JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 10, pp. 105-111, October 2024
https://doi.org/10.9708/jksci.2024.29.10.105

Named entity normalization for traditional herbal formula mentions

Ho Jang*

*Senior Researcher, KM Data Division, Korea Institute of Oriental Medicine. Department of Korean Convergence Medical Science, University of Science and Technology, Daejeon, Korea

[Abstract]

In this paper, we propose methods for the named entity normalization of traditional herbal formula found in medical texts. Specifically, we developed methodologies to determine whether mentions, such as full names of herbal formula and their abbreviations, refer to the same concept. Two different approaches were attempted. First, we built a supervised classification model that uses BERT-based contextual vectors and character similarity features of herbal formula mentions in medical texts to determine whether two mentions are identical. Second, we applied a prompt-based querying method using GPT-4o mini and GPT-4o to perform the same task. Both methods achieved over 0.9 in Precision, Recall, and F1-score, with the GPT-4o-based approach demonstrating the highest Precision and F1-Score. The results of this study demonstrate the effectiveness of machine learning-based approaches for named entity normalization in traditional medicine texts, with the GPT-4o-based method showing superior performance. This suggests its potential as a valuable foundation for the development of intelligent information extraction systems in the traditional medicine domain.

▸ **Key words**: Named entity recognition, Named entity normalization, GPT, BERT, Traditional herbal formula

[요 약]

본 논문에서는 의학 텍스트에 기술된 한의 처방명의 개체명 정규화 방법을 연구하였다. 구체적으로, 주어진 텍스트에서 개체명으로 인식된 처방 명칭과 처방의 약어 등 처방 멘션들이 동일한 처방 개념을 가리키는지를 판단하는 방법론을 연구하였다. 이를 위해 두 가지 접근 방식을 시도하였다. 먼저, 의학 텍스트에 등장하는 처방 멘션에 대해 BERT 기반의 문맥 벡터와 멘션의 문자 유사도 정보를 기계 학습 모델의 특징으로 사용하여, 두 멘션의 동일 여부를 판별하는 지도 학습 기반 분류 모델을 구축하였다. 다음으로, GPT-4o mini 및 GPT-4o 기반의 프롬프트 질의 방식을 활용하여 동일한 작업을 수행하였다. 두 방법 모두 Precision, Recall, F1-score에서 0.9 이상의 성능을 보였으나, GPT-4o 기반 방법이 가장 높은 Precision과 F1-Score를 기록하였다. 본 연구의 결과는 한의학 텍스트에서 개체명 정규화를 위한 기계 학습 기반 접근 방식이 유의미한 성능을 달성할 수 있음을 보여주며, 특히 GPT-4o 기반 방법이 뛰어난 Precision과 F1-Score를 보임으로써 향후 한의학 도메인에서 지능형 정보 추출 시스템 개발에 중요한 기초 자료로 활용될 수 있을 것으로 기대된다.

▸ **주제어**: 네트워크 약리학, 한의학, 중의학, 데이터 통합, 생물정보학

• First Author: Ho Jang, Corresponding Author: Ho Jang
*Ho Jang (jh@kiom.re.kr), KM Data Division, Korea Institute of Oriental Medicine. Department of Korean Convergence Medical Science, University of Science and Technology
• Received: 2024. 09. 03, Revised: 2024. 10. 04, Accepted: 2024. 10. 04.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

정보 추출(Information Extraction, IE) 기법은 비정형 텍스트 데이터를 구조화된 정보로 변환하는 작업으로, 현대의 데이터 중심 환경에서 매우 중요한 역할을 한다. 인터넷, 소셜 미디어, 이메일, 연구 논문, 뉴스 기사 등에서 생성되는 데이터의 대부분은 비정형 텍스트 형식이다. 이러한 비정형 데이터는 분석과 활용이 어려워, 이를 구조화된 형태로 변환하는 것이 중요하다.

개체명 인식(Named Entity Recognition, NER)은 정보 추출의 하위 작업으로, 자연어 처리 및 텍스트 마이닝에 있어 필수적인 기술이다. NER은 텍스트에서 인물, 장소, 조직명, 시간, 날짜 등 관심 있는 정보를 자동으로 식별하고 분류하는 문제에서부터, 특정 도메인 텍스트에서 중요한 고유 명사를 자동으로 식별하고 분류하는 문제까지 다양한 텍스트 데이터에 적용된다. 멘션(Mention)은 텍스트 내에서 관심 있는 특정 개체를 나타내는 단어나 구절을 말한다. 예를 들어, 텍스트에서 "Steve Jobs"라는 단어가 나오면, 이는 "사람"이라는 개체를 가리키는 멘션이 된다. NER의 주요 목표 중 하나는 이러한 멘션을 텍스트에서 정확하게 식별하고, 이들이 어떤 유형의 개체인지를 분류하는 것이다.

