전문용어의 처리에 의한 도메인 온톨로지의 구축
(Domain-specific Ontology Construction by Terminology Processing)

임 수 연 † 송 무 희 †† 이 상 조 †††
(Soo-yeon Lim) (Mu-hee Song) (Sang-jo Lee)

요 약 온톨로지는 특정 도메인에 사용되는 용어들과 그 용어들 간의 관계를 정의하고, 이를 계층구조로 표현한 것을 말한다. 본 논문에서는 전문용어의 처리에 기반한 도메인 특정적인 온톨로지의 반자동 구축방안을 제안하고자 한다. 이를 위하여 도메인 텍스트 내에서 전문용어를 구성하고 있는 명사나 접미사의 패턴을 분류하고, 이에 따라 전문용어를 추출하고 계층구조를 구하는 알고리즘을 제안한다. 실험은 약학 관련 문서를 대상으로 하였으며, 단일어절 전문용어를 인식한 결과 평균 92.57%, 다중어절 전문용어의 경우 평균 66.64%의 정확도를 보였다. 구축된 온톨로지는 의미정보와 함께 전문용어를 구성하는 특정 명사나 접미사를 중심으로 자연스런 의미 군을 형성함으로써 정보검색 등의 전문적인 지식의 접근에 유용하게 쓰일 수 있으며, 검색의 성능을 향상시키기 위한 추론의 기반으로도 이용할 수 있다.

키워드 : 온톨로지, 전문용어, 개념, 관계

Abstract Ontology defines the terms used in a specific domain and the relationships between them and represents them as hierarchical taxonomy. The present paper proposes a semi-automatic domain-specific ontology construction method based on terminology processing. For this purpose, it presents an algorithm to extract terminology according to the noun/suffix pattern of terminology in domain texts and find their hierarchical structure. The experiment was carried out using pharmacy-related documents. As singleton terminology with noun/suffix were identified, the average accuracy was 92.57%. In case of multi-word terminology, the average accuracy was 66.64%. The constructed ontology forms natural semantic clusters with based on suffices and semantic information, so can be utilized in approaches to specific knowledge such as information look-up or as the base of inference to improve searching abilities.

Key words : ontology, terminology, concept, relation

1. 서 론

온톨로지에 관한 연구는 인공지능 분야의 시작과 함께 지식 표현 분야의 핵심으로 활발히 연구가 이루어져 온 분야이다[1]. 시멘틱 웹의 출현과 더불어 온톨로지의 중요성이 인식되면서 이 분야의 연구는 새로운 연구 분야가 아닌 새로운 응용분야로 보는 것이 타당할 것이다. 온톨로지는 주어진 응용 도메인의 특성을 나타내는 관련 개념들의 집합과 정의(definition) 그리고 그들 간의 관계로 이루어진다. 문서나 웹을 검색할 때 온톨로지를 사용하면 중요한 정보가 있는 자원을 빠르게 찾아 사용할 수 있다는 것과 자원을 찾는 정확도를 향상시키는 것이 대표적인 이점이다. 검색엔진은 온톨로지에 정의된 개념과 규칙들을 활용하면서 이를 검색을 향상시키기 위한 추론의 기반으로 이용할 수도 있다.

크고 복잡한 응용 도메인의 경우 온톨로지의 구축작업은 시간이 너무 오래 걸리고 비용이 많이 들며 같은 개념에 대해서도 사람마다 다른 관점을 가지므로 논쟁의 여지가 많다. 이들은 대부분 수작업으로 구축되어 왔지만 이 방법은 상당한 시간과 비용이 들므로 최근에는 온톨로지를 반자동으로 구축되기 위한 방안이 활발히 연구되고 있다.

시멘틱 웹의 성공은 온톨로지의 확장정도에 달려있다. 온톨로지를 구축하고 갱신할 때의 시간과 비용을 줄이기 위해 온톨로지의 학습은 유용하며, 해당 도메인의 개념들과 그들 간의 의미관계를 추출하는 텍스트 마이닝(text mining)기술이 매우 중요하다[2]. 온톨로지 학습

† 학생회원 : 경북대학교 컴퓨터공학과
nadalsy@hotmail.com
†† 학생회원 : 경북대학교 컴퓨터공학과
mhsong@knu.ac.kr
††† 종신회원 : 경북대학교 컴퓨터공학과 교수
sjlee@knu.ac.kr
논문접수 : 2003년 9월 18일
심사완료 : 2003년 12월 5일

은 전혀 구조화되지 않았거나 반구조화, 혹은 완전히 구조화된 여러 가지 데이터 유형들을 대상으로 하여 온톨로지의 구축작업을 반자동으로 이루어지게 해준다. 그러나 자동화된 방법을 사용하더라도 고품질의 의미 지식베이스를 구축하기 위해서는 개념체계를 구축하는 핵심적인 부분이 수작업으로 만들어 진다.

