대한한의학회지 제41권 제3호(2020년 9월)
J Korean Med. 2020;41(3):*1-8*
http://dx.doi.org/10.13048/jkm.20021
pISSN 1010-0695 • eISSN 2288-3339

**Original Article**

# 자연어 처리 및 기계학습을 통한 동의보감 기반 한의변증진단 기술 개발

이승현[1], 장동표[2], 성강경[3*]

[1]한양대학교 공과대학 정보시스템학과, [2]한양대학교 공과대학 생체공학과, [3]원광대학교 한의과대학 한의학과 내과학교실

## *Donguibogam*-Based Pattern Diagnosis Using Natural Language Processing and Machine Learning

Seung Hyeon Lee[1], Dong Pyo Jang[2], Kang Kyung Sung[3*]

[1]Department of Information System, Hanyang University, Seoul, Korea, [2]Department of Biomedical Engineering, Hanyang University, Seoul, Korea, [3]Department of Internal Medicine, College of Oriental Medicine, Wonkwang University, Iksan, Korea

**Objectives:** This paper aims to investigate the Donguibogam-based pattern diagnosis by applying natural language processing and machine learning.
**Methods:** A database has been constructed by gathering symptoms and pattern diagnosis from *Donguibogam*. The symptom sentences were tokenized with nouns, verbs, and adjectives with natural language processing tool. To apply symptom sentences into machine learning, Word2Vec model has been established for converting words into numeric vectors. Using the pair of symptom's vector and pattern diagnosis, a pattern prediction model has been trained through Logistic Regression.
**Results:** The Word2Vec model's maximum performance was obtained by optimizing Word2Vec's primary parameters —the number of iterations, the vector's dimensions, and window size. The obtained pattern diagnosis regression model showed 75% (chance level 16.7%) accuracy for the prediction of Six-Qi *pattern diagnosis*.
**Conclusions:** In this study, we developed pattern diagnosis prediction model based on the symptom and pattern diagnosis from Donguibogam. The prediction accuracy could be increased by the collection of data through future expansions of oriental medicine classics.

***Key Words*** *: Word2vector, Differentiation and Pattern Identification of Symptoms, Word Embedding, Natural Language Processing, Donguibogam*

## 서 론

최근 세계적으로 통합의학 연구가 활성화되면서 전통 한의학에 대한 관심이 높아지고 있다. 한의진단에서 중요한 변증(辨證, Differentiation and Pattern Identification of Symptoms)은 정체적 관점에서 질병의 성질, 부위, 세력(勢力) 등과 환자의 상태를 변별하는 한의학적 진단방법으로, 보고, 묻고, 듣고, 맥

• Received : 30 June 2020 • Revised : 25 July 2020 • Accepted : 27 July 2020
• Correspondence to : Kang Kyung Sung
Department of Internal Medicine, College of Oriental Medicine, Wonkwang University,
460 Iksan-daero, Sin-dong, Iksan 54538, Korea
Tel : +82-62-670-6412, Fax : +82-62-671-6414, E-mail : sungkk@wonkwang.ac.kr

을 짚는 사진과정을 통하여 수집된 증상들 사이에 내재하는 상호관계와 의미를 분석하는 과정을 통해서 이루어진다. 변증 진단 지표들은 임상적 유용성이 있음에도 불구하고 직관적 혹은 주관적인 방식으로 이루어져 왔기 때문에 객관적이고 합리적 분석에 기반한 정확하고 재생 가능한 변증형(辨證形)을 도출하는 데 어려움이 있다.

그동안 변증의 객관화를 위해 변증 지표의 표준화 연구들이 활발히 진행되어 변증 모델들이 제시되어 왔다[1-4]. 하지만 모델에서 설정한 모든 질문을 물어보고, 그 응답에 따라 진단하게 됨으로써 대상자가 주로 느꼈던 주요한 증상이 아닌 다른 항목에 응답을 하면서 바이어스가 들어갈 수도 있는 한계도 존재한다[5-8]. 따라서 보다 다양한 형태의 변증 도출 모델들에 대한 연구가 필요한 실정이다.

