대한예방한의학회지 제17권 제2호(2013년 8월)
Kor. J. Oriental Preventive Medical Society 2013 ; 17(2) : 179-187

한의학 증상용어의 형태소 분석을 위한 자연어 표기 분석

김혜은[1] · 성호경[1] · 엄동명[3, 4] · 이충열[5] · 이병욱[1, 2]*
[1] 부산대학교 한의학전문대학원 인문사회의학부, [2] (주)피엔유 동제 메디칼
[3] 원광대학교 한의과대학 원전학교실, [4]한국전통의학연구소
[5] 가천대학교 한의과대학 생리학교실

Analyzing Morpheme of the Natural Language to Express the Symptoms of Korean Medicine

Kim Hye-Eun[1], Sung Ho-Kyung[1], Eom Dong-Myung[3, 4]
Lee Choong-Yeol[5] & Lee Byung-Wook[1, 2]*
[1] Division of Humanities and Social Medicine, School of Korean Medicine, Pusan National University
[2] PNU Dongje Medical Co.,Ltd.
[3] Dept. of Classics, College of Korean Medicine, Wonkwang University
[4] Research Center of Traditional Korean Medicine, Wonkwang University
[5] Dept. of Physiology, College of Oriental Medicine, Gachon University

Abstract

Objectives : In many cases, patient's symptoms have been recorded on EMR in natural language instead of medical terminologies. It is possible to build a database by analyzing the symptoms of Korean Medicine(KM) that indicates patient's symptoms in natural language. Using the database, when doctors record patient's symptoms on EMR in natural language, conversely it'll be also possible to extract the symptoms of KM from those natural language. The database will enhance the value of EMR as a medical data.

Methods : In this study, we aimed to make data structure of the terminologies that represent the symptoms of KM. The data structure is combinations of smallest unit in natural language. We made the database by analyzing morpheme of the natural language to express the symptoms of KM.

Results & Conclusions : By classifying the natural language in 15 features, we made the structure of concept and the data available for morphological analysis.

· 접수 : 2013년 7월 29일 · 수정접수 : 2013년 8월 22일 · 채택 : 2013년 8월 23일
* 교신저자 : 이병욱, 경남 양산시 물금읍 범어리 부산대학교 한의학전문대학원 313호
전화 : 051-510-8451, 전자우편 : omis@pusan.ac.kr

Key words : Symptom, Structure of concept, Natural Language Processing

I. 서 론

지금까지 한의학 분야에서 정보화와 관련된 연구는 주로 '용어의 표준화'를 시작으로 '분류체계', '용어관리 시스템', '한의학 온톨로지' 등의 영역에 이르기까지 지속적인 연구가 진행되어 왔다.

그러나 지금까지 이러한 연구는 문헌에 등장하는 수 많은 용어를 사용한 표준화 연구 혹은 온톨로지 연구였기 때문에 표준선정의 측면이나 온톨로지 방식을 활용한 지식체계화에 있어서 적절한 연구였음에도 불구하고 실제 임상에서 사용하는 용어와 차이가 있었기 때문에 실용적인 측면에서는 아직 부족한 점이 있다.

이는 표준 용어 선정을 위한 연구의 시작시에 진료부에 기록된 용어도 수집을 시도하였으나, 문헌으로부터 수집된 용어에 비하여 개념을 확정 짓기가 어려웠으며, 명료한 명사형이 아닌 자연어 형태의 용어가 많았기 때문에 표준 용어로 선정될 가능성이 낮았다[1].

이로 인하여 실제 진료부 상에 표현되는 내용을 바탕으로 용어관리 시스템에 기재된 용어리스트와 비교시 차이가 많았다는 최근의 연구 결과[1]가 있었으며, 그 연구의 결론에서 실용적인 측면에서의 자연어처리[44]를 위해서는 일반용어 및 서술어 부분의 연구가 계속되어야 한다고 하였다. 뿐만 아니라 한의학계에서는 아직 자연어처리에 관련한 연구 결과가 발표된 것이 없지만, 한의학 이외의 분야에서는 국내에서도 많은 연구 결과[2), 3)]가 발표되고 있다.

