한국어 의학 문서에 대한 영문 MeSH 키워드의 자동 부여 - 띄어쓰기 변이 처리 효과를 중심으로

이재성, 김미숙, 이영성
충북대학교 컴퓨터교육과, 교육대학원 정보컴퓨터 전공, 의학과
jasonl@cbu.ac.kr, htkms@naver.com, yslee@cbu.ac.kr

Automatic English MeSH keywords assignment to Korean medical documents - spacing variant effect

Jae Sung Lee, Mi Suk Kim, Young Sung Lee
Dept. of Computer Education, Chungbuk National University
Dept. of Health Informatics and Management, Chungbuk National University

요 약

본 논문에서는 한국어 의학 논문의 요약문으로부터 자동 영문 MeSH 키워드 제안 시스템을 소개하고, 띄어쓰기 변이(spacing variant) 문제를 해결할 수 있는 방법을 제안한다. 띄어쓰기 변이란 표준 한글 맞춤법에 비해 다르게 띄어쓰기된 것을 말한다. 이를 위해 시소러스에는 생성 가능한 모든 띄어쓰기 변이 대신에 최대 띄어쓰기 어구만을 저장하고, 문서에서 K-MeSH 용어를 찾기 위해 음절단위 부분문자열 검색을 사용한다. 이 방법으로 한국어 의학 논문의 요약문에서 K-MeSH 용어를 추출한 후, TF-IDF 순위 함수를 이용하여 상위 10위내의 키워드를 저자가 선정한 영문 키워드와 비교한 결과 58%가 일치하였다. 이는 기존 방법에 비해 42%정도의 시소러스 크기가 축소되었고, 상위 10위내에서 영문 MeSH 키워드 추천 재현률이 약 7.8% 증가한 것으로 효과적인 방법임을 보여주었다.

1. 소 개

MeSH(Medical Subject Headings)는 한 가지 개념에 같은 의미를 가지는 모든 용어의 집합이며, 개념의 계층 구조를 제공하는 시소러스이다[10]. MeSH의 주제어는 MEDLINE을 포함한 의학 문서 색인을 위해 사용된다. 일부 연구에서는 MeSH 용어를 사용하는 것이 검색 성능을 향상시킨다고 보여주었다[7, 12].

Korean MeSH(이하 K-MeSH)는 MeSH를 한글화한 것으로 MeSH 뿐만 아니라 한국어 동의어, 변이체 등을 포함한다. 이는 의학 용어에 대한 영어와 한국어의 표제어를 추천해 주는 시스템에 적용할 수 있다.

많은 한국어 의학 저널들은 MeSH 용어의 중요성을 고려하여 한국어 논문에서 영문 키워드를 선정할 때 MeSH 용어를 사용하도록 요구하고 있다. 따라서 K-MeSH는 한국어 의학 용어에 대한 정확한 MeSH 용어를 선정하는데 사용할 수 있다. 이를 위해 사용자가 검색을 원하는 용어를 보다 쉽고 정확하며 빠르게 찾을 수 있는 프로그램이 필요하다.

본 논문에서 구현한 K-MeSH 용어 추천 프로그램(K-MeSH Term Access Program : 이하 KTAP)은 사용자가 입력한 문서에서 K-MeSH 용어를 손쉽게 검색할 수 있도록 개발하였다. 그리고 입력한 의학 논문의 요약문에서 색인 가능한 영문 키워드를 자동으로 추천할 수 있도록 개발하였다. KTAP은 사용자가 입력한 의학 논문의 요약문을 받아 K-MeSH Tree에서 검색한 K-MeSH 또는 관련된 용어에 대한 해당 K-MeSH 표제어를 추출해 낸다. 이는 키워드 검색이 아닌 입력 문서에 대한 일괄적인 검색 방법으로, 사용자가 적절한 K-MeSH 용어를 검색하는 데 있어서 매우 편리한 기능이다. 또한 각 용어들은 중요 순위로 출력되기 때문에 사용자들은 색인을 위한 키워드를 손쉽게 선정할 수 있다.

K-MeSH의 필드에는 한국어 뿐만 아니라 영문 MeSH 용어들이 포함되어 있다는 것을 이용하여 KTAP은 한국어 의학 문서로부터 색인에 필요한 영문 MeSH 키워드를 추천할 수 있도록 개발하였다. 영문 의

학 문서에 대한 MeSH 키워드 추천 방법에 대한 연구는 이미 진행되어 왔고, MeSH 키워드를 사용하는 것이 검색 성능을 향상시킨다고 보여주었다[8, 6]. 또한 문장에서 추출한 일반 키워드와 MeSH 키워드를 혼합하여 관련도 순위 계산에 사용할 경우 검색 성능이 향상되었다[7, 12]. 특히 영어 문맥에서 약어, 축약어, 동의어, 파생적 변이와 같은 변이들에 대해 같은 단어로 처리할 경우 정보 검색의 성능이 향상되었다[5]. 그러나 이러한 연구 결과들은 한국어에 대해 참고할 수는 있으나, 언어의 차이 때문에 K-MeSH 용어 추천 시스템에는 직접적으로 적용 할 수 없다. 더욱이 한국어 문서로부터 다른 언어인 영문 MeSH 키워드를 추천하는 방법은 본 연구에서 새로 시도하는 것이다.