최근 빅데이터 및 머신러닝 기술의 발달과 더불어 자연어 처리(Natural Language Processing)에 대한 관심도 높아졌다. 특히 인터넷상의 대량의 문서나 SNS 자료로부터 최신 동향을 파악하기 자연어 처리 기술이 많이 적용되고 있으며, 이에 대한 기술도 급속이 발전하고 있다. 자연어 처리는 컴퓨터가 인간이 사용하는 언어를 이해하고, 분석할 수 있게 하는 기술을 총칭하는 말이다. 자연어 처리를 위해서 사람이 사용하는 언어나 글의 문장을 해석하여 형태소 단위의 단어로 정리하고, 이러한 단어와의 관계를 이용하여 컴퓨터가 인식 분석할 수 있도록 숫자화 하는 작업이 필요하다. 자연어 처리 기술 중 최근 워드2벡터(Word2vec)모델이 개발되었는데 이는  단어가 가지는 의미를 다차원 공간의 벡터화값으로 표현하는 것이다. 단어들의 의미자체를 벡터화하여 수치화하기 때문에 단어를 이용한 연산이 가능하기 때문에 자연어 처리에 널리 사용되고 있다.

한국의 대표 한의학 서적인 동의보감은 질병을 치료함에 있어 환자의 형색을 구분하여 장부의 상태를 확인하고, 이를 기반으로 병인을 분석함으로써 치료를 하는 방식으로 다양한 임상 기록이 담겨있으며 여러 가지 증상에 대한 서술과 함께 병인을 분석하는 중요한 특징으로 변증형이 제시되고 있다.

본 논문에서는 이러한 동의보감의 모든 증상과 변증을 데이터베이스화하고, 증상에 대한 표현 및 문장에 자연어 처리 기술과 워드2벡터 모델을 적용함으로써 변증과 증상과의 관계를 분석하고자 한다. 특히 증상과 변증의 관계분석에 머무르지 않고 더 나아가 증상에 기반한 변증예측 시스템을 구성함으로써 동의보감 고전에 기반한 객관적인 변증형 도출모델을 만들고자 한다.

## 본 론

### 1. 동의보감 증상-변증 데이터베이스화

동의보감에서 증상들과 이에 대한 변증을 모아 데이터베이스(database, DB)를 구성하기 위해 한의학 고전 데이터(https://mediclassics.kr/)의 동의보감 자료를 바탕으로 작업을 진행하였다[8]. 증상 및 변증 DB는 Fig. 1A처럼 각 증례의 증상들의 리스트를 증상 번호로 정리한 후, 증상번호들의 조합으로 하나의 변증 혹은 처방명으로 정리하였다. 예를 들어 Fig. 1A에서 볼 수 있듯이, 동의보감의 “혈이 중초에 쌓이고, 몸이 누렇게 되고, 어혈이 말라 뭉친다.” 라는 3가지 증상조합은 ‘혈축중초’라는 변증으로 낸다. 이렇게 모든 증상조합과 변증 정리를 내경편<26편>, 외형편<26편>, 잡병편<36편>에 대해 적용하였다.

### 2. 증상형태소 분석

증상은 질환으로 인해 나타나는 현상을 언어로 기술한 것으로써 이 증상을 이용하여 어떤 변증으로 유추하는 알고리듬이나 소프트웨어를 구현하기 위해서 증상 문장 자체를 작은 단위인 형태소로 구분하여 분석하는 작업이 필요하게 된다. 문장을 의미를 갖는 최소단위인 형태소로 나눠주는 과정을 토큰화 과정이라고 한다. 토큰화 과정에서는 문장을 형태소 단위로 나누는 작업을 거치는데 이를 파싱(parsing)