이에 본 연구에서는 문어 기반 자연어처리의 과정 중 첫 번째 단계인 형태소 분석에 관한 연구를 진행하고자 한다(그림1)[3]. 이러한 형태소 분석은 장문수의 연구에 따르면, "입력되는 텍

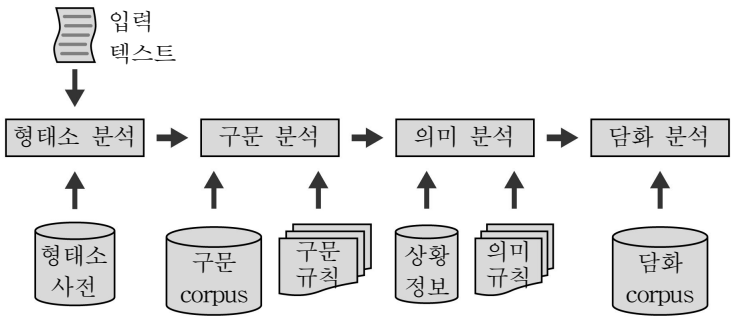

Fig 1. Natural Language Processing

44) 네이버 지식백과. http://terms.naver.com/entry.nhn? cid=391 &docId=863247&mobile&categoryId=391 "컴퓨터를 이용하여 사람 언어의 이해, 생성 및 분석을 다루는 인공 지능 기술. 자연어 이해는 일상 생활 언어를 형태 분석, 의미 분석, 대화 분석 등을 통하여 컴퓨터가 처리할 수 있도록 변환시키는 작업이며, 자연어 생성은 컴퓨터가 처리한 결과물을 사람의 편의성에 입각하여 텍스트, 음성, 그래픽 등을 생성하는 작업이다. 문서 처리, 색인 작성, 언어 번역, 질문 응답 등 많은 응용 분야에서 활용한다." 2013.07.10.

스트를 단어나 어절 단위로 나누고, 각 단어에 대해서 여러 가지 형태소 사전을 참조하여 형태소를 분리한다. 따라서 형태소 분석 단계에서는 문장이나 구와 같은 큰 범위의 언어 단위는 의미가 없기 때문에 구절이나 심지어 비문을 입력하더라도 정상적인 형태소 분석이 이루어질 수 있다. 형태소 분석의 이러한 특징은 초기의 정보 검색 엔진에서 키워드를 추출하는 데에 유용하게 사용되었다"[3]고 하였다.

수집된 한의학용어 중 증상에 해당하는 용어의 의미를 자연어로 풀이하여 형태소 분석을 하였으며, 자연어 설명에 들어가는 의미들의 조합과 증상용어의 관계에 대한 구조를 분석하여, 한의학 전자진료부 작성을 위한 용어시스템의 자연어 표현 데이터 구조에 대해서 연구하고자 한다.

II. 연구방법

1. 연구대상

본 연구에서는 '의무기록용 표준 한의학 임상용어(병증편)'의 연구 과정에서 수집한 한의학회 소속의 각 분과학회에서 제출한 병증용어 관련 목록 6,900여개를 이용하였고, 그 중 출전과 사용빈도를 고려하여 표준 용어 후보에 선정된 증상 분류의 용어 1,269개에 대하여 연구를 진행하였다.

1) 증상 분류 용어의 추출

병증 분류에 속한 용어는 '변증', '병명', '증상', '변증 + 증상', '기타정보 + 증상' 형태의 용어가 선정되어 있었다. 이 중 본 연구에서는 환자에게 나타나는 하나의 현상에 해당하는 내용만 증상으로 추출하였다. 또한 '脚跟痛', '脚膝疼痛' 같은 증상 용어는 다시 '脚痛', '跟痛', '脚疼痛' '膝疼痛'으로 분석하여 '脚跟痛'과 '脚膝疼痛'은 분석하지 않고 '脚痛', '跟痛', '脚疼痛' '膝疼痛'으로 세분화된 증상에 대해서 본 연구의 참여 인력이 8개월 동안 수작업과 논의를 통하여 분석을 시도하였다.

2) 데이터 테이블의 구조

(1) 한자 테이블

의무기록용 한의학 임상용어 중 병증 분류에

Fig 2. Hanja Table

Fig 3. Natural Language Table

속한 용어의 ID, 한자 표기, 간단한 설명, 연구제외 여부, 제외 이유를 기록한 데이터 테이블이다.