한국어 의학 용어들은 대체로 복합어로 이루어져 있다. 띄어쓰기 규칙에 의하면, 전문 용어인 의학 용어는 붙여 쓰는 것이 원칙이지만, 의미가 명확해 지도록 띄어 쓰는 것도 허용하고 있다. 또한 전문 용어에 대한 정의가 모호하여 일반 명사와 섞어 쓸 경우 어느 범위까지가 전문 용어인지를 구분하지 못하는 경우도 있다. 따라서 띄어쓰기는 저자에 따라 다를 수 있다. 이러한 일관성 없는 띄어쓰기 문제는 KTAP에서 K-MeSH 용어를 추출하는 데 있어서 어려움을 발생시킨다. 본 연구에서는 앞으로의 설명을 위해서 이를 띄어쓰기 변이(spacing variants)라 부르고, "띄어쓰기는 다르지만 표준어와 같은 순서의 구성요소를 가진 변이체"로 정의한다.

한국어 복합명사의 띄어쓰기 문제는 정보 검색에서 검색의 성능을 향상시키기 위해 많이 연구되어져 왔다[1, 4]. 기존 논문들에서는 주로 복합명사 띄어쓰기 변이를 통제되지 않은 단어로 정규화시켜 처리한다. 그러나 본 논문에서는 복합명사 띄어쓰기 변이를 통제된 K-MeSH 용어로 정규화시켜 처리한다.

다음 장에서 KTAP을 소개하고, 특히 띄어쓰기 변이에 초점을 맞추어 그 해결책을 제안한다. MeSH 용어 추천을 위해서 변이는 가능한 많이 찾아 내어 같은 뜻의 단어로 인식해야 한다. 본 연구에서는 한국어 의학 논문에서 사용된 변이의 추출 비율을 측정하였고, 저자가 선정한 키워드를 KTAP 시스템이 어느 정도 찾아 추천해 주는가를 비율로 계산하여 KTAP의 성능을 측정하였다.

2. 시스템 개요

KTAP은 웹인터페이스로 구현하였으며, 다수의 웹사용자들에 의한 동시 접속을 고려하여 빠른 서비스를 제공할 수 있도록 최적화하였다. 그림1은 처리 결과의 예를 보여주며 결과는 두 가지로 나타난다.

윗부분에는 검색된 K-MeSH 용어에 하이퍼링크 설정된 입력 문서가 나타난다. 텍스트내의 하이퍼링크 된 구(phrase)를 클릭하면 바로 K-MeSH 용어에 대한 정보를 얻을 수 있다. 이것은 사용자들이 문맥 내에서 표제어나 적절한 용어를 선택하는 데 있어서 매우 편리한 기능이다. 아랫부분에는 문서 내에서 많이 사용된 빈도순에 의한 K-MeSH 키워드 리스트가 나타난다. 우선순위는 여러 방법으로 계산되어 질 수 있는데, 이 예에서는 간단히 용어의 빈도를 사용했다. (그러나 영문 MeSH 키워드 추천 시에는 용어 빈도와 역문헌 빈도를

입력하신 내용의 결과입니다.

목적: H. pylori의 감염은 위점막 상피 세포의 apoptosis를 유도할 뿐만 아니라 세포 증식을 변화시킨다. Prostaglandine(PG)의 생성을 유도하는 cyclooxygenase-2(COX-2)는 상피세포의 증식를 변화시킬 수 있다. 본 연구에서는 H. pylori 감염에 의한 COX-2 유전자의 발현이 위상피 세포 apoptosis에 영향을 미칠 수 있는지를 검사하였다. **방법:** Hs746T 인체 위상피 세포에 H. pylori를 감염 시킨 뒤, COX-2 mRNA 발현 분자수와 단백질은 각각 정량적 역전사 중합효소법과 Westen blot으로 검사하였다. 배양 상청액의 PGE_2는 방사면역법으로 측정하였다. 위상피 세포의 apoptosis 정도는 flow cytometry 분석과 cell death detection ELISA를 이용하였다. 또한 caspase-3의 활성은 첨가한 기질 DEVD-pNA로부터 분해되는 p-nitroanilide를 측정하여 결정하였다.

빈도	키워드	타 입	우선어	영 어	고유번호
4	상피 세포	상용어	상피세포	Epithelial Cells	A11.436
3	감염	우선어	감염	Infection	C01.539
3	apoptosis	영어	고사	Apoptosis	G04.335.139.160
1	위점막	우선어	위점막	Gastric Mucosa	A10.615.550.291

[그림 1] K-MeSH 용어 추출의 예(아래 일부분은 편집상 잘림)

사용하여 순위를 좀 더 정확하게 계산하였다.) 그리고 한 행에는 각 K-MeSH 용어에 해당 되는 표제어(또는 우선어), 영문 MeSH 용어, 시소러스 계층 번호 등의 K-MeSH 주요 필드들이 표시된다.

3. 한국어 의학 용어에서의 변이

한국어 의학 용어들은 대체로 영어, 일본어, 중국어, 독일어, 프랑스어 등과 같은 외국어로부터 들어온 차용어로서 번역되기도 하고 원어 그대로 사용되기도 한다. 따라서 동일한 의미의 용어라도 그 어원이나 출처의 차이로 인해 다르게 해석되기도 하고, 번역 과정에서 발생되는 다양한 단어 선택으로 인해 다양하게 표현되기도 한다. 이와 같은 용어 변이의 범위는 크게 아래와 같이 분류될 수 있다.