개체명 정규화(Named Entity Normalization)는 이렇게 인식된 멘션을 표준화된 형태로 변환하는 작업이다. 예를 들어 텍스트에서 "뉴욕"이라는 지명은 "New York", "NYC" 등으로 다양하게 표현될 수 있는데, 이를 "뉴욕"이라는 동일한 개체로 처리하는 작업이다. 이 두 가지 작업을 통해 텍스트에 등장하는 다양한 정보가 구조화된다.

한의학 분야에서도 다양한 텍스트가 축적되고 있다. 일례로 지난 수십 년간 한의학적 치료에 사용되는 한의 처방(Traditional Herbal Formula) 및 약재(Medicinal)가 질병에 미치는 효능을 다룬 세포 실험, 동물 실험, 임상 실험, 리뷰 논문 등이 국제 저널을 통해 발표되고 있으며, 이러한 수천 건의 논문의 영문 초록은 PubMed에서 누구나 열람할 수 있다. 그러나 한의학만의 특성을 고려하여 구조화된 정보로 변환하기 위한 기술 개발은 아직 진행 중이다.

따라서 본 논문에서는 한의 처방에 대한 개체명 정규화 적용의 일환으로, 영문 의학 초록에 등장하는 처방 멘션(Traditional Herbal Formula Mention)들이 동일한 한의 처방을 지칭하는지 여부를 자동으로 판단하는 문제를 해결하기 위한 인공지능 기술의 적용을 시도하였다. Fig. 1은 연구 논문의 초록에서 한의 텍스트 데이터에 대한 NER 및 개체명 정규화 사례를 보여준다[1]. Fig. 1 (a)에서는 텍스트 데이터에서 네 개의 구분된 처방 멘션이 인식된다 (Banhabaekchulcheonma-tang, BT, Oryeong-san, OS). Fig. 1 (b)의 개체명 정규화 단계에서는 Banhabaekchulcheonma-tang과 BT가 반하백출천마탕이라는 동일한 개체를 지칭하고, Oryeong-san과 OS는 오령산이라는 동일한 개체를 지칭하는 것이 구분된다.

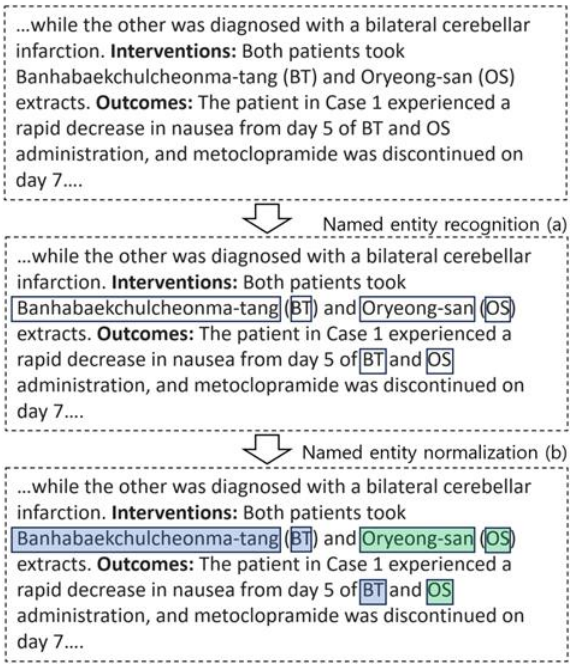

Fig. 1. An example of NER and named entity normalization for traditional herbal formula mentions

