MeSH 시소러스를 이용한 한영 교차언어 키워드 자동 부여

이 재 성[†] · 김 미 숙[††] · 오 영 순[†††] · 이 영 성[††††]

요 약

의학용 시소러스인 MeSH (Medical Subject Heading)는 영어 의학 논문 색인을 위한 통제어 시소러스로서 오랫동안 사용되고 있다. 본 논문에서는 한국어 MeSH를 이용하여 한국어 의학 논문의 요약문에 자동으로 영문 MeSH 색인어를 부여하는 '교차언어 키워드 부여' 방법을 제안하고 색인 전문가 및 저자의 색인 효율과 비교한다. 이 색인어 부여 과정은 우선 한국어 MeSH 용어를 문장에서 인식하여 추출하고, 이 용어를 다시 영어 MeSH 용어로 바꾼 후, 용어의 중요도를 계산하여 상위의 용어를 색인어로 부여한다. 특히, 한국어 MeSH 용어 추출을 위해 효과적으로 띄어쓰기 변이를 처리할 수 있는 방법을 제안한다. 실험 결과, 띄어쓰기 변이를 효과적으로 처리하여 한국어 MeSH의 크기를 약 42% 정도 줄였을 뿐만 아니라, 후보 색인어 추출의 효과도 높였다. 또 이 방법을 이용하여 색인어 자동 부여를 한 후, 색인 전문가 및 저자의 색인 결과를 비교한 결과, 이 자동 색인 방법이 전문가의 색인 능력보다는 부족했지만, 저자의 색인 능력과는 별 차이가 없음을 보였다.

키워드 : MeSH, 한국어 MeSH, 자동 키워드 부여, 자동 색인, 교차언어, 띄어쓰기 변이

Automatic Korean to English Cross Language Keyword Assignment Using MeSH Thesaurus

Jae Sung Lee[†] · Mi Suk Kim[††] · Young Soon Oh[†††] · Young Sung Lee[††††]

Abstract

The medical thesaurus, MeSH (Medical Subject Heading), has been used as a controlled vocabulary thesaurus for English medical paper indexing for a long time. In this paper, we propose an automatic cross language keyword assignment method, which assigns English MeSH index terms to the abstract of a Korean medical paper. We compare the performance with the indexing performance of human indexers and the authors. The procedure of index term assignment is that first extracting Korean MeSH terms from text, changing these terms into the corresponding English MeSH terms, and calculating the importance of the terms to find the highest rank terms as the keywords. For the process, an effective method to solve spacing variants problem is proposed. Experiment showed that the method solved the spacing variant problem and reduced the thesaurus space by about 42%. And the experiment also showed that the performance of automatic keyword assignment is much less than that of human indexers but is as good as that of authors.

Key Words : MeSH, Korean MeSH, Automatic Keyword Assignment, Automatic Indexing, Cross Language, Spacing Variants

1. 서 론

MeSH(Medical Subject Headings)는 미국 NLM(National Library of Medicine)에서 논문의 원활한 검색을 위하여 수작업으로 제작되어 사용하고 있는 계층적으로 통제된 시소러스이다. MeSH는 한 가지 개념에 대해 서로 다른 용어가 존재할 경우, 이를 정리하여 통일하였고, 또, 생명과학에 사용되는 개념을 계층구조로 만들어 상위 개념과 하위 개념을 파악할 수 있도록 하였다. 현재 MeSH 용어는 MEDLINE을 포함한 세계 수많은 기관에서 의료 정보와 논문, 책, 자료 색인을 위해 사용하고 있다[1].

MEDLINE에서는 MeSH 용어를 사용하여 전문가들이 수동으로 색인 및 검색을 하고 있으며, 이런 수동 색인의 효율을 높이기 위해 자동 혹은 반자동으로 MeSH 키워드를 부여하는 연구가 NLM 등을 중심으로 이루어지고 있다. 이 연구에서는 주어진 논문에 대해 색인 전문가에게 키워드를 자동으로 추천해주어 보다 효율적으로 MeSH 색인을 돕도록 하고 있다[2, 3]. 정보 검색 연구 분야에서, 특히 의료 정보 검색 분야에서 MeSH 키워드를 추출하여 색인어로 사용할 경

※ 이 논문은 2004년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었습니다.
† 종신회원 : 충북대학교 컴퓨터교육과 조교수
†† 정 회 원 : (재)충부직업전문학교
††† 정 회 원 : 오창고등학교 교사
†††† 정 회 원 : 충북대학교 의과대학 의학과 의료정보학 및 관리학교실 부교수
논문접수 : 2005년 11월 25일, 심사완료 : 2006년 3월 27일

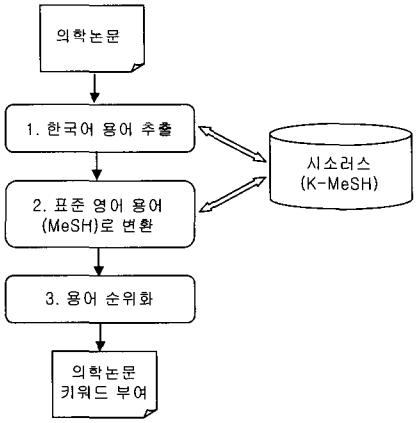
우, 텍스트에서 추출한 색인어만을 사용하는 경우보다 검색 성능이 향상되었다[4, 5]. 이러한 일련의 연구들은 효과적인 MeSH 키워드 부여가 실질적으로 정보 검색의 효과를 높이는데 기여하고 있음을 보여 준다.