1. 번역 변이 : 용어의 의미는 같지만 동의어 혹은 유사어 때문에 번역자에 따라 다르게 번역된다.
2. 음역(transliteration) 변이 : 용어의 발음 그대로 번역되며, 이는 언어의 출처에 따라 매우 다양하게 표현된다. 음역규칙 또한 다양하다.
3. 띄어쓰기 변이 : 띄어쓰기 규칙은 여러 가지 다양한 표기도 인정하므로, 동일한 복합어가 다르게 띄어 쓰여진다.
4. 파생(derivational) 변이 : 명사들 사이에서 '의'(소유격 조사)와 '과'(접속 조사)와 같은 조사가 포함되어 파생된 형식으로 사용된다.
5. 약어(abbreviation) : 용어를 줄인 이름으로도 사용한다.
6. 언어(language) 변이 : 한국어 문서에서 사용되는 영어와 중국 문자들의 변이로, 영어 변이들은 약어와 동의어와 같은 비표준어인 영어 변이가 사용된다.

K-MeSH는 각 한국어 의학 용어들에 관한 표제어와 비표준어인 동의어, 변이, 기타 관련된 용어들을 포함하고 있다. 이전 K-MeSH 검색 프로그램에서는 비표준 용어들에 대한 표제어를 찾기 위해 관계형 데이터베이스 검색 명령어를 사용한다. 이 경우 데이터베이스 내에 용어에 대한 모든 변이들이 완전하게 나열되어 있지 않을 뿐더러 데이터베이스의 검색 명령도 정확하게 일치되는 용어만을 검색하기 때문에, 변이를 모두 찾아내는 비율이 높지 않다. 게다가 신조어를 포함한 모든 변이들을 계속적으로 나열하여 포함시키는 것은 거의 불가능하다.

4. 띄어쓰기 변이

띄어쓰기 변이는 같은 의미를 가졌지만 표제어와는 다르게 띄어 쓴 복합어를 말한다. 예를 들어, '모세혈관저항성'(Capillary Resistance)에 대한 띄어쓰기 변이들은 다음과 같다. (밑줄은 공백을 의미함)

(1) a. 모세혈관저항성
b. 모세혈관저항_성
c. 모세혈관_저항성
d. 모세혈관_저항_성
e. 모세_혈관저항성
f. 모세_혈관저항_성
g. 모세_혈관_저항성
h. 모세_혈관_저항_성

대체로 n개의 단어로 구성된 용어에 대한 모든 띄어쓰기 변이 경우의 수는 2^{n-1}가지이다. 예를 들어, 4개의 단어로 구성된 용어인 경우에는 2^3가지 띄어쓰기 변이가 존재한다. 한국어 의학 용어는 대부분 여러 단어로 구성된 복합어이며, 길이가 긴 경우도 많이 존재한다. 따라서 이전 시소러스처럼 각 용어에 대해서 생성 가능한 모든 띄어쓰기 변이를 등록하게 되면 매우 번거로운 수작업이 필요하며, 또한 시소러스의 크기가 커지게 된다.

이러한 띄어쓰기 변이 문제를 본 논문에서는 간단한 방법으로 해결했다. 시소러스에는 최대로 띄어 쓴 용어만을 등록하고, 검색에서는 해당 용어의 공백을 제거하는 방법을 사용한다. 위의 예인 경우, 최대로 띄어 쓴 (1h)가 시소러스에 등록되면, (1)에 표시한 모든 용어들이 검색 될 것이다. 하지만 다음의 (2)와 같이 띄어 쓴 용어들의 경우에는 검색되지 않을 것이다.

(2) a. 모세혈_관저항성
b. 모_세혈관저항성

부정확하게 결합된 단어가 검색될 가능성이 있긴 하지만, 이 방법은 띄어쓰기 변이 문제를 해결함으로써 검색의 재현율을 증가시킬 것이고, 시소러스의 크기를 줄이게 될 것이다. 실제 본 실험 결과, 이전 시소러스에 비해 K-MeSH 용어가 총 47,100개에서 약 27,300개로 줄었으며, 크기 또한 약 58%로 축소되는 성능을 보였다.

5. 키워드 검출을 위한 분리

한국어는 단어의 끝에 조사나 어미등의 접미사가 붙는 교착어이다. 어절은 기본 단어에 접미사가 붙은 것으로 공백으로 구분한다. 본 논문에서 분리(segmentation)란 K-MeSH 용어를 검색하기 위해 각 어절을 용어와 접미로 나누는 것을 말한다.

본 논문에서는 K-MeSH 용어를 검색하기 위해 명사일치(noun matching) 방법과 부분문자열 일치(substring matching) 방법을 사용하였다. 명사일치 방법은 명사추출기(NE)에 의해 전처리된 명사에서 K-MeSH 용어를 추출 해 낸다. 명사추출기는 형태소 분석에 의해 띄어쓰기 단위인 어절 단위로 명사를 추출한다. 명사추출기의 일반적인 문제점은 분리 애매성과 미등록어 문제이다. 분리 애매성의 예를 들면, (3a)의 어절은 (3b), (3c), (3d)처럼 3가지 형식으로 분리될 수 있다. (3b)는 '은'이 조사로 분석되어 명사 '금'을 출력하게 될 것이다. (3d)가 하나의 명사로 분석된 반면에 (3c)는 두 개 명사의 결합으로 분석된다. KTAP에서는 가장 긴 단어 일치인 최장 일치 방법을 사용하는데, 만약 시소러스에 '금은'이 등록된다면, (3a)의 결과는 (3c)가 아닌 (3d)가 출력된다.