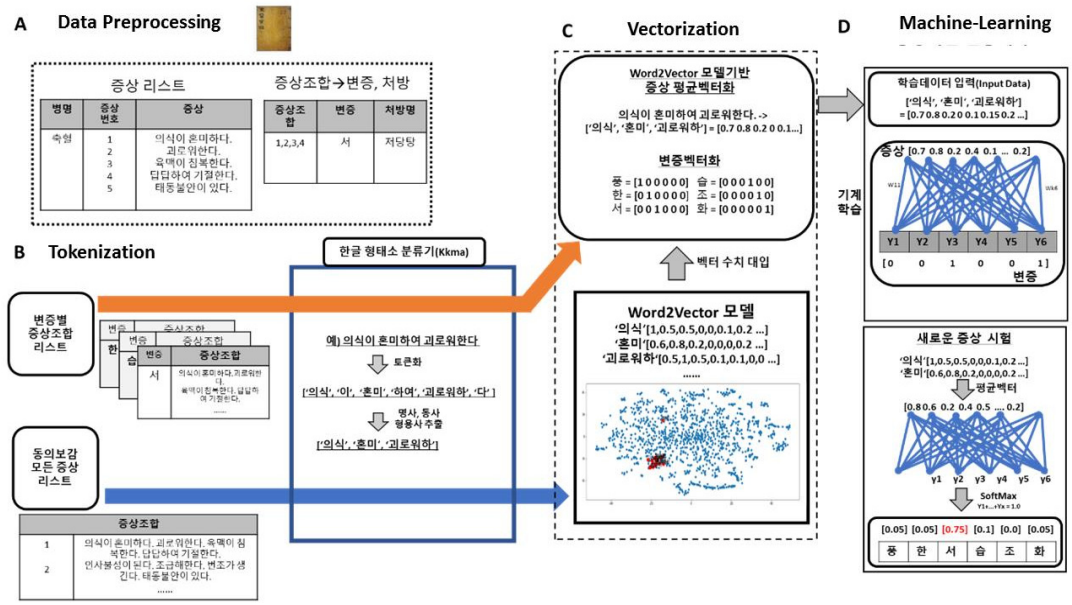


**Fig. 1.** Natural Language Processing and Pattern Diagnois Prediction Model Based on *Donguibogam*.

이라고 한다. 본 연구에서는 형태소 분석 Konlpy 패키지 중에서 토큰화 과정에서 한글형태소 분석에 흔히 사용되고, 사전(dictionary)를 사용자 임의로 교정할 수 있는 장점을 가진 꼬꼬마(kkma)패키지를 사용하였다[9]. 동의보감에서 증상표현에는 상대적으로 고어가 많고, 한의학적 표현이 많기 때문에 이에 대한 보완이 없는 경우 형태소가 분석에 오류가 빈번히 발생하여 정확도가 떨어지는 단점이 있다. 예를 들어 고어 및 한의학적 사전이 적용되지 않으면 [상한리증이 있다. 양명병이 있다.]라는 古語는 [‘상한’, ‘리’, ’증’, ‘이’, ‘있다’, ‘양’, ‘명’, ‘병’, ‘이’, ‘있다’]로 토큰화되는 문제점이 있다. 따라서 한의학 사전 및 오래된 언어 및 단어 등을 추가하여 형태소분석 성능을 높일 수 있는 꼬꼬마형태소가 적합하다고 판단하여 이를 추가하였다.

본 연구에서는 Fig. 1B와 같이 두 가지 형태의 증상리스트에 대해서 형태소 분석을 적용하였다. 첫 번째는 동의보감에 나오는 증상들의 조합을 모아서 증상리스트로서 Fig. 1C처럼 형태소 분석 후 워드2벡터 모델 구성에 사용되었다. 동의보감의 모든 단어(형태소)를 가지고 모델을 구성해야 하기 때문에 병은 진단하는데 사용한 모든 증상조합데이터에 대해 적용하였다. 두 번째는 변증별로 모아진 증상조합리스트에 대해 형태소분석을 적용하였다. 예를 들어 육기변증의 “풍한서습조화” 중 “풍”에 관련된 증상조합리스트, “한”에 관련된 증상조합리스트 등을 따로 따로 형태소분석을 진행하였다.