국내에서도 의학 학술지의 대부분이 저자가 논문의 키워드를 선정할 때 세계적인 문헌 정보 등록을 위해 영어 MeSH 용어를 사용하도록 투고 규정에 권장하고 있다. 그러나, MeSH 키워드를 입력하려고 할 경우, 저자나 색인 전문가가 표준화된 MeSH 용어를 쉽게 기억하지도 못할 뿐만 아니라, 여러 개의 가능한 MeSH 용어 중 필요한 중요 용어를 선택하는 일도 쉽지는 않다[6]. 즉, 한국어 용어를 사용하여 작성된 한국어 논문의 경우, 대응되는 영어 MeSH 용어를 키워드로 선택하기 위해서는 많은 용어에 대한 지식이 필요하며, 또 용어의 선택을 위한 많은 시소러스 참조 작업이 필요하다. 미국 NLM의 연구가 주로 영어 키워드 추천에 관한 것이므로, 한국어에서 영어 용어를 추천해 주는 교차언어(Cross Language) 키워드 추천에 대한 연구가 필요하다. 본 연구에서는 1. 한국어로 작성된 논문의 요약에서 통제된 영어 키워드, 특히 MeSH 키워드를 자동으로 추출하는 일반적인 과정을 제안하고 2. 그 과정에서 특히 한국어 키워드 추출의 성능을 높이기 위한 띄어쓰기 변이 처리 방법을 제안하며, 3. 실험을 통해 한국어 용어 추출 능력이 가장 좋은 방법을 찾아내고, 이 방법으로 키워드 자동 부여 프로그램(KAP: Keyword Assignment Program)을 구현하고, 영문 MeSH 키워드 부여 성능이 색인전문가나 저자의 능력과 어떤 차이가 있는가를 통계적으로 검증한다.

논문 구성은 다음과 같다. 2장에서는 전체적인 한영 교차언어 키워드 자동부여 과정에 대해 설명하고, 3장에서는 한국어 띄어쓰기 변이 처리 방법을 구체적으로 설명한다. 또, 4장에서 두 가지의 실험, 즉, 한국어 MeSH 용어의 추출 성능 비교 실험과 한영 교차언어 키워드 자동부여 실험을 하고, 그 결과를 평가한다. 이어 5장에서 결론을 짓는다.

2. 한영 교차언어 키워드 자동부여의 단계

논문의 내용을 대표할 수 있는 중요 키워드를 추출하기 위해서는 우선 논문 내용의 이해가 필요하다. 하지만, 현재의 자연언어처리 기술로는 완벽한 이해를 할 수 없으므로, 정보검색에서 주로 사용되는 통계적 방법을 이용하여 키워드를 추출한다. 즉, 문장 내에 사용된 모든 단어의 빈도를 계산하고, 이를 이용하여 중요 단어를 선별해 낸다[7]. 본 연구에서는 MeSH 용어만을 처리하기 때문에 MeSH 용어들만을 문장에서 추출하여 빈도를 계산한다. 한국어 문서에서 영어 키워드를 추출하기 위한 교차언어 키워드 추출은 (그림 1)과 같이 3단계로 이루어 질 수 있다.



(그림 1) 한영 교차언어 키워드 부여 단계

1 단계인 한국어 용어 추출은 논문에서 사용된 K-MeSH (한국어 MeSH) 용어를 찾아내는 것이다. (K-MeSH는 영어 MeSH를 가능한 한 용어들 사이에 구분이 가능하도록 번역하여 각 용어에 대한 번역어를 넣고, 한국어에 해당되는 다양한 동의어, 유사어 등을 추가로 등록한 것이다. 따라서, K-MeSH에는 영어 용어뿐만아니라, 한국어 용어가 같이 존재한다[8].) 간단하게 용어를 찾아내는 방법으로는 시소러스에 등록된 엔트리 용어를 단순한 문자열 검색으로 찾아내는 것이다. 하지만, 모든 변형된 유사어들이 시소러스에 있지 않고, 만약 넣는다고 하더라도 시소러스의 용량이 매우 커져 관리에 문제가 있을 수 있다. 또한 시소러스에 있는 문자열과 일치하더라도 잘못된 용어를 추출해 낼 수도 있다. 왜냐하면, 문맥 내에서 사용된 좌우의 다른 어미나 조사 등을 용어의 일부로 잘못 추출해 낼 수도 있기 때문이다. 이러한 문제점을 해결하기 위한 방법은 제 3장에서 논의한다.

2 단계는 한국어 용어를 시소러스 표준 영어 용어로 바꾸는 작업으로, 한국어 용어를 시소러스에서 찾고, 그 시소러스 노드에 있는 영어 용어를 간단히 추출하면 된다. 만약 한국어 용어가 의미 모호성이 있어, 시소러스내의 여러 엔트리와 일치할 경우는 의미 모호성 해결을 위한 방법이 필요하다[9]. 본 논문에서는 단순한 방법으로 대응되는 시소러스 엔트리 중 임의로 한 엔트리를 선택하여 처리한다. 또 다른 방법으로는 시소러스를 이용한 용어 변환시, 바로 영어로 바꾸지 않고, 일단 한국어 표준 용어로 바꾸고, 중요 키워드를 선택한 후, 다시 영어 표준어로 바꾸는 방법이 있을 수 있다. 하지만, 두 번씩 시소러스를 접근하여 시소러스 의미 모호성도 해결하고 영어 표준어로 변환할 때 생길 수 있는 번역 모호성도 해결해야 하기 때문에 더 복잡하고 비효율적으로 여겨지므로 고려하지 않았다.

