까지 성명 미상의 '무기선생(無忌先生)'의 저작으로 알려져 있었다. 그러나 새로운 필사본 사료가 발견되면서 『주촌신방(舟村新方)』을 저술한 신만(申曼)이라는 주장이 제기되었다.[7] 또 전통사회에서는 책의 권위를 세우기 위해 대중에게 잘 알려진 인물을 저자로 세우는 일이 많았다. 이때 책에서 밝힌 저자가 실제 저자와 부합하는지 확인하는 문제가 제기된다. 정약용(丁若鏞)을 전면에 내세운 『정다산선생소아과비방(丁茶山先生小兒科秘方)』이 유사한 예이다. 이 책은 최규헌(崔奎憲)의 『소아의방(小兒醫方)』의 다른 판본이지만 '정다산'을 표제로 내세워 오해를 불러일으켰다.[8] 다행스럽게도 책 내용을 보면 최규헌의 저작임을 알 수 있지만, 명시적인 단서가 없다면 실제 저자가 누구인지 판별해야 하는 문제가 발생한다. 이처럼 한의학 고문헌 텍스트에도 저자가 누구인지에 대한 문제가 상존한다.

다음으로 같은 책 내에 존재하는 여러 문서를 저자나 성립 시대로 구분하는 문제이다. 한의학 고문헌 텍스트 중에는 서로 다른 저자에 의해 저술된 글들이 하나의 책으로 묶인 경우가 있다. 이때 각 편의 저자나 성립 시대를 구분하는 문제가 제기된다. 『영추(靈樞)』와 『소문(素問)』 각 편의 선후관계를 규명하는 문제, 『상한론』 텍스트의 구분 문제[9] 등이 이에 해당한다. 이와 유사한 문제로 중국 고대 소설 『홍루몽(紅樓夢)』의 저자 판별에 대한 연구가 있다.[10]

---

1) 저자 확정(authorship identification)이라고 하기도 한다.
2) Patrick Juola. Authorship Attribution. Foundations and Trends in Information Retrieval. 2006. 1(3). pp.233-334.
3) 최지명. 기계학습 알고리즘을 이용한 한국어 텍스트 저자 판별. 석사학위논문(연세대). 2015. pp.5-6.
4) 최지명. 앞의 논문. pp.5-6.
5) 최지명. 앞의 논문. pp.5-6. 재인용.
6) 강남준, 이종영, 최운호. 『독립신문』 논설의 형태 주석 말뭉치를 활용한 논설 저자 판별 연구 - 어미 사용빈도 분석을 중심으로. 한국사전학. 2010. 15. pp.73-101.
7) 양승률. 주촌 신만의 『보유신편(保幼新編)』 편찬과 『주촌신방(舟村新方)』. 장서각. 2011. 25. pp.52-77.
8) 이가은, 안상우. 소아의방(小兒醫方)의 판본비교(板本比較) 및 편제(篇第) 고찰(考察). 한국의사학회지. 2004. 17(1). pp.163-176.
9) 박경모, 최승훈. 『강평 (康平) 상한론 (傷寒論)』 의 고증을 통한 『상한론(傷寒論)』 과 『황제내경(黃帝內經)』 의 비교연구. 대한한의학원전학회지. 1995. 9. pp.265-301.
10) 『홍루몽(紅樓夢)』의 저자 판별에 대해서는 아래와 같은 다양한 연구가 진행된 바 있다.
Bing-Cho Chan. The authorship of the Dream of the red chamber based on a computerized statistical study of its vocabulary. Joint Publishing Co Ltd., Hong Kong. 1986.
Qing-Xiang Yu. Applications of Statistical methods to Dream of the Red Chamber. Journal of National Cheng-Chi University. 1998. 76. pp.303-327.
Hsieh-Chang Tu, Jieh Hsiang. A Text-Mining Approach to the Authorship Attribution Problem of Dream of the Red Chamber. Digital Humanities. 2013. pp.441-443
Hu, Xianfeng, Yang Wang and Qiang Wu. Multiple authors Detection: a Quantitative Analysis of Dream

www.kci.go.kr

한의학 텍스트에서 제기되는 이상의 문제들은 '자세히 읽기'와 같은 전통적인 방법으로 연구되어 왔다. 그러나 텍스트 안팎에 명시적인 증거가 없거나, 있더라도 완전히 신뢰하기 어려운 경우가 있다. 이 때 계량적인 방법을 적용하여 가능성을 검토해 볼 수 있다. 또 전통적인 방법으로 도출된 연구 결과를 지지하거나 비판하여 보다 진실에 가까운 답을 얻을 수 있다. 이를 위해서는 먼저 한의학 텍스트를 대상으로 한 계량적 저자 판별 방법이 수행 가능한 것인지, 가능하다면 적합한 방법이 무엇인지 검토하는 과정이 필요하다. 이에 본고에서는 이러한 방법에 대해 검토하고자 한다.