(2) 자연어 테이블

한자 테이블의 용어 중 증상에 속한 용어에 대하여 전문용어를 고려하지 않고 최대한 일반인 입장에서 표현한다는 가정하에 최대한 풀어서 표기한 내용을 기록한 데이터 테이블이며, 한자 테이블과 ID 필드를 통하여 연결된다.

(3) 개념분리 테이블

증상에 해당하는 용어를 자연어로 표기했을 때 자연어에 포함된 내용을 최소한의 의미단위로 쪼개어 분석한 결과를 입력하는 데이터 테이블이다.

이 테이블은 증상을 표현하는 데 사용된 문장의 의미를 최대한 살리기 위하여 여러 필드를 만들어 작성하였으며, 초기에 구성된 필드의 구조를 작업 과정중에 수시로 추가하는 형태로 연구를 진행하였다.

Fig 4. Properties of Concept Table

(4) 테이블 사이의 관계

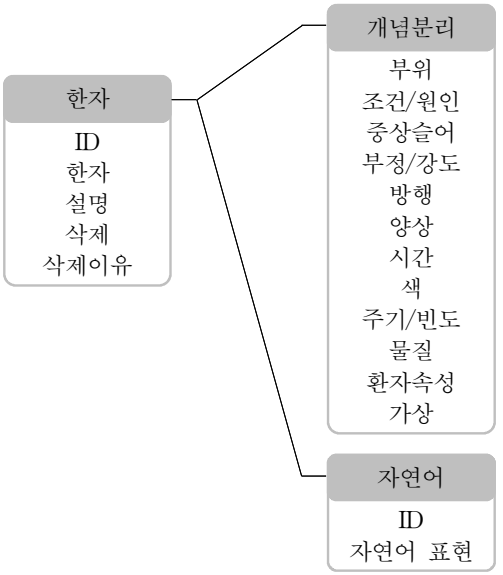

Fig 5. Relations of Tables

Ⅲ. 본론

1. 자연어 개념 분리를 위한 속성

본 연구에서는 증상을 자연현상의 하나로 보았다. 자연현상은 그 현상이 나타나는 공간과 시간이 있으며, 현상 자체를 기술하는 정보가 필요하고 현상이 구현되어 나타나는 공간속에 존재하는 실체가 있어야 한다. 때문에 본 연구 과정에서는 증상 표현과 관련된 자연어 문장을 공간위치, 시간빈도, 행위현상, 실체로 구분하여 분석하고자 하였으며, 다시 다음과 같이 세분화 하였다.

1) 공간위치

(1) 부위

증상이 나타나는 신체 부위 정보.

예) 등(背), 다리(脚), 입술(脣) 등.

(2) 방향(위치)

증상이 나타나는 신체 부위의 추가적인 방향 정보.

예) 한쪽, 주변, 위, 외측 등.

(3) 방향(이동)

증상이 나타나는 현상의 방향에 대한 정보.

예) 한쪽으로, 뒤로, 앞으로, 외측으로 등.

2) 시간빈도

(1) 시간

증상 표현에 사용되는 시간 정보.

예) 후, 저녁, 자는 동안, 생리후 등.

(2) 주기(빈도)

증상 표현에 사용되는 발생 주기 혹은 빈도에 관한 정보.

예) 항상, 오래, 빈번하게, 두달에 한번 등.

3) 행위현상

(1) 양상

증상 표현에 사용되는 정도 혹은 모양에 대한 정보.

예) 화끈하다, 찌는 듯, 쥐어짜는 듯 등.

(2) 색

증상 표현에 사용되는 색깔 정보.

예) 홍색, 파란색, 탁, 창백 등.

(3) 기능(현상)

증상 표현에 포함되는 인체의 고유 기능 혹은 자연 현상.

예) 힘, 호흡, 트림, 출산 등.

(4) 조건(또는 원인)

증상이 나타나거나 심해지는 조건에 해당하는 정보.

예) 누르다(按), 임신, 놀라다 등.

(5) 서술어

증상 표현에 사용되는 서술어 정보.