NER의 경우 Huggingface (https://huggingface.co/)와 같은 공개된 기술에 쉽게 접근할 수 있어, 한의학 도메인을 위한 모델 구현이 가능하다. 그러나 개체명 정규화의 경우 일반적으로 적용할 수 있는 방법은 부재한 실정이다. 자연어 처리 분야에서는 최근 획기적인 기술의 발전으로, Transformer 기반의 BERT(Bidirectional Encoder Representations from Transformers)[2]와 OpenAI GPT (Generative Pre-trained Transformer)[3]가 주목받고 있다. BERT는 Transformer 아키텍처를 기반으로 양방향(Bidirectional) 텍스트 처리를 통해 사전 학습(Pre-training)된 모델을 파인튜닝(Fine-tuning)하여 텍스트 분류, 개체명 인식 등에 널리 쓰인다. BERT는 단어, 문장, 텍스트 전체를 동적으로 임베딩(Embedding)할 수 있는 능력을 가지고 있다. GPT는 Transformer의 디코더 부분을 기반으로 한 자동 회귀 언어 모델로, GPT-1을 시작으로 현재의 GPT-4에 이르는 대규모 언어 모델이다. GPT 아키텍처는 방대한 양의 텍스트 데이터를 학습한 신경망 모델로, 인간의 언어를 이해하고 생성할 수 있다. 수

억에서 수천억 개 이상의 파라미터를 통해 언어 패턴, 문맥, 구조 등을 학습하며, 텍스트 생성, 이해, 요약 등에 활용된다.

본 논문에서는 이러한 두 가지 NLP 기술을 한의 처방 개체명의 정규화에 적용할 가능성을 연구하였고, 특정 전문 도메인에서도 이러한 최신 기술이 충분히 활용될 수 있음을 확인하였다. 논문의 구성은 다음과 같다. 2장에서는 연구 방법, 3장에서는 결과, 4장에서는 논의, 5장에서는 연구 결과 요약을 기술하였다.

II. Material and methods

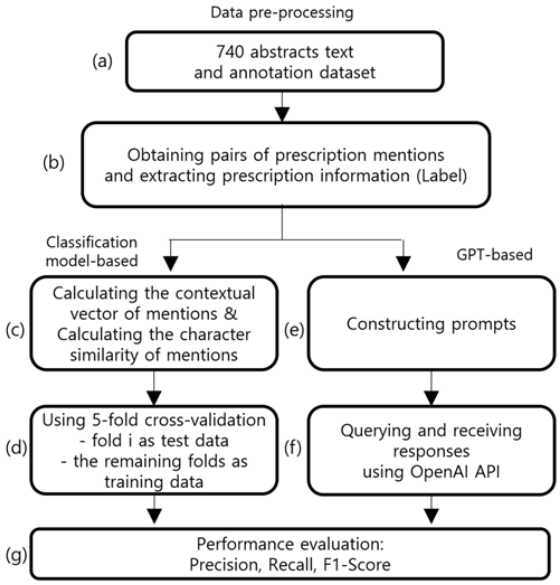

Fig. 2. The steps of this study

1. Research problem, data description, and performance measurement

Fig. 2는 개체명 정규화 성능 평가를 위한 연구의 흐름을 보여준다. 본 연구에서는 기존에 구축된 한의 처방과 질병 간의 관계 추출을 위한 코퍼스를 활용하였다(Fig. 2 (a)). 이 코퍼스는 선별된 740개의 PubMed 초록에서 여러 명의 한의학 도메인 전문가들이 직접 텍스트를 읽고 처방명의 멘션을 태깅한 결과물이다. 어노테이터들이 독립적으로 작업한 데이터는 이후 큐레이션 과정을 통해 하나의 통합된 데이터 셋으로 정제되었다. 740개의 초록에는 총 6,211개의 멘션이 포함되어 있으며, 이 코퍼스는 GitHub (https://github.com/KIOM-AIDoc/TFDR/tree/main)에서 열람할 수 있고, 현재 논문 심사 중에 있다.

본 연구는 텍스트에 하나 이상의 한의 처방 멘션이 있을 경우, 동일한 개체인지 여부를 판별하는 기술을 연구하는 것이다. 이를 위해 740개의 데이터 중 하나 이상의 구분되는 한의 처방 멘션이 있는 텍스트로 범위를 한정하였다. 예를 들어, 코퍼스에 포함된 소청룡탕(小靑龍湯)에 대한 연구[4]의 초록에는 So-Cheong-Ryong-Tang과 그 약어인 SCRT가 포함된 두 개 이상의 멘션이 존재하여 본 연구에서 사용되었다. 반면, 코퍼스에 포함된 온담탕(溫膽湯)에 대한 연구[5]의 초록에는 Wen-Dan Decoction이라는 단일 멘션만 존재하여 본 연구에서는 제외되었다.