(3) a. 금은
b. 금/noun(gold) + 은/particle
c. 금/noun(gold) + 은/noun(silver)
d. 금은/noun(gold and silver)

두번째 문제점은 (4a)에서 '콜레라(cholera)'와 같은 미등록어가 있을 경우의 문제이다. (4b)가 정확한 분석이지만, 명사의 마지막 음절인 '라'가 (4c)처럼 조사의 일부로 분석되어 질 수 있다.

(4) a. 콜레라는
b. 콜레라/noun + 는/particle
c. 콜레/noun + 라는/particle

명사추출기는 대체로 문법상 애매성이 없다면, 위에서 언급한 (3)과 (4)처럼 잘못 분리된 부분 문자열 일치를 피하기 위해 단순 명사만 추출한다. 그러나 이것은 복합어 '작은 창자'(small intestine)와 '큰 창자'(large intestine)처럼 형용사와 명사로 결합된 미등록어 검색은 실패할 것이다.

부분문자열 일치 방법은 전처리기 없이 입력된 문서에서 시소러스의 가장 긴 K-MeSH 용어를 추출해 낸다. 이 방법은 다른 단어와 연결된 부분 문자열까지 추출해 낸다. 따라서 추출된 용어 중에는 입력 문장에서 미등록어와 함께 결합된 복합어의 일부일 가능성도 있다. 예를 들어, '항체역가를 측정하여'에서 '항체역가를'인 경우 시소러스에 '항체'라는 용어만 등록된다면 '항체'를 추출해 낸다. 이와 같이, 대부분 추출된 2음절이상의 용어인 경우 잘못된 경우가 적으므로, 본 논문에서는 유효한 K-MeSH 용어로 보았다. 그러나 단음절어에 대해서는 조건에 따라 처리했으며, 자세한 내용은 7절에서 다룬다.

6. 검색 전략

본 논문에서는 띄어쓰기 변이를 고려해서 어절단위 검색(word phrase based search; WP)과 음절단위 검색(syllable based search; SYL) 두 가지 방법으로 K-MeSH 용어를 검색한다. 어절단위 검색은 검색된 어절의 바로 다음 어절부터 검색을 시작한다. 이 방법은 사용된 모든 용어들이 공백으로 명확히 구분되어 있고, 반드시 어절의 처음 부분에 있다고 가정한다. 음절단위 검색은 검색된 용어의 바로 다음 음절부터 검색을 시작한다. 이 방법은 한국어 의학 용어들이 다른 용어들과 함께 붙여 써서 사용될 수 있다는 것을 고려한 것이다. 따라서 이 방법은 복합어 내에 사용된 K-MeSH 용어들도 추출해 낼 수 있다.

앞 절에서 설명한 띄어쓰기 변이를 고려하여 검색 할 경우, 각 검색 전략의 결과는 다르게 수행될 것이다. (5)는 각각 A, B, C, D로 표현된 4개의 단어로 이루어진 가능한 조합 형태를 보여준다. A단어가 검색되어 졌고, D는 다른 단어 부분이며, B_C가 시소러스에 등록되어 있다고 가정한다. 이때 음절단위 검색 전략은 용어의 시작이 공백 다음이건 아니건 관계없이 검색되므로, 띄어쓰기 변이를 고려할 경우에는 (5)에 나타난 B_C와 BC가 사용된 모든 경우의 키워드를 추출할 것이다. 반면에 띄어쓰기 변이를 고려하지 않을 경우에는 (5a), (5b), (5e), (5f)만 추출할 것이다. 어절단위 검색 전략에서는 용어의 시작이 공백 다음에 나타난 것만을 추출한다. 이때 띄어쓰기 변이를 고려할 경우에는 (5a), (5b), (5c), (5d)만 추출할 것이고, 띄어쓰기 변이를 고려하지 않을 경우에는 (5a)와 (5b)만 추출할 것이다. 또한 BC가 시소러스에 등록된 경우, 음절단위 검색 전략에서 띄어쓰기를 고려할 경우에는 (5c), (5d), (5g), (5h)만 추출할 것이고, 띄어쓰기를 고려하지 않을 경우에는 (5g)와 (5h)만 추출할 것이다. 그리고 어절단위 검색 전략에서는 띄어쓰기를 고려할 경우와 고려하지

않을 경우 모두 (5c)와 (5d)만 추출할 것이다.