예) 아프다, 저리다, 붓다 등.

(6) 부정

증상 표현에 사용되는 서술어 부분을 부정하는 표현에 관한 정보.

예) 못, 안, 불(不) 등.

(7) 강도

증상 표현에 사용되는 서술어 부분의 행위 강도에 관한 정보.

예) 잘, 약간, 조금 등.

4) 실체

(1) 물질

증상 표현에 사용되는 인체 물질 혹은 일반 사물에서 기원한 물질 정보.

예) 소변, 피, 가래, 땀, 콧물 등.

(2) 환자속성

특정 환자에게만 나타나는 증상에 대한 환자 속성 정보.

예) 젖먹이, 임신부, 여성, 아이.

(3) 추상

증상 표현에 사용되는 정보 중 환자가 명확하게 인지하지 못하는 것, 혹은 다양한 형태의 대상으로 하나로 표현하기 어려운 대상 정보.

예) 헛 것, 죽은 사람, 이상한 소리, 무엇인가 등.

2. 속성 조합 유형 분석 결과

본 연구과정에서 사용한 증상 용어에서는 앞의 15개 속성의 값에 대한 조합을 분석한 결과 92종의 조합 형태가 나타났으며, 그 중 상위 10개의 조합의 형태에 대하여 기술하면 아래와 같다.

1) 부위+서술어의 조합

증상을 표현하는 자연어의 표현을 분석한 결과 가장 많은 빈도로 나타난 조합의 형태이다. 주로 증상이 나타나는 신체 부위를 나타내는 개념용어와 증상을 서술하는 서술어 부분으로 이루어져 있다.

예) 口苦: 입이 쓰다.

2) 서술어+기능(현상)의 조합

증상을 자연어로 표현하는 가운데 인체의 고유의 기능현상을 의미하는 용어와 서술어의 조합으로 이루어진 경우이다.

예) 神昏: 정신이 흐리다.

3) 부위+서술어+기능(현상)의 조합

증상을 자연어로 표현하는 가운데 인체 부위와 인체 고유의 기능현상 및 서술어의 조합으로 표현되는 경우이다.

예) 項軟: 목에 힘이 없다.

4) 부위+서술어+방향(위치)의 조합

증상을 표현하는 데 인체 부위와 방향 그리고 서술어가 조합되어 표현되는 경우이다.

예) 項痛: 목 뒤가 아프다.

5) 서술어

증상을 표현하는 데 서술어만으로 구성된 경우이다.

예) 吐: 토하다. 怒: 화나다

6) 부위+서술어+물질의 조합

증상을 표현하는 데 인체의 부위와 서술어 그리고 물질을 뜻하는 의미가 조합되어 표현되는 경우이다.

예) 頭汗: 머리에 땀이 나다.

7) 서술어+물질

증상을 표현하는 데 물질과 서술어가 조합되어 표현되는 경우이다.

예) 喀痰: 가래가 나온다.

8) 부위+서술어+부정

증상을 자연어로 표현하는 가운데 인체의 부위와 서술어 그리고 不定 또는 서술어의 强度를 표현하는 의미의 조합으로 이루어진 경우이다.

예) 口不開: 입을 못 열다.

9) 부위+서술어+양상

증상을 자연어로 표현하는 가운데 인체의 부위와 서술어 그리고 서술어로 표현되는 현상에 대한 양상이 추가적으로 함께 기술되는 경우이다.

예) 毛折: 머리카락이 쉽게 부서진다.

10) 서술어+강도

증상을 자연어로 표현하는 가운데 서술어와 不定 또는 서술어의 强度를 표현하는 의미의 조합으로 이루어진 경우이다.

예) 健忘: 잘 잊는다.

3. 문제점

1) 개인에 따른 다양한 표현

증상을 표현하는데 있어서 실제 자연어의 활용의 측면에서 본다면, 본 연구에서 분류한 속성의 조합은 동일한 의미를 전달하는 과정에 얼마든지 다른 형태의 표현이 있을 수 있으며, 이러한 표현은 개인의 취향에 따라 데이터 값이 다양하게 입력될 가능성이 매우 높다. 이러한 경우에 대한 예를 들면 다음과 같다.