앞서 선별된 각 문헌에서 등장하는 처방 멘션들 간의 이항 조합(Pair Combination)을 구한 후, 두 멘션이 동일한 개체를 지칭하는지 여부를 판별하는 것이 본 연구의 핵심 문제이다. Label Yes (Positive)는 두 처방 멘션이 동일한 처방을 지칭하는 경우이고, No (Negative)는 두 멘션이 서로 다른 개체를 지칭하는 경우이다. Table 1은 Fig. 1에서 도출된 4개의 멘션 조합으로부터 나온 6가지 데이터 사례를 보여준다. 성능 평가는 Precision, Recall, F1-Score를 통해 측정되었다(Fig. 2 (g)).

Table 1. Example data

Mention A	Mention B	Label
Banhabaekchulcheonma-tang	Oryeong-san	No
BT	OS	No
Banhabaekchulcheonma-tang	OS	No
BT	Oryeong-san	No
Oryeong-san	OS	Yes
Banhabaekchulcheonma-tang	BT	Yes

2. Entity disambiguation between mention pairs based on contextual vectors and character similarity

앞 단계에서 생성된 데이터를 활용하여 두 멘션이 동일한 개체를 지칭하는지 여부를 판별하기 위해 분류 모델 기반 연구를 진행하였다(Fig. 2 (c-d)). 조사 결과, 개체명 정규화 또는 멘션 일치 판별을 위한 널리 통용되는 방법론은 없는 것으로 확인되었고, 이에 따라 본 연구의 도메인에 적합한 방법론을 자체적으로 개발하였다. 우리는 텍스트를 BERT에 입력하여 처방 멘션에 해당하는 문맥 벡터(Context vector)를 생성하였으며, 이를 위해 Huggingface에서 제공하는 'bert-base-uncased'를 사용하였다. 단일 멘션에 대해 생성된 문맥 벡터의 크기는 768이며, 두 멘션의 문맥 벡터가 모델의 특징(feature)으로 사용되었다.

또한, 두 멘션의 문자열 유사도를 추가 특징으로 사용하였다. 문자열 유사도를 측정하기 위해 Jaro-Winkler 거리를 사용하였는데, Jaro 거리는 두 문자열 간의 공통 문자와 그 문자의 위치 차이를 기반으로 유사도를 계산한다. Jaro-Winkler 거리는 여기에 더해, 문자열이 처음부터 얼마나 일치하는지를 평가하여 유사도를 더 중요하게 반영한다. 즉, 두 문자열이 앞부분에서 많이 일치할수록 유사도가 더 높게 평가된다[6]. 이 방법은 철자가 약간 틀리거나 오타가 있는 경우에도 두 문자열 간의 유사성을 효과적으로 계산하는 것으로 알려져 있다.

지도학습 기반 분류 모델의 입력 특징 크기는 1,537이다(Fig. 3). 우리는 Random Forest(RF), XGBoost(XGB), SVM(Support Vector Machine)기법을 사용하여 분류 모델을 생성했다. 모델 구축에는 파이썬의 sklearn 패키지를 사용하였다. 성능 평가는 5-fold Cross-validation을 통해 학습 데이터와 테스트 데이터를 분리하고, 분류 성능을 평가하였다. 특히 본 논문에서 SVM을 사용한 방법을 Context Vector Character Similarity SVM (cvcsSVM)으로 명명하였다.

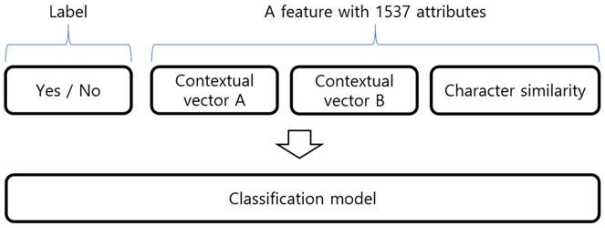

Fig. 3. Label and features used by classification models to determine the same formula given two mentions

3. Entity disambiguation between mention pairs using GPT-4o Mini and GPT-4o

다음으로, GPT 기반 방법을 적용하여 개체명 정규화를 수행하였다(Fig. 2 (e-f)). GPT 기반으로 일치 여부를 판별하기 위해 아래와 같은 프롬프트를 설계하였다(Fig. 4). MENTION_A는 텍스트 내에서 인식된 처방 멘션 중 하나이며, MENTION_B는 또 다른 인식된 처방 멘션이다. TEXT는 판별을 위해 제공된 텍스트로 논문 제목과 초록을 포함한다. ChatGPT는 주어진 질문에 대해 간단히 "yes" 또는 "no"로 답변한다. OpenAI의 API 서비스는 파이썬 openai 패키지를 이용해 호출하였다.