3 단계인 키워드 순위화 방법은 앞에서 계산된 각 용어들의 중요도를 적절히 조절하여 많은 용어 중 중요한 용어를 우선 제시하는 방법으로 기존에 개발되어 가장 일반적으로 사용되고 있는 용어 빈도와 역문헌 빈도(term frequency and inverse document frequency)에 근거한 순위 계산 방식을 적용하였다[7]. 여기에서 용어 빈도 TF는 주어진 요약문에 나타난 용어의 빈도이다. 또, 전체 문헌수(대개 학습용 문헌수)가 N이고, 그 용어의 문헌 빈도수를 DF라고 할 때, 역문헌 빈도 IDF는 (식 1)과 같이 주어진다. 이를 이용하여 한 용어의 중요도는 간단히 TF*IDF로 계산된다. 정보검색에서 사용되는 TF는 여러 문서를 비교해야 하므로 한 문서내의 최대

TF값으로 정규화를 하는 것이 일반적이다. 하지만, 여기에서 TF는 한 요약문내에서 중요 용어 선정에만 필요하기 때문에 정규화를 하지 않고, 용어의 빈도수를 바로 사용하였다.

$$IDF = \log\frac{N+1}{DF} \quad \text{(식 1)}$$

3. 한국어 용어 추출을 위한 띄어쓰기 변이 처리

영어의 경우, 다양한 변이체의 용어를 인식하는 것이 키워드 자동부여의 성능에 영향을 주고 있다[10]. 이는 용어의 빈도가 용어 선정시 중요한 계산 요소로 작용하기 때문이다. 따라서, 한국어-영어 키워드 자동부여에서도 다양한 변이체의 용어 인식이 필요하다고 볼 수 있다. 특히, 본 논문에서 다루는 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. 이러한 변이는 번역, 음역, 새로운 조사가 붙어서 발생되는 파생, 약어 등으로 인해 발생되고, 복합어의 경우, 다양한 띄어쓰기로 인해 발생할 수 있다. 특히, 한국어 의학 용어들은 대체로 긴 복합어로 이루어져 있으므로 띄어쓰기 변이체가 많이 존재한다.

시소러스는 많은 유사어를 포함하고는 있지만, 실제 다양한 어휘적 변이체를 모두 포함하기에는 그 공간 및 처리 시간의 부담이 크다. 따라서, 중요한 유사어를 등록해 두고, 유사한 변이체는 근사 일치로 찾는 방법이 효과적이며, 실제 영어의 경우에는 효과적인 키워드 추출을 할 수 있었다[10]. 특히 한국어에서의 띄어쓰기 변이체는 간단한 규칙으로 검색이 가능하므로, 모든 띄어쓰기 변이체를 시소러스에 저장할 필요가 없다.

용어 추출에 많이 사용되는 기존의 한국어 형태소 분석기들은 대개 한 어절(띄어쓰기 단위) 단위로 분석을 수행하므로 띄어쓰기 변이를 처리하기 위해서는 수정이 필요하다. 띄어쓰기 처리 문제는 주로 한국어 복합명사의 띄어쓰기 문제로 정보 검색에서 많이 연구되어져 왔다[11, 12]. 기존 논문들에서는 주로 복합명사 띄어쓰기 변이를 찾기 위해 간단한 문장 구조 분석을 한 후, 이들 명사들을 정규화시켜 같은 단어로 인식한다. 그러나 본 논문에서는 문장에 사용된 복합명사 띄어쓰기 변이체가 K-MeSH 시소러스 엔트리에 존재하는지를 검색하면 되므로 좀 더 쉽고 효과적으로 용어 추출을 할 수 있다. 또한 본 논문에서는 띄어쓰기 변이체를 인식하고 난 후, 형태소분석의 어미 분석 기능을 필요한 곳에서 사용하여 형태소 분석 방법의 잇점도 갖도록 하였다.

3.1 띄어쓰기 변이의 인식

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. 이전 시소러스(K-MeSH 1999년판)에서는 가능한 모든 띄어쓰기 변이를 문자열 일치 프로그램으로 찾아내기 위해 가능한 모든 변이를 등록하고 있다. (본 논문에서는 이 방법을 **일반 사전 방법(N-방법)**이라고 부른다.) 따라서, 이를 관리하기 위해 매우 번거로운 수작업이 필요하며, 또한 시소러스의 크기도 상대적으로 커진다.

이를 해결하기 위해 본 논문에서는 시소러스 사전에는 최대로 띄어 쓴 용어만을 등록하고, 검색에서는 해당 용어의 공백을 제거하는 방법을 사용한다. (이 방법을 **압축사전 방법(C-방법)**이라고 부른다.) 이를 처리하기 위해서는 문자열 일치 알고리즘을 간단히 변형하면 된다. (그림 2)는 변형된 문자열 일치 알고리즘을 C 프로그래밍 언어 형식으로 표현한 것이다. 주의해야 할 것은 t가 임의의 길이의 텍스트를 포인트하므로, s의 길이를 기준으로 문자열을 비교한다. 즉, s가 t 문자열의 앞부분에서 모두 일치해야 한다. 또한, 최장 일치 용어를 찾기 위해, 시소러스 사전에서 가장 길게 t와 일치하는 s를 찾도록 이 알고리즘을 반복 호출하여 사용한다.

(1)의 예인 경우, 최대로 띄어 쓴 (1h)가 시소러스에 등록되면, (1)에 표시한 모든 용어들이 검색될 것이다. 하지만 다

```
space_variant_string_match(s, t, sMax)
{
// s points to dictionary entry
// t points to target text
// sMax is the end of the entry word pointed by s

    while (s < sMax) {
    if (s* == t*) {
        s++; t++;
        }
    else
     if (s* == blank) s++;
     else return fail
    }
    if (s == sMax) return match
    else return fail
}
```

(그림 2) 띄어쓰기 변이를 고려한 문자열 일치 알고리즘

음의 (2)와 같이 띄어 쓴 용어들의 경우에는 검색되지 않을 것이다.