(1) 물질 + 서술어 = 기능 + 서술어, 서술어

자연어에서 서로 다른 표현 중 물질과 서술어의 조합이 기능과 서술어의 조합과 동일한 의미를 가지는 경우가 있다.

예) 出産: 자식을 낳다(물질+서술어), 출산을 하다(기능+서술어), 출산하다(서술어).

(2) 색 + 서술어 = 서술어

증상을 표현하는 과정에서 색과 서술어의 조합형태인 '붉은색 이다'로 표현하거나 서술어 단독 형태인 '붉다'로 표현할 수 있다.

예) 面白: 얼굴이 흰색이다.(부위+색+서술어), 얼굴이 희다.(부위+서술어)

2) 개인에 따른 속성 판단의 차이

이는 자연어 문장의 표현은 동일하지만, 개념을 분리하여 속성에 데이터를 넣는 과정에서 서로 다른 속성에 배치하는 경우에 해당된다.

예를 들면, '갑자기 아프다'라는 표현에 대하여 연구자에 따라서 '갑자기'를 '시간'의 속성으로 생각한 경우와 '양상'으로 생각하는 경우가 있다. 또한 감각과 관련된 표현에서 '입이 쓰다'는 의미의 표현중 '입에 쓴맛이 난다'라고 표현

하는 경우 '쓴맛'을 '현상'으로 넣는 경우와 '쓴맛(의 느낌)'으로 해석하여 '양상'으로 넣는 경우가 있었다.

3) 문맥에 따른 어휘의 의미 변화

이는 자연어 문장의 표현은 동일하지만 함께 조합되는 다른 어휘에 따라서 의미가 변하는 경우이며, 서술어 속성에 포함되는 경우와 인체 부위의 속성인 경우가 있었다.

(1) 나다: 열이 나다 --> 현상이 나타났다.
피가 나다 --> 바깥으로 배출되다.
(2) 있다: 열이 있다 --> 현상이 나타났다.
목에 무엇인가 있다 --> 有
(3) 쓰다: 입이 쓰다 --> 맛이 쓰다
손을 쓰지 못 하다 --> 사용하다
(4) 시원하다: 시원한 것을 좋아하다 --> 차갑다
소변이 시원하지 않다 --> 快
(5) 보다: 눈이 보이지 않는다 --> 視
대변을 보지 못 하다 -- 배출하다
(6) 목: 목이 아프다. --> 목구멍
목이 쉬다 --> 목소리
뒷 목이 뻣뻣하다 --> neck

4) 하나의 어휘가 여러 어휘의 조합과 의미가 같은 경우

이는 자연어의 표현에서 흔하게 발생되는 경우로 한자어와 순수 한국어를 함께 사용하는 언어환경에서 기인한다.

예) 식사하다 --> 밥을 먹다
內踝 --> 안쪽 복사뼈

5) 1) + 4)의 유형

이는 개인의 언어 습관에 따라서 증상의 표현에 다른 증상 용어를 포함하여 간단하게 표현하는 경우와 한자어를 사용하지 않고 순수 우리말을 사용하여 표현하는 경우에 나타난다.

예) 難産: 출산이 어렵다, 아이 낳기가 어렵다.

4. 앞으로의 연구 방향

이상의 과제는 본 연구를 진행하는 과정에서 구축한 단일테이블 형태의 데이터 구조에서는 해결하기 어려운 문제이다. 따라서 본 연구결과를 바탕으로 의미관계를 파악하여 연구자 개인의 특성이나 어휘가 가지는 다양한 의미에 대한 유연한 대응이 가능한 새로운 데이터 구조가 필요하다.

앞으로의 연구에서는 데이터 테이블에서 어휘의 분화관계를 표현할 수 있는 구조를 만들어 구축하여 해결해야 할 것이다. 예를 들면 '出産'이라는 개념에 'A001'이라는 코드를 부여한다면, 'B001', 'B002', 'B003'의 동의어 코드를 여러 개 생성하고 각각의 동의어 코드에 대하여 (B001, 자식), (B001, 낳다) 등의 조합으로 표현하여 'A001'에 도달하는 방법을 추구해야 할 것이다.