Ⅱ 본론

1. 연구방법

저자 판별을 위한 계량적인 방법들은 매우 다양하다. 일반적으로 자연어처리(NLP ; Natural Language Process), 기계학습(ML ; Machine Learning)에 사용되는 방법이 적용될 수 있다. 이러한 방법 가운데 어떤 방법이 가장 효과적인지에 대해서는 여전히 연구가 계속 진행 중이다.

계량적인 방법으로 저자 판별을 수행할 때 핵심적인 문제는 '텍스트의 문체를 어떤 특성(feature)으로 나타낼 것인가'와 '텍스트 사이의 관계를 어떤 분류 방법(classification methods)으로 구분할 것인가'를 결정하는 일이다. 효과적인 저자 판별 방법을 모색한 선행연구 2가지를 살펴보자. 터키 신문을 대상으로 한 저자 판별 연구에서는 '기능어를 제외한 단어'를 특성(feature)으로 하고 '서포트 벡터 머신(SVM ; Support Vector Machine)'을 분류 방법으로 한 경우, 그리고 '기능어(FW ; Function Word)'를 특성(feature)으로 하고 '베이지안 분류기(Bayesian Classifier)'와 같은 가우시안 모델을 사용한 경우에 가장 좋은 성능을 보였다.[11] 또 영어로 이루어진 C50 코퍼스와 엔론(Enron) 코퍼스를 대상으로 한 연구에서는 '유니그램(unigram)'을 특성(feature)으로 하고 SVM을 분류 방법으로 한 경우가 좋은 성능을 보였다.[12] 이런 연구들은 텍스트에 사용된 언어, 텍스트의 형식, 텍스트가 다루고 있는 주제 등에 따라 결과를 조금씩 달리한다. 한의학 텍스트는 한자로 이루어져 있고 한의학과 관련된 주제를 다루고 있으므로 이러한 선행연구를 참고할 수는 있지만 반드시 일치하는 결과가 나타난다고 볼 수 없다.

저자의 문체를 나타내는 특성(feature)은 다양할 수 있지만, 기본은 텍스트에 사용된 어휘이다. 어휘는 사용된 의미에 따라 구체적인 내용을 전달하기 위한 '내용어(content words)'와 문법적인 기능이나 분위기를 전달하기 위한 '기능어(function words)'로 구분할 수 있다. 또 어휘를 구분하는 방법에 있어서 단어 하나하나를 독립적으로 보는 '단어(single word)' 방식과 단어 사이의 영향을 고려한 '연어(collocation)' 방식으로 나눌 수 있다.

본 연구에서는 실험을 통해 저자 판별에 가장 효과적인 방법을 검토해 보고자 한다. 실험에 사용할 텍스트를 선택한 뒤 저자 판별에 비교적 널리 사용되는 SVM을 분류 방법으로 고정하고[13], 내용어와 기능어, 단어와 연어가 분류 결과에 어떤 영향을 미치는지 검토해 보았다. 아울러 텍스트를 수학적으로 공간에 표시할 때, 단순빈도(TF ; Term Frequency)를 사용하는 경우와 여기에 역문서빈도(IDF ; Inverse Document Frequency)를 가중치로 적용한 TFIDF (Term Frequency - Inverse Document Frequency)

---

of the Red Chamber. Advances in Adaptive Data Analysis. 2014. 6. pp.1-19.

11) İlker Nadi Bozkurt, Özgür Bağlıoğlu, Erkan Uyar. Authorship attribution: performance of various features and classification methods. 22nd International Symposium on Computer and Information Sciences, ISCIS 2007. IEEE. 2007. pp.158-162.

12) Smita Nirkhi, R.V.Dharaskar, V.M.Thakare. Authorship Identification using Generalized Features and Analysis of Computational Method. Transactions on Machine Learning and Artificial Intelligence. 2015. 3(2). pp.41-45.

13) Matthew L. Jockers, Daniela M. Witten. A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing. 2010. 25(2). pp.2105-223.

www.kci.go.kr

를 사용한 경우의 차이도 함께 살펴보았다.