(2) a. 모세혈_관저항성
b. 모_세혈관저항성

이 방법은 띄어쓰기 변이를 하나만 저장하고도 가능한 띄어쓰기 변이를 모두 추출할 수 있기 때문에 검색의 재현률을 증가시킬 뿐만아니라, 시소러스의 크기를 줄이게 될 것이다.

3.2 문맥 내에서의 띄어쓰기 변이 인식

일반적으로 한 단위의 용어는 어절의 첫 부분(즉, 빈칸 다음)부터 시작된다. 하지만, 띄어쓰기가 규칙적이지 않아 다른 용어나 접두사를 앞에 붙여 쓰는 경우, 우리가 찾는 용어가 어절 중간에서부터 시작될 수도 있을 것이다. 본 논문에서는 이러한 두 가지 경우를 비교하기 위해 어절단위 검색과 음절단위 검색의 두 가지 방법으로 K-MeSH 용어를 검색한다. **어절단위 검색(word phrase based search; W-방법)**은 검색된 어절의 바로 다음 어절부터 검색을 시작한다. 이 방법은 사용된 모든 용어들이 공백으로 명확히 구분되어 있고, 반드시 어절의 처음 부분에 있다고 가정한다. **음절단위 검색(syllable based search; S-방법)**은 검색된 용어의 바로 다음 음절부터 검색을 시작한다. 이 방법은 한국어 의학 용어들이 다른 용어들과 함께 붙여 써서 사용 될 수 있다는 것을 고려한 것이다. 따라서 이 방법은 복합어 내에 사용된 K-MeSH 용어들도 추출해 낼 수 있다. (당연히, 음절 단위 검색이 더 많은 계산 시간을 필요로 하겠지만, 여기에서는 용어 추출의 성능을 더 중요시하여 비교한다.)

앞 절에서 설명한 띄어쓰기 변이를 고려하여 검색할 경우, 각 검색 전략의 결과는 다르게 수행될 것이다. (3)은 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 (3)에 나타난 B_C와 BC가 사용된 모든 경우의 키워드를 추출할 것이다. 반면에 띄어쓰기 변이를 고려하지 않을 경우에는 (3a), (3b), (3e), (3f)만 추출할 것이다. 어절단위 검색 전략에서는 용어의 시작이 공백 다음에 나타난 것만을 추출한다. 이때 띄어쓰기 변이를 고려할 경우에는 (3a), (3b), (3c), (3d)만 추출할 것이고, 띄어쓰기 변이를 고려하지 않을 경우에는 (3a)와 (3b)만 추출할 것이다. 또한 BC가 시소러스에 등록된 경우, 음절단위 검색 전략에서 띄어쓰기를 고려할 경우에는 (3c), (3d), (3g), (3h)만 추출할 것이고, 띄어쓰기를 고려하지 않을 경우에는 (3c)와 (3d)만 추출할 것이다. 그리고 어절단위 검색 전략에서는 띄어쓰기를 고려할 경우와 고려하지 않을 경우 모두 (3c)와 (3d)만 추출할 것이다.

(3) 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

3.3 인식된 용어의 검증

문장 내에서 용어에 대응되는 문자열을 찾아내더라도 그 문자열이 반드시 그 용어라고 결정할 수는 없다. 그 이유는 그 문자열 뒤에 다른 문자열이 붙어 다른 용어로 사용될 수도 있기 때문이다. 예를 들어 K-MeSH 용어의 '위'(胃, stomach)는 (4b)와 (5b)처럼 다른 단어의 일부를 잘못 인식할 수도 있다.

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

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

이런 현상은 짧은 용어에서 많이 나타나며, 특히 단음절어(One Syllable Word: OSW)에서 많이 나타난다. 이러한 문제를 해결하기 위해서는 찾은 문자열(용어 후보)이 다른 명사나 어미 등의 일부인가를 확인해야 한다. 본 논문에서는 이러한 문제를 비교적 간단하게 해결하기 위해 이음절어 이상으로 이루어진 문자열이 발견될 경우는 올바른 용어로 간주하여 처리하고, 단음절어일 경우는 조건에 따라 선택적으로 용어를 추출했다.

정확한 단음절어 추출은 쉽지 않을 뿐더러 논문의 키워드 선택에서도 그렇게 비중을 많이 차지하지 않는다. 단음절어를 다루기 위해 본 논문에서는 3가지 선택 방법, 1) **단음절어 전부 무시(X-방법)**, 2) **단음절어 모두 선택(A-방법)**, 3) **조건부 단음절어의 선택(O-방법)**을 두었다. 3)에서 조건이란, 단음절어 앞에 다른 명사나 접두사가 붙어 있지 않고, 또 단음절어 뒤에 다른 단어(어미나 접미사)가 붙어 있지 않거나 붙어 있다면 유효한 접미사가 붙어 있는 경우이다.

선택 방법 3)의 예를 들면 다음과 같다. (6)은 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, (6a)의 B가 추출될 것이고, (6b)의 B는 C가 유효한 접미사라면 추출될 것이다. 그러나 (6c)와 (6d)의 B는 단음절어 B앞에 A가 붙어있으므로, 추출되지 않는다.

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

위와 같은 방법으로 단음절을 처리하더라도 여전히 문제는 있다. 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 높이를 나타내는 위(上, above)나 지위를 나타내는 위(位, position)를 의미하기도 한다. (7)은 어절 분리가 정확하게 되었더라도 그 의미가 두 가지로 애매하여 불분명함을 보여준다. 문맥상으로는 (7c)의 경우일지라도 K-MeSH 용어 (7b)로 잘못 선택될 수도 있다. 애매성의 해결은 전체 문장, 단락, 심지어 전체 문서를 이해할 수 있는 자연언어 기술이 필요하므로 본 논문에서는 다루지 않는다.