(5) a. A_B_C_D
b. A_B_CD
c. A_BC_D
d. A_BCD
e. AB_CD
f. AB_C_D
g. ABC_D
h. ABCD

7. 단음절어(OSW)

단음절어(One Syllabel Word : OSW)란 하나의 음절로 이루어진 단어이다. 대체로 단어들은 중의성이 있기 때문에 K-MeSH 용어와는 다른 의미를 가진 단어가 추출될 수 있다. 예를 들어, '위'는 K-MeSH 용어에서는 위(胃,stomach)를 의미하지만, 일반적인 용어로는 위(上, above)나 위(位, position)를 의미하기도 한다. (6)은 어절 분리가 정확하게 되었더라도 그 의미가 두 가지로 애매하여 불분명함을 보여준다. 문맥상으로는 (6c)의 경우일지라도 K-MeSH 용어 (6b)로 잘못 선택될 수도 있다. 애매성의 해결은 전체 문장, 단락, 심지어 전체 문서를 이해할 수 있는 자연언어 기술이 필요하다. (7)과 (8)은 단음절어 '위'가 잘못 분리된 예들을 보여준다.

(6) a. 위에서
b. 위/noun +에서/particle (in the stomach)
c. 위/noun +에서/particle (in the above)

(7) a. 상위시대(a higher rank period)
b. 상+위/noun(stomach) + 시대
c. 상위+ 시대

(8) a. 위성통신(satellite communication)
b. 위/noun(stomach) + 성 + 통신
c. 위성 + 통신

정확한 단음절어 추출은 쉽지 않을 뿐더러 논문의 키워드 선택에서도 그렇게 비중을 많이 차지 하지 않는다. 단음절어를 다루기 위해 본 논문에서는 3가지 선택 방법, 1) 단음절어 전부 무시, 2) 단음절어 모두 선택, 3) 조건에 맞는 단음절어의 선택을 두었다. 3)에서 조건이란 두 가지 조건을 가지는데, 첫 번째 조건은 단음절어에 다른 단어가 붙어 있지 않은 경우, 두 번째 조건은 유효한 접미사로 간주되는 다른 단어가 붙어 있는 경우이다.

선택 3)의 예를 들면 다음과 같다. (9)는 각각 A, B, C로 표현된 3개의 단어로 이루어진 가능한 조합 형태를 보여준다. A가 검색되어 졌고, C는 다른 단어부분이라고 가정한다. B가 시소러스에 등록된 단음절어라고 할 때, (9a)의 B가 추출될 것이고, (9b)의 B는 C가 유효한 접미사라면 추출될 것이다. 그러나 (9c)와 (9d)의 B는 검색 전략과는 상관없이 추출되지 않을 것이다.

(9) a. A_B_C
b. A_BC
c. AB_C
d. ABC

8. 실험 및 결과

8.1 실험 방법

실험은 298개 한국어 의학 저널[9]에서 44,285개 요약문으로 구성된 한국어 의학 데이터베이스(KMBASE)를 사용하였다.

성능은 각 방법의 가능한 모든 조합에 대해서 K-MeSH 용어 추출과 영문 MeSH 키워드 추천 두 가지를 측정하였다. 읽기 쉽게 하기 위해 아래의 약어들로 각 방법을 표현한다.

1. 띄어쓰기 변이 처리 : 'C'는 압축 사전을 이용한 띄어쓰기 변이를 다룬 방법을 나타내고, 'N'은 일반 사전을 이용한 띄어쓰기 변이를 다루지 않은 방법을 나타낸다.
2. 검색 전략 : 'W'는 어절단위(WP) 검색 전략, 'S'는 음절단위(SYL) 검색 전략을 나타낸다.
3. 단음절어(OSW) 처리 : 'X'는 모든 단음절어를 무시한 것, 'A'는 모든 단음절어를 선택하는 것, 'O'는 유효한 단음절어만 선택하는 것을 나타낸다.

본 실험에서는 위의 약어인 각 3개의 글자를 조합하여 한 가지 실험 방법을 표현한다. 예를 들어, 'CSA'는 띄어쓰기 변이를 다루고, 음절단위 검색 전략으로 추출되는 모든 단음절어를 선택한 방법을 의미한다. 그리고 단지 한 글자로 표현하는 경우는 그 글자를 가지고 조합할 수 있는 모든 방법을 의미한다. 예를 들어, 'C-방법'은 CSA, CSO, CSX, CWA, CWO, CWX 방법을 집합적으로 표현한다. 그리고 'S-방법'은 CSA, CSO, CSX, NSA, NSO, NSX 방법을 표현한다.

8.2 K-MeSH 용어 추출 실험

이 실험은 각 방법에 대해 K-MeSH 용어 추출을 평가하기 위한 것으로 결과는 표1과 같다. 띄어쓰기 변이 성능 실험을 위해서, 각 방법은 임의적으로 선택된 5개의 요약문을 실험 평가하였다. 대체로 어절단위 검색 전략 방법들이 높은 정확률을 보인 반면에 음절단위 검색 전략 방법들은 높은 재현율을 보였다.