1.1. 데이터 선정

실험에서 사용할 코퍼스(Corpus)를 선택하기 위해 다음 2가지 기준을 고려하였다. 첫째, 코퍼스를 이루는 문서들(Documents)이 서로 다른 몇 명의 저자에 의해 작성되어 있을 것. 저자 판별을 시험하기 위해 여러 저자의 글이 포함되어 있어야하기 때문이다. 둘째, 코퍼스에 속한 문서들은 유사하거나 동일한 주제에 대해 다루고 있을 것. 서로 상이한 주제를 다루고 있을 경우, 문서에 나타나는 특성(feature)이 저자가 달라 생겨난 것인지 주제가 달라 야기된 것인지 구분하기 어려울 수 있다. 이러한 변수를 줄이기 위해 가급적 동일한 주제에 대해 쓴 글을 찾고자 했다.

이 기준에 따라 실험 코퍼스로 『난경집주(難經集註)』를 선택하였다. 『난경집주』는 『난경(難經)』에 대한 여러 주석가들의 주석을 모아 보기 편하게 정리한 책이다. 이 책 「발문(跋文)」에 따르면, "명나라 왕구사(王九思) 등이 오나라의 여광(呂廣), 당나라의 양현조(楊玄操), 송나라의 정덕용(丁德用)·우서(虞庶)·양강후(楊康侯)의 주해(注解)를 수집하여 기록"[14]하였다고 한다. 또한 『난경집주』는 『난경』이라는 동일한 주제에 대해 서로 다른 주석가가 작성한 텍스트이다. 따라서 주제 차이에서 올 수 있는 오류를 최소화할 수 있다.

『난경』은 81난(難)으로 이루어져 있으므로, 각각을 하나의 장(章)으로 볼 수 있다. 각각의 장 속에는 다시 난경 본문과 난경 본문에 대한 주석들이 혼합되어 있다. 분석에 앞서 각 텍스트를 원문(이하 약어 사용, 'O'), 여광의 주석('L'), 양현조 혹은 양강후의 주석('Y')[15], 정덕용의 주석('Z'), 우서의 주석('W')으로 나누었다. 이렇게 원문을 포함하여 저자 레이블(label)이 붙은 81장 × 5종의 '문서'로 이루어진 분석 대상 코퍼스를 준비하였다. 다만 문서의 크기가 너무 작은 경우 분석하기 어렵다. 한자 텍스트의 경우 4-5자로 하나의 구(句)가 이루어지며, 다시 4-5자의 구가 모여 의미를 전달한다고 보고 20자 미만의 문서는 분석에서 제외하였다.

『난경집주』의 전자 텍스트는, 한국한의학연구원에서 상해(上海) 함분루(涵芬樓)에서 영인한 일존총서본(佚存叢書本)을 저본으로 하여 한의학고전DB(mediclassics.k)를 통해 공개하고 있는 데이터[16]를 사용하였다.

1.2. 분석 설계

정량적 분석을 하기 위해서는 코퍼스를 연산이 가능하도록 정량적인 형태로 재구성해야 한다. 이를 텍스트 임베딩(embedding)이라고 하며, 일반적으로 텍스트를 그 속에 나타난 '특성(feature)'을 기준으로 정량화하게 된다. '특성(feature)'은 분석에 따라 다양하게 정의될 수 있으나 일반적으로 의미의 최소 단위라고 할 수 있는 어절 혹은 단어를 용어로 삼는다.[17] 예를 들어 문서에 등장하는 단어를 특성(feature)으로 본다면, 임베딩 결과는 문서에 사용된 단어의 빈도가 기반이 된다.

텍스트 임베딩 방식에는 카운트 기반 방법(Counting-based word embedding)과 예측 기반 방법(Prediction-based word embedding) 등이 있다.[18] 본 연구에서는 텍스트 분석에서 기본적으로

14) "難經集注五卷, 明王九思等集錄, 吳呂廣、唐楊玄操、宋丁德用、虞庶、楊康侯注解者。" MEDICLASSICS [homepage on the Internet]. Korea Institute of Oriental Medicine; 2015 [cited 30 Jan 2020]. Available from: https://mediclassics.kr/books/149/volume/5#content_148

15) 책에는 "楊曰"과 같이 문장 앞에 누구의 주장인지 밝히는 형식으로 주석자를 표시하고 있다. 그런데, 양현조와 양강후는 모두 "楊曰"로 표시되므로 구분할 수 없었다. 이에 부득이 하나의 문서로 처리하였다.

16) MEDICLASSICS [homepage on the Internet]. Korea Institute of Oriental Medicine; 2015 [cited 30 Jan 2020]. Available from: https://mediclassics.kr/books/149

17) 한자텍스트의 경우 하나의 글자가 하나의 의미를 나타내기 때문에 하나의 글자를 용어로 사용할 수 있다. 이를 유니그램(unigram, 1-gram) 방식이라고 한다. 하나 이상의 용어가 함께 특정한 의미를 나타내는 경우를 고려하기 위해 2가지 글자를 하나의 용어로 적용할 수 있다. 이를 바이그램(bigram, 2-gram) 방식이라고 지칭한다.