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

4. 실험 및 결과

4.1 실험 방법

실험 데이터로는 298개 한국어 의학 저널[8]에서 44,285개 요약문으로 구성된 한국어 의학 데이터베이스(KMBASE)를 사용하였다. 각 실험에 따라 필요한 요약문을 선택하여 사용하였다.

실험은 크게 2단계로 진행하였다. 첫 번째 단계에서는 앞장에서 제시한 여러 가지 방법을 사용하여 K-MeSH 용어를 추출하여 비교하고, 두 번째 단계에서는 앞 단계에서 사용한 방법 중 성능이 가장 좋은 방법으로 키워드 추천을 한 후, 이를 전문가와 저자가 추출한 키워드와 비교하였다. 앞장에서 설명된 각 방법을 다시 정리하면 다음과 같다.

1. 사전과 문자열 일치 방법:
N-방법 : 모든 띄어쓰기변이를 포함한 사전으로 기존의 문자열 일치 방법을 사용
C-방법 : 압축 사전을 이용하여 띄어쓰기 변이 처리
2. 문장내 검색 위치 선정 방법:
W-방법 : 어절의 처음에서만 검색 시작
S-방법 : 어절내 임의의 음절에서 검색 시작
3. 단음절어(OSW)의 유효성 처리 방법:
X-방법 : 모든 단음절어를 무시하여 추출하지 않음
A-방법 : 모든 단음절어를 무조건 추출함
O-방법 : 접미사를 검사하여 유효한 단음절어만 선택

위의 방법에 대한 조합가능한 방법은 모두 12가지(2x2x3) 방법이고, 약어로 사용된 영어 알파벳 3개의 글자를 조합하여 한 가지 실험 방법을 표현한다. 예를 들어, 'CSX'는 C-방법, S-방법, X-방법을 조합하여 처리하는 방법을 나타낸다. 또, 하나의 약어만을 사용하는 경우, 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.

4.2 K-MeSH 용어 추출 실험

용어 추출 성능 실험을 위해서, 수작업으로 먼저 요약문으로부터 K-MeSH의 모든 용어를 추출했고, 이를 앞에서 설명한 각 방법의 프로그램 결과와 비교하였다. 용어 추출의 정확도를 확인하기 위한 것이므로 5개 정도의 요약문으로도 충분하다고 판단되어, 임의로 선택된 5개의 요약문에 대해서만 평가를 수행하였다. 이때 사용된 각 요약문의 평균 어절수는 178개이고, 전체 어절수는 891개이다. 평가시 띄어쓰기 성능 평가에 초점을 맞추기 위해 K-MeSH에 우선어가 등록된 용어만을 한정하여 비교하였다. 각 방법에 대해 K-MeSH 용어 추출의 평가 결과는 <표 1>과 같다. 여기에서 사용된 재현률과 정확률 및 이를 통합한 F-값은 아래의 (식 2), (식 3), (식 4)으로 계산된다. 본 실험의 평가에서 (식 4)의 F-값 계산시 α값을 0.5로 하여 정확률과 재현률의 중요도를 같은 비율로 반영하였다.

$$\text{재현률} = \frac{\text{정확하게추출한용어수}}{\text{전체용어수}} \quad \text{(식 2)}$$

$$\text{정확률} = \frac{\text{정확하게추출한용어수}}{\text{추출한용어수}} \quad \text{(식 3)}$$

$$F\text{-값} = \frac{1}{\alpha\frac{1}{\text{정확률}} + (1-\alpha)\frac{1}{\text{재현률}}} \quad (\text{단}, 0 \le \alpha \le 1) \quad \text{(식 4)}$$

C-방법(압축 사전 방법)은 모든 경우에 있어서 항상 N-방법(기존 사전 방법)보다 약간 좋게 수행되었다. 이는 이전 K-MeSH(1999년 번역판)가 가능한 모든 띄어쓰기를 입력했음에도 빠뜨린 것이 있음을 의미하며, 띄어쓰기 변이를 찾기 위한 방법이 유효함을 보여준다. 더욱이 이 방법으로 시소러스를 재구성하여 이전 K-MeSH(1999년 번역판)에 비해 K-MeSH 용어가 총 47,100개에서 약 27,300개로 줄었으며, 크기 또한 약 58%로 축소되어 공간 절약에 매우 효과적임을 보였다.

S-방법(음절단위 검색 전략)의 재현률은 항상 W-방법(어

<표 1> K-MeSH 용어 추출 결과

방법	평가 결과(%)		
	재현률	정확률	F-값
CSA	98.9	42.0	59.0
CSO	88.8	82.7	85.3
CSX	87.4	94.3	90.7
CWA	77.6	61.2	68.4
CWO	70.8	81.4	75.7
CWX	70.2	96.3	81.2
NSA	98.6	41.9	58.8
NSO	87.7	82.5	85.0
NSX	87.1	93.9	90.4
NWA	77.3	61.0	68.2
NWO	70.5	81.3	75.5
NWX	69.9	96.3	81.0

절단위 검색 전략)보다 높았다. 그러나 정확률은 단음절어에 대한 유효성을 검사했을 때만 좋았다. 즉 CSO(82.7%)가 CWO(81.4%)보다 높고, NSO(82.5%)가 NWO(81.3%)보다 높았다. 음절단위 검색을 할 경우에는 단음절어가 틀리게 추출되는 경향이 있음으로 유효성 검사를 하는 것이 더 정확함을 의미한다.

보다 적은 단음절어를 추출할수록 높은 정확률을 얻을 수 있는 반면에, 보다 많은 단음절어를 추출할수록 높은 재현률을 얻을 수 있었다. 또, 비록 유효한 단음절어를 얻기 위한 접미사 검사 프로그램을 사용한다 할지라도 기대하는 것만큼의 재현률을 향상시킬 수는 없었다. 본 실험에서 사용한 접미사 검사 프로그램은 가능한 조사 및 접미사의 조합을 검사해서 접미사를 판단해 주는 것이다[13]. 그러나 과도생성에 의해 잘못된 조합도 접미사로 처리하는 오류가 있어, 이를 향상시킬 필요가 있다.