[표 1] K-MeSH 용어 추출 결과
(R=Recall, P=Precision, F=F-measure, 단위 : %)

	명사 일치			부분문자열 일치		
	R	P	F	R	P	F
CSA	80.0	46.1	58.5	98.9	42.0	59.0
CSO	71.5	73.7	72.6	88.8	82.7	85.3
CSX	71.0	81.6	75.9	87.4	94.3	90.7
CWA	66.6	54.5	59.9	77.6	61.2	68.4
CWO	60.0	70.5	64.8	70.8	81.4	75.7
CWX	59.4	79.3	68.0	70.2	96.3	81.2
NSA	80.0	46.1	58.5	98.6	41.9	58.8
NSO	71.5	73.7	72.6	87.7	82.5	85.0
NSX	71.0	81.6	75.9	87.1	93.9	90.4
NWA	66.6	54.5	59.9	77.3	61.0	68.2
NWO	60.0	70.5	64.8	70.5	81.3	75.5
NWX	59.4	79.4	68.0	69.9	96.3	81.0

8.3 영문 MeSH 키워드 추천 실험

여러 용어들 중 각 논문의 키워드로는 논문의 핵심을 표현할 수 있는 용어를 선택해야만 한다. 따라서 키워드를 선택하는 데 있어서 손쉽게 선택하기 위해 추출된 용어들을 순서화 할 필요성이 있다. 키워드 선택에 있어서 띄어쓰기 변이 효과를 살펴보기 위해, 본 실험에서는 가장 인기 있는 순위 방법 중의 하나인 용어 빈도 및 역문헌 빈도(TF-IDF)[11]를 사용했다.

한국어 의학 데이터베이스의 여러 저널들은 한국어 요약문 뿐만 아니라 영어 요약문을 포함하고 있으며, 키워드 또한 한국어나 영어 또는 둘 다 포함하고 있다. 본 실험을 위해서 한국어 의학 데이터베이스의 전체 44,285개 논문 가운데 한국어 요약문이면서 영문 키워드를 포함하고 있는 논문들을 실험 대상으로 하였다. 실험 대상은 25,729개이며 그 중 90%(23,156개)는 IDF 값을 얻기 위한 학습 데이터로 사용되었고, 10%(2,573개)는 실험 데이터로 사용되었다.

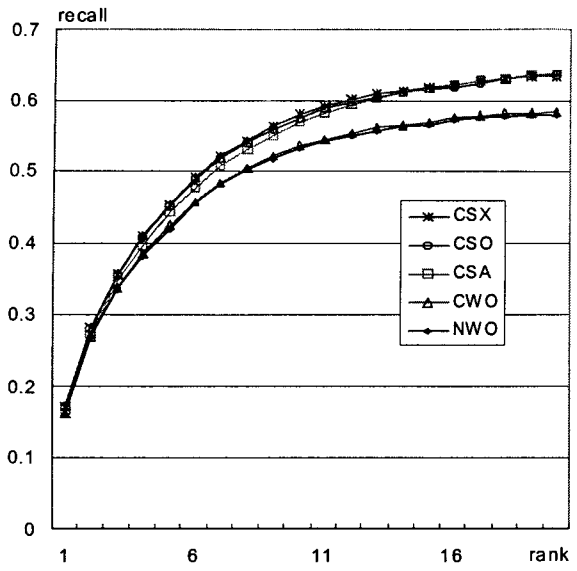
한국어 요약문에서 추출된 K-MeSH 용어들을 K-MeSH 시소러스의 영문 MeSH 용어로 변환시키는데, 이를 본 실험에서는 '생성된 키워드' (generated-keyword)라고 명명한다. 가장 좋은 표준 영문 MeSH 키워드 선택을 위해서 저자가 직접 입력한 영문 키워드 중 MeSH 용어만을 선별하는데 이를 본 실험에서는 '정답 키워드'(answer-keyword)라고 명명한다. 실험 결과 실험 대상 문서에서 영문 키워드 중 40.1%가 MeSH 용어임이 나타났다. 이는 한국 가정 의학 협회 저널의 논문에서 유효한 영문 MeSH 키워드의 사용 실태를 조사한 결과인 약간의 철자 차이는 같은 것으로 처리하여 얻은 44%[3]에 가까운 수치이다.

표2는 각 방법에 있어서 추출 결과의 순위에 따른 정답 키워드의 최대 재현율을 보여준다. S-방법은 W-방법보다 훨씬 더 높고, C-방법은 N-방법보다는 약간 더 높다. 단음절어 처리에 대해서는 A-방법, O-방법, X-방법 순으로 나타났다.

[표 2] 영문 키워드 추천 성능 – 각 순위내 정답 키워드 추출 재현율 (단위 : %)

	5위이내	10위이내	20위이내	최대
CSA	44.4	56.9	63.7	65.6
CSO	45.2	57.4	63.6	64.6
CSX	45.3	58.0	63.4	64.2
CWA	41.9	53.2	58.4	59.5
CWO	42.4	53.7	58.4	59.1
CWX	42.7	54.1	58.2	58.8
NSA	44.2	56.6	63.5	65.4
NSO	45.0	57.2	63.3	64.4
NSX	45.1	57.8	63.1	63.9
NWA	41.6	52.9	58.1	59.2
NWO	42.0	53.4	58.1	58.8
NWX	42.4	53.8	58.0	58.4

그림2는 5가지 방법들에 대한 상위 20위내의 재현율을 보여준다. C-방법과 N-방법사이에는 많은 차이점이 없지만, 상위 순위에서 S-방법은 W-방법 보다 훨씬 더 높다. CWO와 NWO는 거의 비슷하게 나타났고, S-방법인 CSA, CSO, CSX 보다는 낮게 나타났다. 각 방법에서의 단음절어의 차이는 그렇게 크지 않았지만 20위쯤에서는 CSA가 CSO보다 약간 더 높게 나타났다. 그러나 20위 이후의 높은 재현율 순서는 CSA, CSO, CSX로 변했다.