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

www.kci.go.kr

사용되는 카운트 기반 방법 가운데 1차 벡터(1st order vector)로 문서를 표시하는 방법을 사용하였다.

본 연구에서는 저자 판별에 영향을 미칠 것으로 추정되는 특성(feature)의 3가지 요소를 검토하였다. 첫째로 기능어와 내용어, 둘째로 단어와 연어, 셋째로 TF방식과 TFIDF방식이 그것이다.

첫째, 기능어와 내용어가 저자의 문체 특징을 얼마나 잘 드러내는지 확인하기 위해 특성(feature) 추출에서 전체 어휘를 대상으로 한 경우(A①), 기능어(function words)를 대상으로 한 경우(A②), 그리고 기능어를 제외한 내용어(content words)만을 대상으로 한 경우(A③) 이렇게 3가지로 나누어 분석을 진행하였다. 기능어는 문법이나 어감을 나타내기 위해 사용되는 용어로서 한자 텍스트에서는 주로 허사(虛辭)를 의미한다. 유니그램 기능어로는 '也', '之', '者', '其', '不', '而', 바이그램 기능어로는 '何以', '假令', '奈何', '者也', '所以' 등이 그 예이다. 본 연구에서는 한문 해석 관련 전문서[19]에서 다루고 있는 928종의 허사를 기능어로 보았다.

둘째, 단어와 연어를 차이를 살펴보기 위해 특성(feature) 추출에서 유니그램(unigram)을 사용한 경우(B①), 바이그램(bigram)을 사용한 경우(B②), 양자를 모두 합친 경우(B③) 이렇게 3가지로 나누어 분석하였다. 유니그램은 보통 단어 하나, 바이그램은 인접한 2가지 단어를 의미한다. 그러나 한문(漢文)의 경우 하나의 글자가 하나의 단어 역할을 수행하므로, 한자(漢字) 한 글자를 유니그램, 인접한 두 글자를 바이그램으로 간주한다. 예를 들어 "十二經皆有動脈"이라는 텍스트가 있을 때, 유니그램은 ['十', '二', '經', '皆', '有', '動', '脈'], 바이그램은 ['十二', '二經', '經皆', '皆有', '有動', '動脈'], 혼합방식은 ['十', '二', '經', '皆', '有', '動', '脈', '十二', '二經', '經皆', '皆有', '有動', '動脈']이 특성(feature)이 된다.

셋째, TF방식과 TFIDF방식의 차이를 살펴보았다. TF는 단순 용어 빈도(TF ; Term Frequency)를 의미한다(C①). 문서 내에 해당 용어가 나타난 빈도 그대로를 사용하는 방식이다.[20] TFIDF는 단순 용어 빈도(TF)에 역문서빈도(IDF ; Inverse Document Frequency)를 곱해준 값을 사용하는 방식이다(C②). 많은 문서에 등장할수록 문서의 특징을 나타내기 어렵다는 경험적 사실에 기반을 둔 방법이다.

자주 사용되는 용어는 높은 빈도로 나타나는 데 반해 대다수의 용어는 1-2회 정도의 낮은 빈도로 출현한다. 따라서 용어의 빈도를 기준으로 문서를 공간상의 벡터로 표현하면 대부분이 0의 값을 가지는 성근(sparse) 형태의 벡터가 된다. 따라서 통상 차원을 축소하여 이러한 문제를 해결한다. 차원 축소의 결과는 문서에 잠재된 특성을 드러내는 데도 의미가 있다. 본고에서는 잠재의미분석(LSA ; Latent Semantic Analysis) 방법을 이용하여 64차원으로 축소하였다.

1.3. 성능 검토

성능 검토를 위해서 5-fold 교차검증(cross validation) 방식을 사용하였다. 전체 코퍼스에 속한 문서를 4:1의 비율로 학습 데이터(training data)와 테스트 데이터(test data)로 나누고 학습과 테스트를 5회 반복하는 방법이다. 학습 데이터로 모델을 학습시키고 이를 테스트 데이터에 적용해 각 회차별로 정확도(accuracy)[21]를 측정한다. 이렇게 되면 5회에 걸쳐 5개의 정확도가 도출되는데, 이를 평균 내어 최종 결과로 삼는다.[22]

2. 분석결과

2.1. 텍스트 기본 통계량

연구에 사용된 실험용 코퍼스는 모두 299개의 문서로 이루어졌다. 『난경집주』에서 20글자 미만으로

19) 김원중. 한문 해석 사전. 글항아리. 2013.