CSA 방법은 모든 띄어쓰기 변이를 찾아낼 것으로 예상했었지만, 분리 오류 때문에 재현률이 100%가 아닌 98.9%를 결과로 내놓았다. (8)은 이때 발생한 오류의 예를 보여준다. '호흡'(respiration), '기질'(temperament), '질환'(disease)이 시소러스에 등록되어져 있을 때, (8b)와 (8c)의 두가지 방법으로 분리되어 질 수 있다. 본 실험에서는 시소러스에서 최장일치방법으로 추출하기 때문에, 비록 (8c)가 정확한 것이더라도 (8b)가 추출된다. 이런 경우는 많지는 않지만, 더 정확한 용어 추출을 위해 앞으로 연구되어져야 한다.

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

논문의 중요한 키워드를 계산하기 위해서는 관련된 용어를 많이 정확하게 추출하여야 한다. 하지만, 일반적으로 자동시스템의 경우, 많은 용어를 추출하다보면, 잘못된 용어가 많이 포함되고, 정확한 용어만을 추출하다보면 용어를 많이 빠뜨리게 된다. 따라서 정확한 평가를 위해서 재현률과 정확률 사이의 균형이 필요하다. 본 실험에서는 F-값 계산에 정확률과 재현률을 같은 비율로 고려하였고, 이 F-값이 가장 좋은 것을 키워드 추출 계산에 사용하였다. <표 1>에 나타난 실험결과에 따르면 기본적으로 쉽게 쓸 수 있는 방법인 NWA (압축되지 않은 사전을 사용하여 어절단위로 용어를 추출하고 모든 단음절어를 포함하여 추출하는 방법)는 68.2%이고, 가장 효과적인 방법은 CSX로 F-값이 90.7%이었다. 이는 약 22.5%의 성능이 향상된 것이다. 또한 CSX 방법이 다른 방법에 비해 가장 우수하므로, 이 방법을 교차언어 키워드 부여시 용어 추출 방법으로 선택했다.

4.3 한영 교차언어 키워드 부여의 정확성 평가

교차언어 키워드 부여 프로그램(KAP: Keyword Assignment Program)은 CSX방법을 한국어 용어 추출 방법으로 사용하고, 이 용어에 대응하는 영어 용어를 K-MeSH에서 찾아내어, 이 용어들의 중요도를 계산한 후, 이를 키워드로 제시한다. 용어의 중요도는 2장에서 설명한 바와 같이 TF와 IDF의 곱으로 계산된다. TF는 한 문서내에서 계산되므로, 특별히 전처리가 필요하지 않지만, IDF는 전체 문서 집합을 대상으로 계산되어져야 한다. 본 실험에서는 IDF를 한국어 의학 데이터베이스 (KMBASE)[8]에 있는 논문 25,729중 90%인 23,156개를 선택하여 계산하였다. IDF 계산에 사용되지 않은 논문 중 임의로 20개를 선택하여 이를 키워드 부여의 정확성 평가를 위해 사용했다. 평가를 하기 위해 수작업을 해야 하기 때문에 비교적 적은 수의 논문을 선정하였고, 이를 보완하기 위해 통계 분석을 통해 결과가 유의미한지를 검증하였다.

평가용으로 선택된 20개의 논문에 대해 색인 전문가 3인에게 가능하면 K-MeSH용어를 사용하여 각각 색인하도록 의뢰했다. (실제 색인 전문가들이 추출한 색인어는 분야(qualifier) 등의 정보가 포함되어 있으나, 이 실험에서는 순수 키워드만을 대상으로 비교하였다.) 일반적으로 색인은 전문가마다 차이가 있을 수 있으므로, 색인 전문가 3인중 2인 이상이 선택한 키워드만을 모아 정답으로 가정하였다. 이 결과와 저자의 키워드, KAP 프로그램으로 생성한 상위 10개의 키워드를 비교하였다.

평가는 이 정답에 대한 정확률(precision)과 재현률(recall)로 계산한 후, 다시 F-값을 계산하여 측정하였다. F-값은 정확률과 재현률을 통합하여 하나의 측정값으로 계산해주며, 정확률과 재현률의 중요도에 따라 차등하여 계산할 수 있으나, 4.2절에서와 같이 중요도가 같은 것으로 하여 계산하였다.

KAP 프로그램은 순수하게 K-MeSH 우선어만을 키워드로 사용하는데 반해, 색인 전문가나 논문 저자는 K-MeSH 우선어가 아닌 다른 용어를 키워드로 사용할 수도 있다. 이를 고려하여 색인 전문가나 논문 저자들이 사용한 용어 중 동의어나 유사어는 K-MeSH 우선어로 바꾸고, 그 이외 용어는 제외한 후 결과를 측정하였다.

<표 2>는 20개의 실험용 요약 문서 각각에 대한 색인주체별 F-값을 나타낸 것이다. 첫 행인 "일반용어 허용"은 일반 용어 색인을 허용하여 색인한 결과로 색인 전문가들의 색인이 저자나 KAP에 비해 월등하게 좋았다. 저자의 색인도 KAP에 비해 훨씬 점수가 높았다. 둘째 행인 "K-MeSH 용어 제한"은 K-MeSH 우선어로 한정하여 측정한 F-값이다. 여기에서도 색인 전문가들의 색인 점수가 저자나 KAP에 비해 훨씬 높았다. 하지만, 저자와 KAP의 차이는 거의 없었다.