## 3. 증상 및 변증의 벡터화

모든 증상조합리스트의 형태소 분류기 적용 후 조사 및 부사 등은 빼고 명사, 동사, 형용사만을 수치화하여 워드2벡터 모델을 적용하였다. 워드2벡터는 단어들 간의 유사도를 벡터화한 것이다. 워드2벡터 모델을 만들기 위해서는 핵심단어를 수학적으로 표현 즉 벡터화 과정을 거쳐야 하는데, 이 과정을 워드임베딩이라 한다[10]. 워드2벡터 모델에 적용하게 되면

동의보감에서 증상을 표현하기 위해 사용되는 명사, 동사, 형용사의 단어들은 다차원공간의 맵핑되게 되는데, 증상조합에서 함께 자주 사용되는 단여들은 연관성이 높은 것으로 판단하여 비슷한 공간에 놓이게 됨으로써 비슷한 벡터값으로 나타나게 된다[11]. 따라서 최적화된 워드2벡터 모델을 만들기 위해서는 워드2벡터의 하이퍼파라메터(Hyper-Parameter)에 따라 모델의 정확도가 영향을 받는다. 워드2벡터 모델링 과정에서 중요하게 고려되는 하이퍼파라메터는 크게 세 가지이다. 첫 번째 파라메터인 벡터차원은 표현되어지는 단어의 벡터의 차원의 크기를 나타낸다. 벡터차원크기를 10으로 설정하면 한 단어 당 갖는 고유의 벡터값은 10차원에 분포된다. 두 번째 파라메터인 윈도우(Window)는 증상 문장에서 연관성 학습시 고려되는 주변단어의 수를 의미한다. 예를 들어 윈도우 1 이라면 중심단어를 기준으로 앞의 한 단어와 뒤의 한 단어를 주변단어로 묶어서 함께 사용되어 높은 유사성을 갖는 것으로 학습시킨다. 마지막 파라메터로는 학습의 반복횟수이다. 훈련 데이터를 반복함으로써 모델의 학습횟수가 많아질수록 단어들이 더 정교하게 고유의 벡터값을 가지는 경향이 있게 된다. 증상을 입력하여 변증을 예측하는 시스템을 구성하기 위해서는 비슷한 의미의 단어들이 비슷한 벡터로 표현될 수 있도록 동의보감 기반의 정확한 워드2벡터 모델이 구축이 필수적이다. 따라서 본 연구에서는 파라메터를 변화시켰을 때 서로 같은 변증에서 많이 사용되는 단어들이 서로 비슷한 벡터를 가지는 측정하기 위해 코사인 유사도(Cosine Similarity)를 사용하였다. 서로 비슷한 두 단어의 벡터의 코사인 유사도는 높은 값을 가지고, 다른 단어들의 벡터의 코사인 유사도는 낮은 값을 가지게 된다. 따라서 최적화비율은 유사도가 높은 단어의 평균 코사인 유사도와 임의 단어의 평균 코사인 유사도 값의 비로 계산하였다. 예를 들어, 육기변증의 중심단어 중 하나인 '습'을 기준으로 유사도가 높은 단어들을 5개(습지, 습열, 풍습, 이슬, 훈증)와 임의의 단어 5개를 뽑아 최적화비율 계산하였다.

모든 증상조합리스트를 이용하여 워드2벡터모델을 구성한 후, 변증 별 증상조합리스트를 이용하여 증상기반 변증예측 기계학습을 위한 훈련입력데이터를 구성하였다. 예를 들어, '의식이 혼미하여 괴로워한다' 증상조합이 육기변증의 '서'로 변증이 정의되어 있는 경우, 증상의 경우 '의식', '혼미', '괴로워하다'의 각 벡터들의 평균벡터 [0.7 0.8 0.2 0.0 0.1]로 표현되고, '서'는 [0.0 0.0 1.0 0.0 0.0 0] 벡터로 변환되어 증상과 변증 벡터 쌍을 구성하였다.