20) 문서의 길이가 길어질수록 특성(feature, 여기서는 글자 혹은 단어)의 TF 값이 더 크게 나타난다. 이를 보정하기 위해 본 연구에서는 단순 빈도 그대로를 TF로 사용하지 않고, $1 + \log(tf)$ 함수를 적용하고 L2 방식으로 정규화(Normalize) 하였다.

21) 정확도(accuracy)는 테스트 데이터에서 예측값(Predicted Labels)과 참값(True Labels)이 일치하는 케이스 개수를 전체 케이스 개수로 나눈 값이다.

22) 본 연구에서는 분석을 위해 python(3.6) 및 python의 데이터 분석 패키지인 scikit-learn(0.20.1)을 사용하였다.

www.kci.go.kr

구성된 문서를 제외한 결과이다. 이 문서들을 저자에 따라 나누어 문서의 개수와 문서를 구성하는 총 글자 수를 살펴보면, O 81종(총 13,024자), L 24종(총 4,177자), Y 59종(총 10,569자), Z 79종(총 13,595자), W 56종(총 14,301자)이었다.

코퍼스 전체에서 1회 이상 나타난 유니그램은 1,480종, 바이그램은 20,075종이었고, 이 가운데 기능어는 유니그램 268종, 바이그램 77종이었다. 저자에 따라 사용된 유니그램, 바이그램, 기능어 가운데 빈도가 높은 10가지 특성(feature)을 살펴보면 <Table 1>과 같다.

Table 1. The 10 Most Frequent Features for Documents Belonging to Corpus. (UG: unigram, BG: bigram, AF: all features, FF: function features, CF: content features. value: feature count)

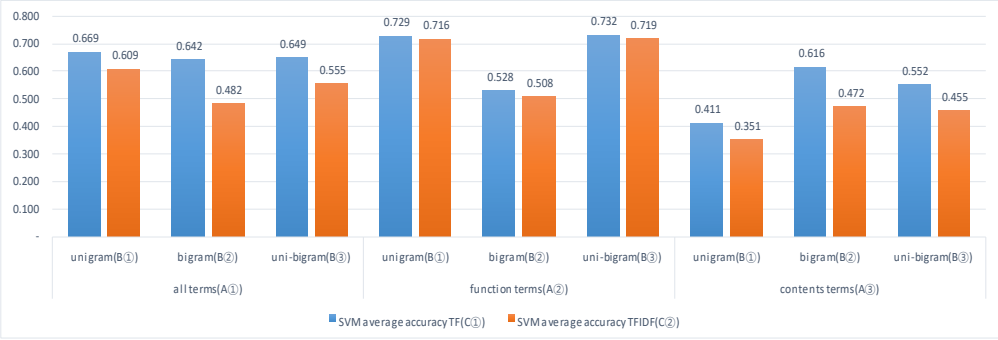
Feature		Author	1		2		3		4		5		6		7		8		9		10	
UG	AF	O	也	422	之	406	者	382	脈	251	其	200	不	198	有	192	而	191	陰	179	陽	176
		L	也	186	脈	159	者	125	故	117	之	104	心	83	其	77	病	67	爲	56	陽	56
		Y	也	481	之	346	者	225	故	202	爲	202	氣	186	脈	169	陽	137	其	130	陰	128
		Z	也	561	之	427	者	359	其	290	陽	250	陰	249	是	234	脈	233	故	221	爲	194
		W	之	524	也	434	脈	292	氣	251	故	226	陰	211	於	202	陽	184	曰	173	爲	163
	FF	O	也	422	之	406	者	382	其	200	不	198	有	192	而	191	爲	174	曰	172	一	143
		L	也	186	者	125	故	117	之	104	其	77	爲	56	一	41	不	40	在	36	言	36
		Y	也	481	之	346	者	225	爲	202	故	202	其	130	而	123	以	118	於	116	不	115
		Z	也	561	之	427	者	359	其	290	是	234	故	221	爲	194	而	162	不	155	於	133
		W	之	524	也	434	故	226	於	202	曰	173	爲	163	其	146	者	143	一	141	不	134
	CF	O	脈	251	陰	179	陽	176	病	161	氣	156	五	134	十	123	經	114	心	97	藏	92
		L	脈	159	心	83	病	67	陽	56	氣	55	主	55	陰	52	肺	42	藏	39	腎	37
		Y	氣	186	脈	169	陽	137	陰	128	病	105	五	104	經	95	藏	89	人	82	心	79
		Z	陽	250	陰	249	脈	233	氣	176	經	146	病	140	五	124	主	121	寸	112	心	98
		W	脈	292	氣	251	陰	211	陽	184	五	161	經	146	病	124	中	104	上	101	心	100
BG	AF	O	難曰	81	也然	67	五藏	41	何謂	40	經言	38	謂也	38	者爲	33	者何	32	何以	29	曰經	28
		L	其脈	26	故言	22	脈也	21	也心	18	故曰	16	故其	14	胃氣	14	之脈	13	也此	13	其氣	12
		Y	故曰	42	五藏	37	脈也	33	也言	31	故云	29	所以	24	之所	22	按之	22	陰陽	20	陽氣	19
		Z	者是	47	也其	46	陰陽	45	五藏	39	者謂	37	也此	34	是也	33	此是	32	所以	31	三陽	25
		W	故曰	62	五藏	37	之脈	36	陰陽	36	太陰	35	起於	28	厥陰	27	少陽	27	三焦	26	少陰	26
	FF	O	何以	29	何也	23	是謂	18	假令	17	奈何	9	者也	9	所以	6	者邪	5	何所	5	更相	5
		L	所以	6	是謂	5	如此	2	非獨	1	是以	1	當復	1	於此	1	所爲	1	也夫	1	以爲	1
		Y	所以	24	如此	13	至於	9	然後	6	以爲	6	假令	6	是謂	3	可以	3	由此	3	如其	3
		Z	所以	31	是謂	13	是故	7	也夫	7	然後	6	是以	5	遞相	5	無所	5	其諸	5	有所	4
		W	如是	12	所以	11	乃如	10	如此	9	因而	8	無所	8	假令	7	於此	6	更相	6	可以	5
	CF	O	難曰	81	也然	67	五藏	41	何謂	40	經言	38	謂也	38	者爲	33	者何	32	陰陽	28	曰經	28
		L	其脈	26	故言	22	脈也	21	也心	18	故曰	16	胃氣	14	故其	14	之脈	13	也此	13	其氣	12
		Y	故曰	42	五藏	37	脈也	33	也言	31	故云	29	之所	22	按之	22	陰陽	20	陽氣	19	十二	18
		Z	者是	47	也其	46	陰陽	45	五藏	39	者謂	37	也此	34	是也	33	此是	32	故曰	25	太陰	25
		W	故曰	62	五藏	37	陰陽	36	之脈	36	太陰	35	起於	28	厥陰	27	少陽	27	少陰	26	三焦	26