이에 대한 통계적 유의성을 검증하기 위해 논문 저자와 KAP의 결과를 아래와 같은 가설로 독립표본 t검정을 시행하였다. 검증 결과, 유의수준 5% (p=0.906)로 H_0를 기각하지 못하여, 저자와 KAP의 결과가 차이가 없음을 보였다.

H_0: 저자의 색인 점수가 KAP의 점수와 같다.
H_1: 저자의 색인 점수가 KAP의 점수가 차이가 있다.

NLM의 경우, F-값 계산시, 재현률을 정확률에 비해 2배 더 비중을 두어 계산하였다. 일반 용어로 할 경우, 45.5%, 논문의 요점을 나타내는 MeSH의 중요 키워드(IM: Index Medicus)

로만 한정할 경우는 26% (R=81%, P=11%)이었다[2]. 같은 방법으로 KAP의 결과를 F-값으로 계산하면 일반 용어는 21.4%, K-MeSH 우선어로 한정할 경우 22.2%가 된다. 영어의 경우, 25개의 결과를 추출한 후, 색인 전문가의 색인어와 비교하였고, 영어 문서에서 바로 영어 색인어를 처리했으므로 본 논문의 경우와 정확하게 비교할 수는 없다. 하지만, 제한된 용어로 색인을 한정할 경우, 약 4%정도밖에 차이가 나지 않으므로, MeSH 한글화 과정의 오류나 불완전함 및 한글 처리의 복잡성 등을 고려해 볼 때, 어느 정도 의미있는 결과라고 볼 수 있다.

<표 3>은 실제 추출한 키워드를 보여준다. 추정 정답은 4개가 나왔고, 이중 '니코란딜' (Nicorandil)은 K-MeSH에 '니코랜딜'로만 등록되어 있다. 따라서, KAP은 'Nicorandil'의 음차 표기인 '니코란딜'을 인식하지 못해 영문 키워드 'Nicorandil'을 추출하지 못했다. 반면에 색인 전문가들은 모두 '니코랜딜'이나 '니코란딜'을 하나의 용어로 인식하여 영문 키워드 'Nicorandil'을 추출해 냈다. 만약 '니코란딜'이 '니코랜딜'의 유사 외래어로 K-MeSH 등록되었다면, KAP도 이 용어를 추출했을 것이다. 현재의 비교는 K-MeSH에 등록된 용어만으로 한정하므로, 일단, '니코란딜'은 K-MeSH 색인 비교에서는 제외시켰다.

추정 정답 색인어 중, 'Adenosine(아데노신)'과 'Blood Flow Velocity (혈류속도)'는 저자와 KAP이 모두 추출해 냈다. 'Coronary Circulation (관상동맥순환)'은 저자와 KAP이 모두 추출하지 못했는데, 이는 논문 요약에 이 용어가 포함되어 있지 않았기 때문이다. 하지만, 색인 전문가들은 주어진 용어 외에 그 문서의 특징을 나타내는 새로운 용어를 찾아내어 색인하였으므로, 새로운 용어인 '관상동맥순환'을 추출한 것이다. KAP은 새로운 용어를 제시하지는 못하지만, 시소러스 상에서 그와 유사한 관련 용어들을 제시하고 있다. 즉, 새 용어 'Coronary Circulation' (관상동맥순환: G09.330.163.324)과 시소러스 트리상으로 비교적 가까운 용어인 'Blood Flow Velocity' (혈류속도: G09.330.612.095), 'Diastole' (이완기: G09.330.800.295), 'Systole' (수축기: G09.330.800.880), 'Heart Rate' (심박수: G09.330.612.509) 등을 제시하고 있다.

KAP이 K-MeSH 용어로 된 정답 색인어를 추출하지 못한 이유는 크게 2가지로 종합하여 설명할 수 있다. 첫째로 논문에 사용된 용어가 K-MeSH내의 유사어 (상용어나 동의어)로 등록되지 않은 경우이고, 둘째로 논문내에서 사용한 용어만으로 직접 색인어를 찾을 수 없는 경우이다. 첫째, 유사어로 등록되지 않아 정답 키워드를 찾지 못한 예로는 유사어로 한글 단어나 외래어 이형태 등이 포함되지 않은 경우 ('니코랜딜'에 대한 유사어인 '니코란딜'이 없음), 외래어와 한글을 혼합하여 표기함으로써 찾지 못한 경우 ('그레이브스병'을 'Graves병'으로 사용함), 약어로 나타내서 찾지 못한 경우 ('중화효소연쇄반응'을 'PCR'로 사용함) 등이다. 이를 해결하기 위해서는 다양한 형태를 K-MeSH내에 더 포함시키거나, 이형태를 쉽게 인식할 수 있는 기법이 필요하다. 둘째, 논문내에서 사용한 용어만으로 직접 색인어를 찾을 수 없는 예로는 앞에서 설명한 'Coronary Circulation(관상동맥순환)' 등과 같이, 논문에 나타난 유사한 용어들을 포괄적으로 나타내는 경우이다. 이와 같은 경우를 처리하기 위해서는 전문 색인가와 같이 기존의 지식을 활용하여 추론하거나, 일반화할 수 있는 기술이 필요하다.