### 4. 기계학습을 이용한 증상기반 변증예측 모델

이전 단계에서 구성된 증상벡터와 변증벡터 쌍은 신경망의 학습데이터를 들어가게 된다. 증상벡터를 입력으로 넣었을 때, 출력으로 변증벡터가 결과로 나올 수 있도록 신경망 기계학습을 하였다. 증상조합과 변증 쌍 데이터를 이용하여 로지스틱 회귀분석(Logistic Regression)을 이용한 변증 예측 모델을 훈련하였다. 훈련 후, 새로운 증상조합을 넣으면 변증을 예측할 수가 있게 된다. Fig. 1D처럼, 새로운 증상조합문장은 형태소분석을 거치고, 워드2벡터 모델에서 구해진 평균증상벡터를 입력으로 넣으면 육기변증의 각각의 변증에 해당할 수 있는 확률이 계산되어 출력으로 나오게 된다. 가장 높은 확률값을 가지는 육기변증 중 하나가 변증으로 예측하였다.

## 결 과

### 1. 워드2벡터 최적화

워드2벡터 모델의 성능의 최대치를 측정하기 위해 워드2벡터의 주요 파라메터인 반복횟수, 벡터차원, 윈도우 길이를 조정하면서 성능을 측정했다. 반복횟수의 경우 2의 승수(2, 4, 8, 16, 32, 64, 128회)로 증가시키면서 코사인유사도에 의한 최적화비율을 계산하였다. Fig. 2A에서와 같이 반복횟수가 증가함에 따라서 최적화비율이 높아지며, 반복횟수 32회 이상에