www.kci.go.kr

2.2. 저자 판별 교차검증 결과

저자 판별 교차검증 결과는 <Fig. 1>과 같았다. '기능어/유니-바이그램/TF'를 사용한 경우, 정확도 0.732로 가장 좋은 성능을 보였고, 그 다음으로 '기능어/유니그램/TF' 조합에서 0.729의 정확도를 보였다. 반대로 '내용어/유니그램/TFIDF' 조합이 0.351로 가장 낮은 정확도를 보였다.

Fig 1. The Average Accuracy of 5-fold Cross-validation for Each Case.



3. 고찰

먼저 기능어와 내용어가 저자의 문체 특징을 어떻게 드러내지 확인하기 위해 모든 용어를 사용한 경우(A①), 기능어만 사용한 경우(A②), 내용어만 사용한 경우(A③)를 비교해 보자. 전반적으로 기능어만 사용한 경우에 높은 정확도를 보였고 내용어를 사용한 경우에 낮은 정확도를 보였다. 이를 통해 저자 판별에서 기능어가 중요하다는 점을 확인할 수 있었다. 이는 저자 판별에서 기능어가 중요하다는 선행 연구 결과를 지지하는 결과이다.[23)24)] 다만 기능어에서 바이그램만을 적용한 경우 오히려 성능이 낮게 나타났다. <Table 1>에서 보이듯(BG/FF), 바이그램 기능어 각각의 출현 빈도는 매우 낮다. 따라서 이는 기능어 바이그램이 문서의 문체를 충분히 드러내기 어려웠기 때문으로 추측된다.

다음으로 단어와 연어를 차이를 살펴보기 위해 유니그램(unigram)만 사용한 경우(B①), 바이그램(bigram)만 사용한 경우(B②), 양자를 모두 합친 경우(uni-bigram)(B③)를 살펴보자. 이 차이는 다른 변수에 영향을 받았다. 모든 용어를 사용한 경우에는 각각의 성능 차이가 크지 않았다. 그러나 기능어의 경우에는 바이그램을 사용한 경우 성능이 낮았고, 반대로 내용어의 경우 바이그램을 사용한 경우 성능이 높게 나타났다. 저자 판별에서는 기능어가 중요한 역할을 하지만, 내용어로 한정한다면 내용어의 바이그램이 상대적으로 저자의 문체 특성(feature)을 드러낸다고 할 수 있다.