또한 하나의 테이블의 하나의 레코드를 이용하여 개념을 설명하지 않고, 여러개의 레코드 조합으로 표현하는 방식을 추구한다면, 개인에 따라 속성의 판단이 다르거나, 하나의 어휘가 여러 어휘의 조합으로 표현되더라도 속성값만을 참조하여 의미를 찾아가는 방식을 취한다면 큰 문제없이 표현이 가능할 것이다. 즉 '갑자기 아프다'를 하나의 레코드로 표현할 때에는 시간 속성에 '갑자기'를 넣을 경우와 양상 '속성'에 '갑자기'를 넣을 경우에 검색하는 필드가 달라지므로 결과값이 다르게 출력될 수 있었다. 그러나 문장을 표현하는 테이블의 구조에서 각각의 속성을 필드로 모두 만드는 것이 아니라 '개념', '필드속성', '값'이라는 오직 세 개의 필드만을 가지고 테이블을 구축하여 '갑자기 아프다'를

급통(개념)-시간(필드속성)-갑자기(값), 급통(개념)-서술어(필드속성)-아프다(값)

또는

급통(개념)-양상(필드속성)-갑자기(값), 급통(개념)-서술어(필드속성)-아프다(값)으로 표현한다고 할지라도 검색시에는 '급통'이라는 개념 필드와 '아프다'라는 값 필드만을 검색한다면 두 경우 모두 동일한 결과를 얻을 수 있을 것이다.

Ⅳ. 결론

본 연구에서는 자연어처리의 단계중 형태소 분석단계를 구현하기 위한 연구를 진행하였다. 이를 위하여 한의학용어 중 증상에 해당하는 용어의 의미를 자연어로 풀이하여 자연어 설명에 들어가는 의미들의 조합과 증상용어의 관계에 대한 구조를 분석하고 형태소 분석을 위한 DB를 제작하였으며, 그 결과 한의학 전자진료부 개발을 위한 용어시스템의 자연어 표현 데이터 구조에 대해서 연구하여 다음과 같은 결론을 얻었다.

1. 현재 본 연구 과정에서 증상 표현과 관련된 자연어 문장을 분석하여 공간위치, 시간빈도, 행위현상, 형태로 구분한 후 다시 부위, 방향(위치), 방향(이동), 시간, 주기(빈도), 양상, 색, 기능(현상), 조건, 서술어, 부정, 강도, 물질, 환자속성, 추상으로 세분화 하였다.

2. 하나의 테이블을 이용하여 증상에 대한 자연어 표현에 필요한 속성을 분석하면, 다음과 같은 문제가 발생할 수 있으며 차후 지속적인 연구를 통하여 개선되어야 할 것이다.

1) 증상에 대한 표현은 개인에 따라 다양한 속성의 어휘로 분석될 수 있다.
2) 개인에 따라 동일한 어휘의 속성을 달리 분석할 수 있다.
3) 동일한 어휘가 여러 가지 의미를 가지는 경우가 있다.
4) 하나의 어휘가 여러 어휘의 조합으로 표현되어도 의미는 같다.
5) 1) + 4)의 유형이 나타날 수 있다.

이상의 과제는 본 연구를 진행하는 과정에서 구축한 단일테이블 형태의 데이터 구조에서는 해결하기 어려운 문제이다. 따라서 본 연구결과를 바탕으로 의미관계를 파악하여 연구자 개인의 특성이나 어휘가 가지는 다양한 의미에 대한 유연한 대응이 가능한 새로운 데이터 구조가 추가적으로 연구되어야 할 것이며, 동일한 의미를 가지는 여러 속성들의 조합으로부터 대표증상용어를 연결하는 알고리즘 개발에 대한 추가적인 연구가 필요하다.

참고문헌

1. 이병욱, 백진웅, 김상균, 문헌과 진료기록부에 기록된 용어의 비교, 대한한의학원전학회지. 2011:24(1):1-14.
2. 박경미, 황규백, 자연어처리 기반 바이오 텍스트 마이닝 시스템, 정보과학회논문지. 2011:17(4):205-213.
3. 장문수, 구어에 대한 자연어처리의 문제점과 그 해결을 위한 스스테믹 기능문법적 접근, 한국어학. 2010:46:123-143.