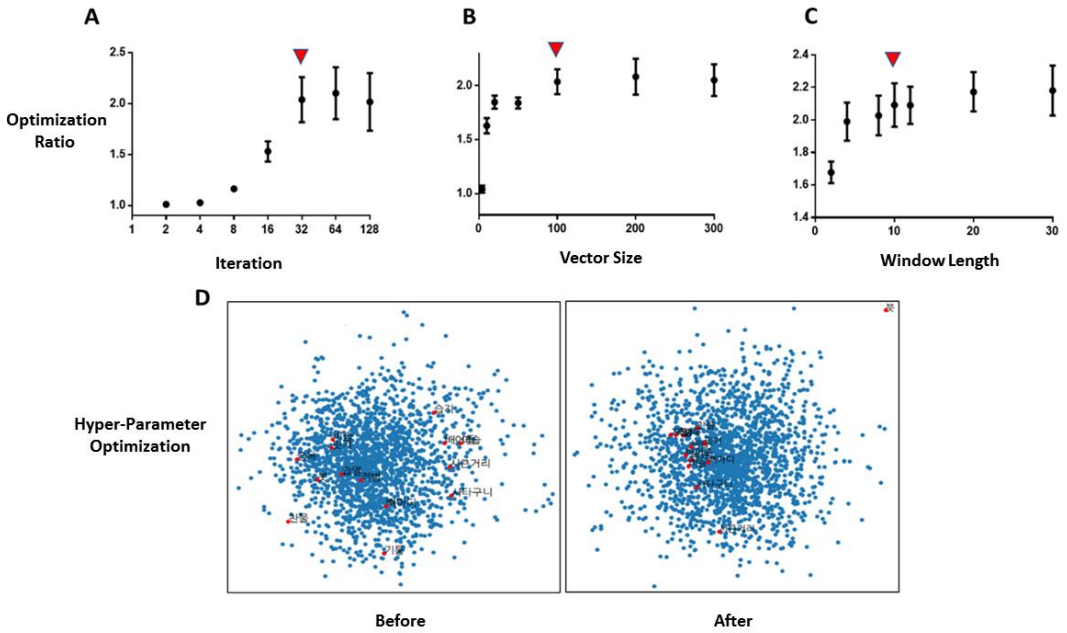


**Fig. 2.** Optimization of Word2Vector Model.

서는 변화가 크지 않은 것으로 나타났다. 따라서 이는 반복횟수 32회 이상에서 워드2벡터 성능의 높다는 것을 나타낸다. 두 번째 단어를 벡터로 나타낼 때 벡터차원의 크기에 대해서는 10, 20, 50, 100, 200, 300차원을 이용하여 최적화비율을 확인해 본 결과, 100차원 이상에서는 최적화비율의 변화의 폭이 적어지는 것을 보였다. 같은 방식으로 윈도우 길이 2, 4, 8, 10, 12, 20, 30에 대해서는 Fig. 2C에서 볼 수 있듯이, 윈도우 길이가 10인 경우 모델이 최적화됨을 보였다. Fig. 2D는 워드2벡터 모델의 성능을 2차원에 맵핑하여 시각화한 것이다. Fig. 2D에서 빨간 점으로 표시한 단어는 육기변증 중 '습'에서 자주 나오는 단어들을 표시한 것이고, 파란 점들은 증상조합리스트에 있는 모든 단어를 나타낸 것이다. Fig. 2D에서 보는 것처럼 파라메터를 최적화하기 전에는 비슷한 의미의 단어가 넓게 분포되어 있지만, 최적화 후에는 비슷한 의미의 단어가 비슷한 공간 벡터로 표시되어 워드2벡터 모델이 잘 구성되었음을 나타내고 있다.

## 2. 기계학습을 이용한 증상기반 변증예측

Fig. 1D에서처럼 동의보감 증상조합 리스트와 육기변증 쌍을 정리했을 때 '풍'에 관련된 증상조합 개수는 총 526개 추출하였고, '한' 593개, '서' 115개, '습' 351개, '조' 55, '화' 271개가 추출되었다. 이후 로지스틱 회귀분석 기반의 기계학습을 진행하였는데, 이 중 90%는 학습훈련으로 사용하였고, 10% 데이터는 정확도를 계산하는데 사용하였다. 안정적인 값을 얻기 위해 훈련 및 정확도 계산을 5회 반복하였다. 또한 워드2벡터 모델은 최적화된 파라메터 반복횟수(Iteration) 32회, 100 차원 벡터(Vector Size), 10의 윈도우길이(Window Length)로 모델의 파라메터를 조정 후 증상조합벡터를 얻었다. 그 결과, 육기변증을 예측하는 정확도는 75% (찬스레벨 16.7%)를 얻