마지막으로 IDF 가중치 적용의 문제(C①, C②)이다. 모든 경우에서 IDF 가중치를 적용하지 않았을 때 더 우수한 성능을 보였다. 따라서 저자 판별에서 IDF 가중치는 큰 의미가 없으며 때에 따라 오히려 기능을 저해할 수 있다.

다만 기능어에서는 TF와 TFIDF의 차이가 가장 적게 나타났다. IDF 가중치는 본래 모든 문서에 나타나는 특성(feature)의 영향력을 줄이기 위해 고안된 방법이다. 내용어와 달리 기능어는 대부분의 문서에 등장하여 IDF 가중치 값이 작으므로 그 영향

23) Shlomo Argamon, Shlomo Levitan. Measuring the Usefulness of Function Words for Authorship Attribution. ACH/ALLC 2005 Conference Abstracts book. 2005. pp.1-3.

24) Mike Kestemont. Function Words in Authorship Attribution From Black Magic to Theory?(Proceedings of the 3rd Workshop on Computational Linguistics for Literature) Association for Computational Linguistics. 2014. pp.59-66.

www.kci.go.kr

이 상대적으로 적기 때문에 이와 같은 결과가 나타난 것으로 이해할 수 있다.

이상의 내용을 아직 일반화하기는 어렵다. 그러나 한의학 고문헌 텍스트를 대상으로 저자 판별을 수행할 경우, 기능어를 특성(feature)로 하고 그 출현 빈도를 IDF 가중치 없이 사용하는 방법이 좋은 출발점이 될 수 있다. 이 경우, 유니-바이그램을 혼합하여 사용하였을 때 가장 좋은 성능을 보였으나, 유니그램만 사용한 경우와의 성능 차이가 크지 않으므로 연산의 복잡도를 고려한다면 '기능어/유니그램/TF' 조합도 합리적인 선택이 될 수 있다. 다만 최고 성능이 0.732로 높지 않기 때문에 문서에 대한 임베딩 방식과 분류 방식을 달리하여 검토해야 하며, 다른 종류의 한의학 고문헌에도 동일한 결과가 나타나는지에 대해서도 후속 연구가 필요하다.

Ⅲ. 결론

저자 판별(authorship attribution)은 텍스트로부터 작자의 특징을 유추하는 작업으로, 텍스트의 저자를 밝히기 위해 과거에는 '자세히 읽기(close reading)'와 같은 전통적인 방법이 사용되어 왔으나 현재는 자연어처리 및 기계학습 등을 통한 계량적인 방법이 연구되고 있다. 저자 판별은 글에 녹아 있는 저자의 글쓰기 습관이나 고유한 특징을 찾아내는 문제로 귀결되는데, 이를 '문체(style)'라고 할 수 있으므로 글의 숨겨진 문체를 포착해 내는 일이기도 하다.

한의학 고문헌은 동아시아 전통사회에서 한문(漢文)으로 작성된 텍스트로, 책의 저자를 알 수 없거나 확인이 필요한 경우, 여러 저자의 글이 혼합되어 있어 이를 구분해야 하는 경우 등의 문제가 존재한다.

본 연구의 목적은 한의학 고문헌 텍스트를 대상으로 한 저자 판별 문제를 계량적으로 수행하기 위해 가장 적합한 '특성(feature)'이 무엇인지 검토해 보는 것이다. 이를 위해 『난경집주(難經集註)』를 실험 코퍼스(Corpus)로 하고 『난경(難經)』 원문 및 주석을 5종의 서로 다른 저자의 문서(Documents)로 보고 실험을 수행하였다. 선행 연구를 바탕으로, 저자 판별 모델은 가장 보편적으로 사용되는 서포트 벡터 머신(SVM ; Support Vector Machine)을 사용하였다. 저자 판별에 가장 적합한 문서의 특성(feature)을 알아내기 위해, 기능어와 내용어의 차이(A) [모든 용어를 사용한 경우(A①), 기능어(function word)만 사용한 경우(A②), 내용어(content word)만 사용한 경우(A③)] , 단어와 연어의 차이(B) [유니그램(unigram)을 사용한 경우(B①), 바이그램(bigram)을 사용한 경우(B②), 양자를 모두 합친 경우(B③)] , IDF 가중치 적용 여부(C) [TF를 사용한 경우(C①), TFIDF를 사용한 경우(C②)] 로 나누어 살펴보았다.