[그림 2] 상위 20위내에서의 영어 키워드 추천 성능의 변화

9. 토 의

명사일치의 성능은 부분문자열 일치에 비해 CSA와 NSA의 정확률을 제외하고는 나쁘게 나타났다. 이것은 명사추출기가 재현율을 희생하면서 단음절어를 정확하게 선택했음을 의미한다. 그러나 단음절어를 무시하거나 적절하게 다룬 경우 부분문자열 일치가 명사 일치보다 항상 성능이 좋게 나타났다. 왜냐하면, 명사추출기는 어절 단위로 처리되기 때문에, 띄어 쓴 한 단위 용어를 찾지 못한다. 그래서 명사추출기는 복합어가 발견되었을 때 문맥의 고려 없이 단지 첫번째 명사만 선택한다. 이것은 간단한 일반적인 명사추출기는 시소러스 용어 추출의 전처리기로는 유용하지 않다는 것을 보여준다. 따라서 이제부터는 부분문자열 일치 방법만 논한다.

C-방법은 모든 경우에 있어서 항상 N-방법보다 약간 좋게 수행된다. 이것은 띄어쓰기 변이를 찾기 위해 조금 더 문자열 계산에 시간을 썼음에도 불구하고 성능이 좋다는 것은 매우 의미 있는 결과이다. 또한 시소러스 크기가 약 42% 축소되는 성능을 보였다.

음절단위 검색 전략의 재현율은 항상 어절단위 검색 전략 보다 높다. 그러나 정확률은 단음절어가 정확히 다루어 졌을 때만 좋다. 즉 CSO(82.7%)가 CWO (81.4%)보다 높고, NSO(82.5%)가 NWO(81.3%)보다 높다. 이것은 다른 단어의 중간에 있는 단음절어가 틀리게 추출 되는 경향이 있음을 의미한다.

단음절어 처리에 관해서 보다 적은 단음절어 추출은 높은 정확률을 얻을 수 있고, 보다 많은 단음절어 추출은 높은 재현율을 얻을 수 있다는 결론을 내릴 수 있다. 비록 유효한 단음절어를 얻기 위한 접미사 검사 프로그램을 사용한다 할지라도 기대하는 것 만큼의 재현율을 향상시킬 수는 없었다. 본 실험에서 사용한 접미사 검사 프로그램은 가능한 조사 및 접미사의 조합을 검사해서 접미사를 판단해 주는 것이다[2]. 그러나 과도생성에 의해 잘못된 조합도 접미사로 처리하는 오류가 있어 이를 향상시킬 필요가 있다.

CSA의 재현율은 분리 오류 때문에 100%가 아닌 98.9%이다. (10)은 오류의 예를 보여준다. '호흡'(respiration), '기질'(temperament), '질환'(disease)이 시소러스에 등록되어 있을 때, (10b)와 (10c)로 분리되어질 수 있다. 본 실험에서는 시소러스에서 최장일치방법으로 추출하기 때문에, 비록 (10c)가 정확한 것이더라도 (10b)가 추출된다.

(10) a. 호흡기질환
b. 호흡(respiration) + 기질(temperament) + 환
c. 호흡(respiration) + 기(organ) + 질환(disease)

더 많은 용어를 추천해 주는 시스템이 사용자에게는 논문의 키워드를 선택하는 데 있어서 더 좋다. 만약 결과에 틀리게 분리되거나 인식된 용어가 포함되어 있다면 사용자들은 불편하게 느낀다. 따라서 재현율과 정확률 사이의 균형이 필요하다. 본 실험에서는 F-measure 계산에 정확도가 재현율만큼 동등하게 가중되었다고 결정하였다. 표1의 F-measure에 의하면 CSX(90.7%)가 가장 높기에 키워드 추천 프로그램(Term Access Program)을 위한 기본적인 방법으로 선택했다.

실험 결과 표2에서 모든 방법들 중 가장 높은 최대 재현율은 CSA(65.6%)로 나타났다. 이것은 정답 키워드에 해당하는 한국어 용어의 약 34.4%가 한국어 의학 논문의 요약문에서 사용되지 않거나, 정답 키워드가 K-MeSH 시소러스에 타당한 한국어 번역이 없어서 변환 되지 않았을 것이라고 추측 할 수 있다.

표2의 각 방법들의 최대 재현율은 표1의 결과와 거의 유사하다. 즉 C-방법은 N-방법보다 약간 좋고, S-방법은 W-방법보다(약 5%정도) 좋다. 이것은 정확한 한국어 MeSH 용어 추출이 영문 MeSH 용어 추출의 성능에 직접적으로 영향을 준다는 것을 보여준다. 띄어쓰기 변이를 고려한 C-방법과 S-방법은 그렇지 않은 방법들보다 성능이 더 우수했다.