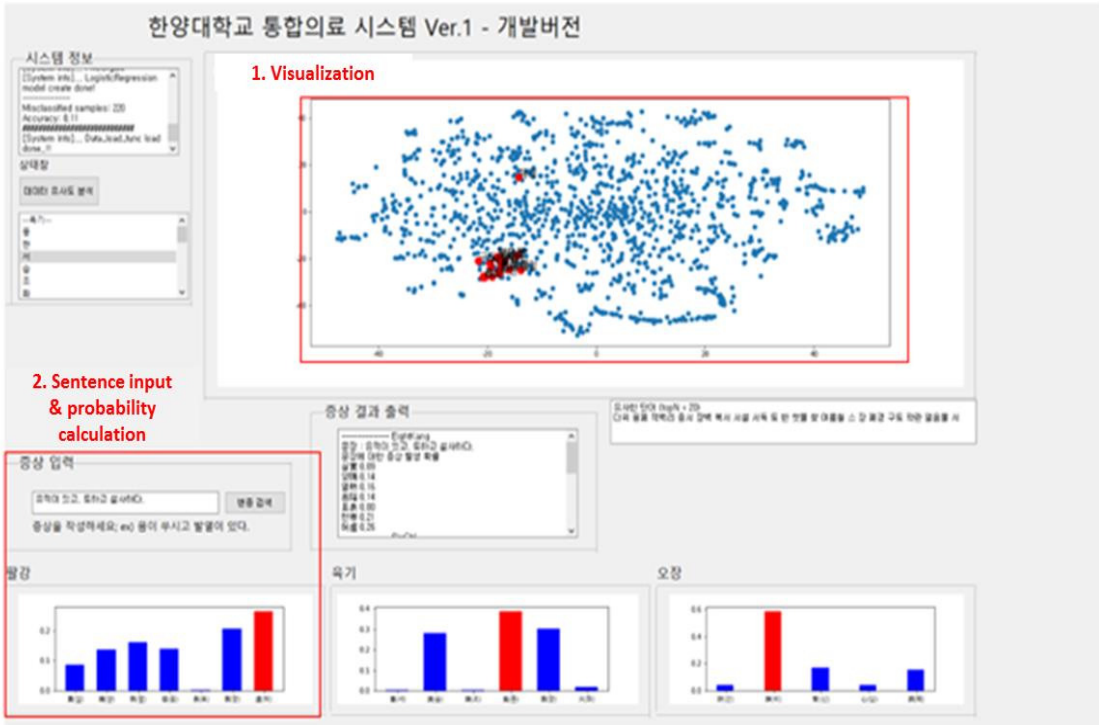


Fig. 3. Word2Vector Model-Based Diagnostic Prediction Program.

을 수 있었다. 이를 기반으로 하여, Fig. 3과 같이 변증 소프트웨어를 구성하였다.

## 고 찰

### 1. 형태소 분석기의 한계 및 고성능의 한의학 사전 구축 필요

古語 및 한의학 전문 용어들을 꼬꼬마 사전에 추가했음에도, 일부 단어들이 원활하게 토큰화되지 않는 문제가 발생했다. 예를 들어, [적백리가 있다.]라는 단어는 [‘적백리’, ‘가’, ‘있다’]로 토큰화 되어야 하지만, [‘적백’, ‘리가’, ‘있다’]로 토큰화 됨을 확인할 수 있었다. 이 문제는 꼬꼬마 형태소 분석기 사전에 있는 ‘적백’, ‘리가’ 라는 단어가, ‘적백리’, ‘가’보다 먼저 토큰화 되었기 때문에 발생한 것이다. 이러한 문제를 해결하기 위해, ‘리가’, ‘기가’와 같이 형태소 분석에 오류를 만들 수 있는 단어들을 사전에서 삭제하였다. 띄어쓰기를 기준으로 형태소를 분석하는 영어 형태소 분석과는 달리, 한글은 교착어의 특성상 어근과 어미, 접미사와 조사 등으로 복잡한 언어구조를 특성으로 한다. 더불어 古語 및 한의학 전문 용어들을 다수 포함한 동의보감 증상조합들은 기존의 한글 형태소 분석기로는 모두 분석할 수 없었다. 이러한 문제점을 해결하기 위해, 한의학 사전에 단어들을 추가하여 해결하였다. 추후에는 한의학 단어를 먼저 인식하도록 할 수 있는 형태소 분석기를 마련하여, 효과적으로 한의학 단어를 인식할 수 있도록 해야 할 것이다.

### 2. 동의보감 증상 및 변증 데이터

 http://dx.doi.org/10.13048/jkm.20021

일반적으로 워드2벡터 모델을 만들 때, 워드2벡터 모델이 수 십 만개에서 수 백 만개의 단어들의 관계를 학습하도록 만든다. 하지만, 현재 모델의 중복 증상조합 리스트를 제외한 문장의 개수는 6천여 개에 불과하다. 모델이 단어들 간의 관계를 더 정확하게 학습하기 위해서는 충분한 증상조합의 리스트의 구축이 필요하기 때문에, 동의보감뿐 만이 아니라, 다양한 한의학 서적들의 변증 증상조합리스트들을 포함하여 모델을 구축하여야 할 필요가 있다. 예를 들어, 육기변증 총 증상조합 1,695개 중에서 '조' 변증에 속한 증상조합의 수는 55개였다. 이는 총 증상조합 대비 약 3.2% 수준 밖에 되지 않았고, 적은 증상조합이 있을 경우, 증상조합의 변증 예측 테스트 및 학습에도 어려움이 있었다. 풍부한 증상 및 변증 데이터의 확보는 현재 75%인 변증예측 정확도를 향후 높이는데 중요한 역할을 할 수 있을 것이라 생각된다.