실험 결과 '기능어(A②)/유니-바이그램(B③)/TF(C①)'를 사용한 경우에 가장 높은 정확도(0.732)를 보였고, '기능어(A②)/유니그램(B①)/TF(C①)' 조합에서 그 다음으로 높은 정확도(0.729)를 보였다. 반대로 '내용어(A③)/유니그램(B①)/TFIDF(C②)' 조합에서 가장 낮은 정확도(0.351)를 보였다.

이를 통해 한의학 고문헌 저자 판별에서 다음과 같은 사실을 알 수 있었다. 첫째, 기능어가 내용어에 비해 중요한 역할을 한다. 둘째, 내용어에서는 상대적으로 연어가 중요했지만, 기능어에서는 단어가 더 중요한 의미를 가진다. 셋째, 일반적인 텍스트 분석에서와 달리 IDF 가중치 적용이 더 좋지 않은 결과를 가져왔다.

마지막으로 본 연구 결과를 일반화할 수는 없으며, 최고 성능이 만족할 만큼 높지 않으므로 이에 대한 후속 연구가 진행되어야 할 것이다.

감사의 말씀

본 연구는 한국한의학연구원 주요사업 "AI 한의사 개발을 위한 임상 빅데이터 수집 및 서비스 플랫폼 구축(KSN2012110)"의 지원을 받아 수행되었습니다.

References

1. 김원중. 한문 해석 사전. 서울. 글항아리. 2013.

www.kci.go.kr

2. 최지명. 기계학습 알고리즘을 이용한 한국어 텍스트 저자 판별. 서울. 석사학위논문(연세대). 2015.
3. 강남준, 이종영, 최운호.『독립신문』논설의 형태 주석 말뭉치를 활용한 논설 저자 판별 연구 - 어미 사용빈도 분석을 중심으로. 한국사전학. 2010. 15.
4. 박경모, 최승훈.『강평 (康平) 상한론 (傷寒論)』의 고증을 통한『상한론(傷寒論)』과『황제내경(黃帝內經)』의 비교연구. 대한한의학원전학회지. 1995. 9.
5. 양승률. 주촌 신만의『보유신편(保幼新編)』편찬과『주촌신방(舟村新方)』. 장서각. 2011. 25.
6. 오준호. 한의학 고문헌 데이터 분석을 위한 단어 임베딩 기법 비교 : 자연어처리 방법을 적용하여. 대한한의학원전학회지. 2019. 32(1).
7. 이가은, 안상우. 소아의방(小兒醫方)의 판본비교(板本比較) 및 편제(篇第) 고찰(考察). 한국의사학회지. 2004. 17(1).
8. Bing-Cho Chan. The authorship of the Dream of the red chamber based on a computerized statistical study of its vocabulary. Hong Kong. Joint Publishing Co Ltd. 1986.
9. Hsieh-Chang Tu, Jieh Hsiang. A Text-Mining Approach to the Authorship Attribution Problem of Dream of the Red Chamber. Digital Humanities. 2013.
10. Hu, Xianfeng, Yang Wang and Qiang Wu. Multiple authors Detection: a Quantitative Analysis of Dream of the Red Chamber. Advances in Adaptive Data Analysis. 2014. 6.
11. İlker Nadi Bozkurt, Özgür Bağlıoğlu, Erkan Uyar. Authorship attribution: performance of various features and classification methods. 22nd International Symposium on Computer and Information Sciences, ISCIS 2007. IEEE. 2007.
12. Matthew L. Jockers, Daniela M. Witten. A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing. 2010. 25(2).
13. Mike Kestemont. Function Words in Authorship Attribution From Black Magic to Theory?(Proceedings of the 3rd Workshop on Computational Linguistics for Literature) Association for Computational Linguistics. 2014.
14. Patrick Juola. Authorship Attribution. Foundations and Trends in Information Retrieval. 2006. 1(3).
15. Qing-Xiang Yu. Applications of Statistical methods to Dream of the Red Chamber. Journal of National Cheng-Chi University. 1998. 76.
16. Shlomo Argamon, Shlomo Levitan. Measuring the Usefulness of Function Words for Authorship Attribution. ACH/ALLC 2005 Conference Abstracts book. 2005.
17. Smita Nirkhi, R.V.Dharaskar, V.M.Thakare. Authorship Identification using Generalized Features and Analysis of Computational Method. Transactions on Machine Learning and Artificial Intelligence. 2015. 3(2).
18. MEDICLASSICS [homepage on the Internet]. Korea Institute of Oriental Medicine; 2015 [cited 30 Jan 2020]. Available from: https://mediclassics.kr/books/149

www.kci.go.kr