> Do [MENTION_A] and [MENTION_B] in the given text designate the same concept? Simply answer "yes" or "no".
>
> [TEXT]

Fig. 4. the prompt designed to determine the same formula given two mentions

III. Results

1. Descriptive analysis of the data

Table 2는 각 처방별 데이터의 분포를 나타낸다. 740개의 초록 중 684개의 문헌은 단일 처방만을 다루었다. 740개의 문헌은 대부분 특정 질병에 대한 특정 처방의 효능을 다룬다.

Table 3은 처방 멘션별 데이터의 분포를 보여준다. 처방 멘션이 등장하는 초록 중 115개의 초록은 단일 처방 멘션만 포함하고 있어 본 연구에서 제외되었다. 가장 많은 경우는 2개의 구분되는 멘션이 등장한 380개의 문헌이었다. 본 논문의 2장에서 설명한 것처럼, 멘션이 2개 이상인 초록에서는 이항 조합으로 데이터를 생성하였다. 740개의 문헌 중 115개를 제외한 625개의 문헌에서 총 1,486개의 데이터를 사용하였다.

Table 2. The numbers of formula entities and corresponding articles

# formulas	1	2	3	4	5	6	7	8	9
# articles	684	34	11	6	3	0	1	0	1

Table 3. The numbers of formula mentions and corresponding data for training and testing

# mentions	1	2	3	4	5	6	7	8	9
# combinations	-	1	3	6	10	15	21	28	36
# articles	115	380	180	45	10	6	2	1	1
# data	-	380	540	270	100	90	42	28	36

2. Performance comparison

Table 4는 지도 기반 분류 모델의 Precision, Recall, F1-score 결과를 보여준다. Precision은 RF, XGB, cvcsSVM 순으로 높았다. 반면, Recall은 RF가 가장 높았고, XGB와 cvcsSVM이 유사한 결과를 보였다. F1-score는 cvcsSVM, XGB, RF 순으로 높았다.

Table 4. The performance evaluation of RF, XGB, and cvcsSVM

	RF	XGB	cvcsSVM
#Rows	1,486		
#Pos	1,166		
#Neg	320		
#TP	1164	1151	1154
#FP	208	124	108
#TN	2	15	12
#FN	112	196	212
Precision	0.848	0.902	0.914
Recall	0.998	0.987	0.989
F1-score	0.917	0.943	0.950

Table 5은 앞에서 가장 높은 F1-Score를 보인 cvcsSVM과 GPT에 기반한 GPT-4o mini, GPT-4o의 사례에 대한 Precision, Recall, F1-score를 보여준다. Precision은 cvcsSVM은 0.914로 가장 낮았고 GPT-4o가 0.996으로 가장 높았다. Recall은 cvcsSVM이 0.989로 가장 높았고, GPT-4o mini와 GPT-4o는 약 0.941로 유사했다. F1-score는 cvcsSVM, GPT-4o mini, GPT-4o가 각각 0.955, 0.948, 0.968로 GPT-4o가 가장 높았고, GPT-4o mini가 가장 낮았다. cvcsSVM이 GPT-4o mini보다 높은 F1-score를 보였으나, 이는 Positive 데이터의 개수가 1,486개 중 1,166개로 78.4%를 차지하므로, cvcsSVM의 높은 Recall 때문으로 보인다.

앞선 성능 평가 결과에서는 Positive 데이터의 비중이 Negative 데이터보다 많았는데, 이는 Table 2에서 확인할 수 있듯이 684개의 문헌에서 단일 처방 개체만 존재하고, 이 데이터들이 모두 Positive로 분류되었기 때문이다. 데이터셋에서 Positive 데이터의 비중이 성능에 미치는 편향을 확인하기 위해, 서로 다른 처방 개체가 등장하는 경우로만 한정하여 성능을 재평가하였다. 예를 들어, 설사(Diarrhea)에 대한 연구[7]에서 2개의 멘션이 등장하지만 Shengjiang Xiexin Decoction과 SXD가 생강사심탕이라는 동일한 개체에 대한 것이므로 제외하였다. 반면, 다낭성 난소 증후군에 대한 연구[8]에서는 Jia-wei-xiao-yao-san과 Si-wu-tang과 각각 가미소요산 및 사물탕이라는 서로 다른 처방 개체에 대한 멘션이므로 분석에 포함되었다.