성능에 영향을 준 가장 중요한 요인은 검색 방법이다. S-방법은 다른 방법들보다 우수했다. 두 번째 중요한 요인은 성능에 영향을 주는 단음절어 처리이다. 논

문에 제시되는 키워드의 수가 대체로 10개 이하라는 점을 고려해 보면, 단음절어를 무시하는 방법이 단음절어를 선택하는 방법보다 우수하다. 이것은 단음절어 처리가 성능 향상을 위해 필요하지만, 그렇게 효과적이지 않다는 것을 보여준다. 본 실험에서 상위 10위내에서는 CSX가 영문 MeSH 키워드를 추천하는데 가장 우수한 방법으로 나타났다. 이는 전형적인 방법인 NWX의 재현율 53.8%를 약 4.2% 포인트(비율은 7.8%) 증가시킨 58%이다. C-방법은 성능을 증가시켰을 뿐만 아니라 시소러스의 크기를 약 42% 축소시켰다. 이것은 이전 버전의(압축하지 않은) 시소러스에 많은 띄어쓰기 변이를 수작업으로 입력했다는 것을 의미한다.

10. 결 론

본 논문에서는 K-MeSH 용어 추출과 영문 MeSH 키워드 추천 방법에서의 띄어쓰기 변이 문제를 위한 해결 방법을 제안했다. 변이는 단어들과 함께 분리되거나 결합된 형식으로 사용되어 지기 때문에 일단 어절로부터 분리해서 시소러스에 있는 K-MeSH 용어와 비교했다.

입력 문서에서의 K-MeSH 용어 추출은 부분문자열 일치 방법으로 검색하였다. 이 방법은 K-MeSH 용어가 공백 없이 다른 단어에 붙여 쓴 경우를 고려한 것이다. 또한 전문 용어 특성상 입력 문서에서 동일한 복합어이지만 다르게 띄어 쓴 띄어쓰기 변이가 발생한다. 이를 위해 시소러스에는 생성 가능한 모든 띄어쓰기 변이를 등록하는 것 대신에 최대로 띄어 쓴 용어만 등록하고 이를 이용하여 모든 변이를 검색해 낸다. 이 결과 시소러스의 크기가 약 42%정도 축소되었다.

본 실험에서는 띄어쓰기 변이를 고려하고 음절단위 검색 전략을 사용하며 단음절어를 무시한 CSX 방법이 K-MeSH 용어 추출에서 F-measure가 90.7%(재현율 87.4%, 정확률 94.3%)로 평가되었다. 반면에 띄어쓰기 변이를 고려하지 않고 어절단위 검색 전략으로 단음절어를 무시한 NWX 방법은 81.0%(재현율 69.9%, 정확률 96.3%)로 나타났다. 또 영문 MeSH 키워드 추천에서 CSX 방법은 상위 10위내의 재현율이 58%로 평가되어 진 반면에 NWX 방법은 53.8%로 나타났다. 이 결과는 본 논문에서 제안한 띄어쓰기 변이를 고려하고 음절단위 검색 전략을 사용하여 단음절어를 무시한 CSX 방법이 교차 언어 간의 MeSH 용어 추천 방법에 효과적임을 보여준다.

본 논문에서는 간단한 부분문자열 일치 방법과 간단한 접미사 검사 프로그램을 사용했으므로 부분 파싱이나 태깅 방법 등과 같은 더 정교한 자연 언어 처리 기술을 사용한다면 성능을 좀 더 향상시킬 수 있을 것이다. 또 시멘틱 정보나 문맥의 정보를 추가로 사용 한다면 용어 추천의 성능을 더 향상시킬 수 있을 것이다.

11. 참고문헌

[1] 강병주, 최기선, 윤준태. 1998. 한국어 정보검색에서 복합명사 색인 실험. 한글 및 한국어 정보처리 학술대회, 130-136.

[2] 강승식. 2002. 한국어 형태소 분석과 정보 검색. 홍릉과학출판사.

[3] 김병선, 김수영. 가정의학회지 논문의 영문 주제어 선택에 있어서 MeSH용어 사용 여부와 선택 정확도. 대한가정의학회지, 1998;19(17) : 531-537.

[4] 윤보현, 김상범, 임해창. 1998. 한국어 정보검색에서 구문적 용어불일치 완화방안. 한글 및 한국어 저보처리 학술대회. 143-149.

[5] Aronson, Alan R. 1996. *The effect of textual variation on concept based information retrieval.* In proceedings of AMIA annual fall symposium, 373-377.

[6] Aronson, Alan R., Bodenreider, Oliver, Chang, H. F Florence, Humphrey, Susan M., Mork, James G., Nelson, Stuart J., Rindflesch, Thomas C., Wilbur, W. John. 2000. *The NLM indexing initiative.* In proceedings of AMIA symposium, 17-21.

[7] Hersh, W., Buckly, C., Leone, TJ. 1994. OHSUMED : *An interactive retrieval evaluation and new large test collection for research.* In proceedings of seventeenth annual international ACM-SIGIR conference on research and development in information retrieval. Dublin, Ireland, Spring-Verlag, 192-201.

[8] Kim, Won, Aronson, Alan R,. Wilbur, W. John. 2001. *Automatic MeSH term assignment and quality assessment.* In proceedings of AMIA symposium, 319-323.

[9] KMbase. 2004. http : //kmbase.medric.or.kr/.

[10] MeSH. 2004. http : //www.nlm.nih.gov/mesh/.

[11] Salton, G. 1989. *Automatic text processing.* Readings, Massachusetts, Addison-Wesley series in computer science.

[12] Srinivasan, P. 1996. Optimal document indexing vocabulary for MEDLINE. *Information Processing & Management,* 32(5) : 503-514.