〈표 2〉 각 문서별 색인주체의 F-값 (20문서 평균)

	전문가A	전문가B	전문가C	저자	KAP
일반용어허용	76.7%	67.9%	85.0%	19.7%	19.6%
K-MeSH 용어 제한	76.3%	69.5%	84.4%	21.3%	20.1%

〈표 3〉 색인주체별 문서1에 대한 추출 키워드 예
(밑줄친 단어는 정답과 일치하는 용어, 이탤릭체의 단어는 K-MeSH 용어가 아닌 일반 용어임)

저자	*Coronary Flow Reserve*, Nicorandil, Adenosine, Doppler Ultrasonography, Blood Flow Velocity
KAP	Arteries, Hematocele, Adenosine, Blood Flow Velocity, Diastole, Systole, Lipids and Antilipemic Agents, Injections, Patients, Heart Rate
정답	Nicorandil, Adenosine, Blood Flow Velocity, Coronary Circulation, Doppler Ultrasonography, Coronary Vessels
전문가 A	Nicorandil, Adenosine, Blood Flow Velocity, Coronary Circulation, Doppler Ultrasonography, Comparative Study
전문가 B	Adenosine, Angiocardiography, Blood Flow Velocity, Coronary Angiography, Coronary Circulation, Coronary Vessels, Electrocardiography, Nicorandil, Vasodilator Agents
전문가 C	Coronary Circulation, Coronary Vessels, Doppler Ultrasonography, Nicorandil, Adenosine, Blood Flow Velocity

5. 결 론

본 논문에서는 한국어 의학 논문에서 자동으로 영문 MeSH 키워드를 부여하는 방법을 제안하고, 이에 수반되는 K-MeSH 용어 추출과 띄어쓰기 변이 문제 처리 방법을 제안했다. 띄어쓰기 변이를 처리하기 위해, 시소러스에 생성 가능한 모든 띄어쓰기 변이를 등록하는 것 대신에 최대로 띄어쓴 용어만 등록하고 이를 이용하여 모든 변이를 검색해 내도록 했고, 문장내에서도 음절 단위로 용어를 검색하도록 했다. 실험 결과, 시소러스의 크기가 약 42%정도 축소되었을 뿐만 아니라, 일반적인 용어 추출 방법에 비해 약 22.5% (F-값) 향상되었다.

또 제안한 방법 (CSX 방법)을 사용하여 교차언어 키워드 부여 실험을 하였다. 색인 전문가 3인의 색인 결과에서 2인 이상이 일치하는 키워드를 정답 색인어로 가정하고 평가를 해본 결과 색인 전문가들이 평균 F-값은 76.7%이고, 저자의 경우는 21.3%, 본 프로그램의 경우는 20.1%이었다. 현재 교차언어 키워드 부여 프로그램(KAP)의 성능은 색인 전문가보다는 훨씬 못했지만, 일반 저자들의 결과와는 통계적으로 별

차이가 없었다.

본 논문에서는 간단한 부분문자열 일치 방법과 간단한 접미사 검사 프로그램을 사용했으므로 성능에 한계가 있었다. 부분 파싱이나 태깅 방법 등과 같은 더 정교한 자연 언어 처리 기술이나 의미 정보나 문맥의 정보를 추가로 사용한다면 교차언어 키워드 부여의 성능을 더 향상시킬 수 있을 것이다. 또한, 다른 여러 가지 이형태 용어를 더 인식할 수 있도록 시소러스를 보완하고, 시소러스내의 용어에 대한 모호성 처리를 할 경우, 더 정교한 키워드 부여를 할 수도 있을 것이다.

감사의 글

본 연구를 위해 실험용 문서 색인을 도와주신 카톨릭 의대 사서 정소나님, 삼성의료원 사서 조혜민님, 연대 의대 사서 김미희님께 감사드립니다.

참 고 문 헌

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

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

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

[4] Hersh, W., Buckly, C., Leone, TJ. 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, pp.192-201, 1994.

[5] Srinivasan, P. Optimal document indexing vocabulary for MEDLINE. Information Processing & Management, Vol.32, No.5, pp.503-514, 1996.

[6] 김병선, 김수영. 가정의학회지 논문의 영문 주제어 선택에 있어서 MeSH용어 사용 여부와 선택 정확도. 대한가정의학회지, Vol.19, No.17, pp.531-537, 1998.

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

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

[9] Manning, Christopher D., Schutze, Hinrich. Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, pp.244-247, 1999.

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

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

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

[13] 강승식. 한국어 형태소 분석과 정보 검색. 홍릉과학출판사, 2002.

이 재 성

e-mail : jasonl@cbu.ac.kr

1983년 2월 서울대학교 컴퓨터공학과(학사)

1985년 2월 한국과학기술원 전산학과(석사)

1999년 2월 한국과학기술원 전산학과(박사)

1985년~1988년 큐닉스컴퓨터 개발부 과장

1988년~1989년 Microsoft(미국), software design engineer

1988년~1993년 마이크로소프트 개발부 차장

1999년~2000년 한국전자통신연구소 선임연구원/팀장

2000년 9월~현재 충북대학교 컴퓨터교육과 조교수

2005년 8월~현재 University of Arizona, research scholar

관심분야 : 정보검색, 자연언어 처리, 한글공학, 컴퓨터교육

김 미 숙

e-mail : htkms@naver.com

2000년 2월 충주대학교 전자계산학과(학사)

2005년 2월 충북대학교 정보컴퓨터교육과 (석사)

2005년 3월~현재 (재)중부직업전문학교

관심분야 : 정보검색, 자연언어 처리, 한글공학, 컴퓨터교육

오 영 순

e-mail : misogiyoo@empal.com

2000년 2월 충북대학교 컴퓨터교육과(학사)

2005년 2월 충북대학교 교육대학원 정보컴퓨터교육과(석사)

현 재 오창고등학교 교사

관심분야 : 정보검색, 컴퓨터교육

이 영 성

e-mail : young@medric.or.kr

1987년 2월 서울대학교 의과대학 의학과 (학사)

1992년 2월 서울대학교 의과대학원 의학과 의료관리학전공(석사)

1996년 2월 서울대학교 의과대학원 의학과 의료관리학전공(박사)

1999년 11월~현재 과학기술부 한국과학재단 지정 의학연구정보센터 소장

충북대학교 의과대학 의학과 의료정보학 및 관리학교실 부교수

관심분야 : 원격의료, 의료정보, 의학교육