Table 5. The performance evaluation of cvcsSVM, GPT-4o mini, and GPT-4o

	cvcsSVM						GPT-4o min	GPT-4o
	F1	F2	F3	F4	F5	Total		
#Rows	306	322	274	319	265	1,486		
#Pos	230	248	225	240	223	1,166		
#Neg	76	74	49	79	42	320		
#TP	226	246	221	239	222	1,154	1,097	1,095
#FP	22	24	21	13	28	108	51	4
#TN	54	50	28	66	14	212	269	316
#FN	4	2	4	1	1	12	69	71
Precision	0.911	0.911	0.913	0.948	0.888	0.914	0.955	0.996
Recall	0.982	0.991	0.982	0.995	0.995	0.989	0.941	0.941
F1-score	0.945	0.949	0.946	0.971	0.938	0.950	0.948	0.968

서로 다른 처방 개체가 등장하는 문헌들로만 한정했을 때의 결과는 Table 6에 나타나 있다. 이번 분석에서는 Positive 데이터보다 Negative 데이터가 더 많아졌으며, Precision은 GPT-4o가 가장 높았고, cvcsSVM은 가장 낮았다. Recall 역시 GPT-4o가 가장 높았으며, GPT-4o mini는 가장 낮았다. F1-score 역시 GPT-4o가 가장 높았고, cvcsSVM이 가장 낮았다. GPT-4o는 Precision, Recall, F1-score 모두 0.9 이상의 성능을 기록하였다. cvcsSVM은 FP가 가장 많았고, TN이 가장 적었으며, 이로 인해 GPT 기반 모델들에 비해 서로 다른 처방 개체 멘션 쌍을 구분하는 데 어려움을 겪는 것으로 보였다. 반면, GPT-4o는 대부분의 서로 다른 개념을 정확히 구분할 수 있었다.

Table 6. The performance evaluation of cvcsSVM, GPT-4o mini, and GPT-4o in cases with more than two distinguished formula entities in the text

	cvcsSVM	GPT-4min	GPT-4o
#Rows	398		
#Pos	78		
#Neg	320		
#TP	68	63	72
#FP	108	51	4
#TN	212	269	316
#FN	10	15	6
Precision	0.386	0.552	0.947
Recall	0.871	0.807	0.923
F1-score	0.535	0.656	0.935

IV. Discussions

우리는 한의 처방의 질병 효능을 다룬 문헌에서 한의 처방 멘션들이 동일한 개체를 가리키는지 여부를 판단하는 자동화 기술을 연구하였다.

비록 양방 의학에 비해 한의학 텍스트를 대상으로 한 정보 추출 기술이 많지는 않지만, 관련 연구들이 점차 보고

되고 있다. 예를 들어, 중의학(Traditional Chinese Medicine)에서 '임상 증상', '증후', '질병', '치료법', '약재 명칭'에 대한 개체명 인식 연구가 있다[9]. 이 연구에서는 단어 수준의 BiLSTM-CRF, BERT-CRF, RoBERTa-c 모델을 적용해 90% 이상의 F1-score를 달성했으나, 처방 개체는 포함되지 않았다. 이에 반해, 한의학에서 처방 명칭의 정규화를 다룬 유의미한 연구는 거의 없는 실정이므로, 본 연구의 처방 명칭 정규화 시도는 매우 의미 있는 연구라 할 수 있다.

개체명 인식에 비해 개체명 정규화에 대한 연구는 많지 않다. 한 연구[10]에서는 질병명과 식물명에 대한 개체명 인식된 멘션과 컨셉을 연결하기 위해 벡터 공간에서 멘션을 표현하여 정규화하는 방법을 제안했다. 약어 처리를 위해 기존 사전을 이용해 멘션과 컨셉 매핑의 성능을 향상시켰고, 약어 처리로 정보 추출 성능을 높인 바 있다. 그러나 한의 처방 명칭은 국가마다 표기가 다르며, 한자를 영어로 음차하면서 생기는 변동성으로 인해 처방에 대한 약어 사전을 구축하는 것은 한계가 있다. 이에 따라 우리는 주어진 텍스트에 기반하여 멘션 간 동일성을 판단하는 방법을 택했다.