## 결 론

본 연구에서는 증상에 기반하여 변증을 예측하는 시스템을 구성함으로써, 동의보감에 기반한 객관적인 변증추측 모델을 만들었다. 인공지능 및 기계학습 분야에서 흔히 사용되는 워드2벡터 모델과 동의보감의 증상조합 및 변증 데이터에 기반하여 변증 예측하는 알고리즘을 구현함으로써 현재 75% 변증 예측률(6가지의 변증)을 보였다. 이 정확도는 다양한 한의학 고전 데이터 확보를 통해 좀 더 높일 수 있을 것이라 생각된다. 더욱이, 본 논문에서 다룬 변증 분류 기술이 임상에서도 적용되기 위해서는 임상에서 사용되는 변증의 실제 예와 동의보감 원문에 나오는 변증들을 비교해야 할 것이라 판단된다.

## 연구 사사

본 연구는 보건복지부 재원으로 통합의료연구지원사업의 지원에 의하여 이루어진 것임. [과제번호 : B0080613000158]

## 참고문헌

1. Kim JK, Seol IC, Lee I, Jo HK, Yu BC, Choi SM. Report on the Korean standard differentiation of the symptoms and signs for the stroke-1. J Physiol Pathol Korean Med. 2006;20(1):229-34.
2. Kang BK, Go HY, Kim JK, Kim BY, Ko MM, Kang KW, et al. Study of concordance rate to measure symptoms in interanl researchers. J Physiol Pathol Korean Med. 2006;20(6):1728-31.
3. Go HY, Kim JK, Kang BK, Kim BY, Ko MM, Kang KW, et al. Report on the Korean standard differentiation of the symptoms and signs for the stroke-1 (KSDSSS-1). J Physiol Pathol Korean Med. 2006;20(6):1789-92.
4. Go HY, Kim JK, Kang BK, Kim BY, Ko MM, Kang KW, et al. Survey of stroke subtype classification. J Physiol Pathol Korean Med. 2007;21(1):318-21.
5. Choi SM, Yang KS. Standardization and unification of the terms and conditions used for diagnosis in oriental medicine. Korean J Orient Med. 1995;1(1):101-25.
6. Yang KS, Choi SH, Choi SM, Park KM, Jeong WY, Ahn KS, et al. Standardization and unification of the terms and conditions used for diagnosis in oriental medicine. II. Korean J Orient Med. 1996;2(1):381-401.
7. Choi SM, Yang KS, Choi SH, Park KM, Park JH, Shim BS, et al. Standardization and unification of the terms and conditions used for diagnosis in oriental medicine III. Korean J Orient Med. 1997;3(1):41-65.
8. KOREA INSTITUTE OF ORIENTAL MEDICINE

(KIOM). 한의학고전DB.
9. 서울대학교 IDS연구실. 꼬꼬마(KKMA) 세종 말뭉치 활용 시스템. 2010. Available from: http://kkma.snu.ac.kr/
10. Eddie. 딥 러닝을 이용한 자연어 처리 입문. 대한민국. Wikidocs. 2020.
11. Yogatama. Learning Word Representations with Hierarchical Sparse Coding. ICML (International Conference on Machine Learning). 2015.

## ORCID

이승현 https://orcid.org/0000-0002-7479-9359
장동표 https://orcid.org/0000-0002-2832-2576
성강경 https://orcid.org/0000-0002-6310-7556