참고할 수 있는 기존 기술이 부족했기 때문에, 우리는 BERT의 문맥 벡터(Context vector)를 활용한 알고리즘을 직접 개발하였으며, 최근 발전한 대규모 언어 모델(LLM)을 활용한 방법도 실험적으로 적용해보았다.

연구에 사용된 1,486개의 데이터에서 Positive 데이터가 1,166개, Negative 데이터는 320개로 Positive 데이터가 대부분을 차지하였다(Table 5). 이는 대부분의 문헌이 단일 처방의 특정 질병에 대한 효능을 다루고 있기 때문에, 동일 처방을 지칭하는 멘션들이 많았기 때문이다. 모든 데이터를 사용했을 경우, 모든 방법이 0.9 이상의 F1-score를 기록했다. 그러나 구분되는 처방이 2개 이상 포함된 경우에는 Negative 데이터가 Positive 데이터를 초과하였고, 이때는 GPT-4o만이 모든 지표에서 0.9 이상의 성능을 기록하였다(Table 6). 즉, 텍스트 내에 여러 처방이 존재할 경우, cvcsSVM은 GPT-4o 및 GPT-4o mini보다 많은 False positive를 보였다. 이는 최근 연구[11]에서도 ChatGPT의 추론 능력이 BERT 모델을 크게 능가한다고 보고된 것과 일치한다.

우리는 문맥벡터와 문자 유사도에 기반한 분류 방법을 직접 고안하였다. BERT는 Transformer 아키텍처의 Encoder를 기반으로하여, 텍스트에서 일부 단어를 마스킹(making) 한 후, 마스킹된 단어를 예측하는 방식으로 학습된다. 이때 각 단어 앞뒤의 문맥을 동시에 고려하여 처리한다. 따라서 멘션에 대응하는 임베딩 벡터는 텍스트에 대한 양방향성의 문맥정보를 포함하고 멘션의 알파벳이 다르더라도 문맥벡터의 유사성 정보를 활용하고자 하는게 이 모델의 주요한 동기였다. 분류 모델로는 Random forest, XGBoost, SVM를 사용했다. 초기에는 두 멘션의 문맥벡터만 분류모델의 입력으로 활용했으나, 두 멘션의 문맥벡터에 더해서, 두 멘션 문자열의 유사도를 같이 넣었을 때 전반적으로 F1-Score의 향상을 확인할 수 있었다. SVM이 가장 높은 성능을 보였는데, 모델의 특징의 개수가 1,537개로 고차원의 공간이므로, 고차원 공간에서 분류 작업을 효율적으로 처리하는 SVM이 성능이 높은 것으로 보인다.

GPT-4는 Transformer의 디코더(Decoder)를 기반으로 하며, 다음 단어를 예측하는 방식으로 학습되어 생성형 작업에 적합하다. 본 연구의 목적은 개체명 정규화였으므로, GPT-4에 간단한 yes/no 답변을 요구하는 프롬프트를 설계하였다. GPT 기반 모델은 cvcsSVM보다 일관성 있게 낮은 FP를 보였다. GPT는 방대한 텍스트 데이터를 기반으로 학습되었으며, 문맥과 논리적 연결을 처리하는 능력이 뛰어나 파인튜닝 없이도 강력한 성능을 발휘할 수 있다. GPT-4o와 GPT-4o mini의 성능 차이는 파라미터 수의 차이에서 기인할 가능성이 있다.

최근 LLM의 급격한 발전으로, 요약 및 질의응답을 넘어서 텍스트 마이닝 연구에도 적용되고 있다. 한 연구[12]에서는 임상 개체명 인식에서 GPT-4가 GPT-3.5보다 우수한 성능을 보였지만, BioClinicalBERT보다는 성능이 낮았다고 한다. 개체명 정규화 작업에서는 아직 LLM이 널리 적용되지 않고 있지만, GPT-4o는 일관성 있는 높은 성능을 보여 활용 가치가 높다. 다만, GPT-4o는 블랙박스 특성으로 인해 출력 결과를 명확히 설명하기 어렵고, 웹 API 호출 방식으로 처리 속도가 제한적이며 비용 문제도 있다.

따라서 전통적인 분류 모델 접근 방식은 독자적인 서비스로 구현할 수 있고, GPT의 활용과 별개로 발전시킬 가치가 있다.

V. Conclusions

우리는 한의 처방 멘션의 동일 개체 여부를 판별하기 위한 기술을 개발하기 위해 임베딩 벡터를 특징으로 하는 지도 학습 모델과 LLM을 활용한 프롬프트 기반 질의 방식을 시도하였다. 그 결과, LLM을 활용한 방법이 더 높은 성능을 보였으며, 본 연구는 한의학 도메인의 개체명 정규화에 유용하게 활용될 가능성이 있다.

ACKNOWLEDGEMENT

This work was funded by the Korea Institute of Oriental Medicine (No. KSN1824130 and No. KSN1923111).

REFERENCES

[1] H. Lee, et al., "Treatment of nausea and vomiting associated with cerebellar infarction using the traditional herbal medicines banhabaekchulcheonma-tang and oryeong-san: Two case reports (CARE-complaint)," EXPLORE, Vol. 19, No. 1, pp. 141-146, January 2023. DOI: 10.1016/j.explore.2021.11.011.

[2] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805

[3] A. Radford, et al., "Improving language understanding by generative pre-training," 2018. https://api.semanticscholar.org/CorpusID:49313245

[4] M. H. Kim, et al., "A multicenter study on the efficacy and safety of So-Cheong-Ryong-Tang for perennial allergic rhinitis," Complementary Therapies in Medicine, Vol. 45, pp. 50-56, August 2019. DOI: 10.1016/j.ctim.2019.05.018.

[5] T. H. Lan, et al., "Systems pharmacology dissection of traditional Chinese medicine Wen-Dan decoction for treatment of cardiovascular diseases," Evidence Based Complementary and Alternative Medicine, Vol. 2018, No. 1, pp. 5170854, May 2018. DOI: 10.1155/2018/5170854

[6] Y. Wang, J. Qin, and W. Wang, "Efficient approximate entity matching using Jaro-Winkler distance," in International Conference on Web Information Systems Engineering, Cham: Springer International Publishing, pp. 231-239, October 2017.

[7] C. Deng, B. Deng, L. Jia, H. Tan, P. Zhang, S. Liu, Y. Zhang, A. Song, and L. Pan, "Preventive effects of a Chinese herbal formula, Shengjiang Xiexin decoction, on irinotecan-induced delayed-onset diarrhea in rats," Evidence-Based Complementary and Alternative Medicine, Vol. 2017, No. 1, pp. 7350251, January 2017.

[8] M.-J. Lin, H.-W. Chen, P.-H. Liu, W.-J. Cheng, S.-L. Kuo, and M.-C. Kao, "The prescription patterns of traditional Chinese medicine for women with polycystic ovary syndrome in Taiwan: a nationwide population-based study," Medicine, Vol. 98, No. 24, pp. e15890, June 2019. DOI: 10.1097/MD.0000000000015890.

[9] Z. Liu, C. Luo, Z. Zheng, Y. Li, D. Fu, X. Yu, and J. Zhao, "TCMNER and PubMed: A novel Chinese character-level-based model and a dataset for TCM named entity recognition," Journal of Healthcare Engineering, Vol. 2021, No. 1, pp. 3544281, August 2021. DOI: 10.1155/2021/3544281.

[10] H. Cho, W. Choi, and H. Lee, "A method for named entity normalization in biomedical articles: application to diseases and plants," BMC Bioinformatics, Vol. 18, pp. 1-12, October 2017. DOI: 10.1186/s12859-017-1975-0.

[11] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT," arXiv preprint, arXiv:2302.10198, 2023.

[12] Y. Hu, Q. Chen, J. Du, X. Peng, V. K. K. Keloth, X. Zuo, Y. Zhou, et al., "Improving large language models for clinical named entity recognition via prompt engineering," Journal of the American Medical Informatics Association, September 2024. DOI: 10.1093/jamia/ocad259.

Authors

Ho Jang received his Ph.D. degree from School of Electrical Engineering and Computer Science at Gwangju Institute of Science and Technology in 2017. Dr. Jang has been working for Korea Institute of Oriental Medicine (KIOM) since 2018. He is currently a senior researcher in KM Data Division at KIOM. His research interests include computational biology and machine learning.