온톨로지를 (반)자동으로 구축하는 방안들은 기존의 시소러스나 사전 등과 같은 기존의 자원을 이용하는 방법[3]과 기존의 자원을 이용하지 않고 텍스트의 분석결과로 얻어지는 단어들의 분포를 이용하여 베이스 온톨로지를 구축하고 확장하는 방법[4] 등이 있다. 전자의 경우에는 개념이 부착된 대용량의 사전을 이미 확보함으로써 추가의 사전작업 없이 바로 활용할 수 있는 지식베이스를 구축할 수 있으며 후자의 경우에는 개념들의 확장이 용이하다. 두 방법 모두 고품질의 의미관계 패턴을 추출하는 것이 매우 중요하다. 보다 많은 의미관계 패턴을 추출하기 위하여 코퍼스 내의 텍스트에 출현하는 용어들의 형태를 분석하였다. 그 결과, 전문용어들이 많이 발견됨에 따라 이들에 대한 처리가 필요하였다.

기존의 전문용어에 대한 연구는 크게 규칙에 기반한 방법과 통계에 기반한 방법으로 나뉠 수 있다. 규칙에 기반한 방법은 전문용어 구성의 패턴을 수작업으로 구축하거나 학습 코퍼스를 만든뒤 이를 이용하여 자동으로 인식 패턴을 구축함으로써 전문용어를 인식한다[5]. 이 때 명사 사전, 접사 사전 등과 같은 일반 사전을 이용하며 주로 사람이 직접 규칙을 기술하므로 비교적 정확한 결과를 보여준다. 통계에 기반한 방법은 학습 코퍼스로부터 인식에 필요한 지식을 학습하는데 은닉 마르코프 모델이나 최대 엔트로피 모델, 문자형, 어휘정보 등과 같은 지식을 이용한다[6-8]. 특히 [9]에서는 기술문서들을 대상으로 다중 어절형태로 이루어진 전문용어들을 추출하고 WordNet[10]에 포함된 개념간의 관련성 정보를 이용하여 이들의 관계를 온톨로지에 구축하고 확장해 나가는 방법을 취하고 있다.

본 논문에서는 한국어 문서 내에 복합명사의 형태로 출현하는 전문용어들의 패턴들을 분류하고 이들의 구조를 분석한다. 그리고 분석한 결과로부터 의미군과 계층구조를 이끌어내어 온톨로지내의 의미관계를 부여해주는 알고리즘을 제안한다. 복합명사를 이루기 위해 첨부되는 접미사나 명사의 형태에 따라 분류된 전문용어들은 관련 도메인 내에서 자연스러운 군집현상을 일으키며 각각의 의미군을 형성한다. 관련도메인에 대한 첨부명사나 접미사군을 잘 활용하고 의미군을 파악할 수 있다면 전문적인 지식의 접근이 가능하다. 예를 들어 "문학, 의학, 공학..." 등과 같이 명사 "학(學)"으로 연결된 개념들은 학문(學文)을 나타내므로 온톨로지 내에서 군집을 형성한다. 제안한 방법에 의해 구축된 온톨로지는 정보검색 등의 전문적인 지식의 접근에 유용하게 쓰일 수 있다. 실험 도메인은 약학 분야로 정하고 약품 매뉴얼에 있는 텍스트를 대상으로 실험을 행하였다.

2. 온톨로지의 구축

온톨로지는 어떤 특정 도메인에서 사용되는 정보들과 그 정보들 간의 관계를 정의해놓은 것으로 해당 도메인 전문가들과의 협의에 의하여 개념들과 관계들의 구조를 정한 뒤 이들을 기반으로 구축하게 된다. 실제의 응용시스템에서는 도메인마다의 특정적인 지식을 포함하는 온톨로지가 필요하다. 왜냐하면 경제, 의학, 공학 등과 같이 각 분야마다 사용되는 개념이 약간씩 다르기 때문이다.

2.1 구축단계

본 논문에서는 관련 도메인 코퍼스 내의 문서들을 학습시킨 결과를 이용하여 온톨로지를 반자동으로 구축하는 방안을 제시하고자 한다. 이 때 학습을 위한 문서들은 약학 도메인 내의 문서들로 한정한다.

그림 1은 제안한 온톨로지 구축의 대략적인 과정을 보여주는데, 네 단계의 과정들로 나뉜다. 첫 번째 단계에서는 관련 도메인내의 웹 문서들을 구조화하여 코퍼스를 형성하고, 두 번째 단계에서는 간단한 자연어처리과정을 거친 뒤, 개념들을 추출한다. 세 번째 단계에서 전문용어들을 추출하고 이들의 구조를 분석한 결과로부터 계층구조를 구한 뒤, 추출한 관계들을 온톨로지에 추가한다. 각 단계에 대한 자세한 설명은 다음과 같다.

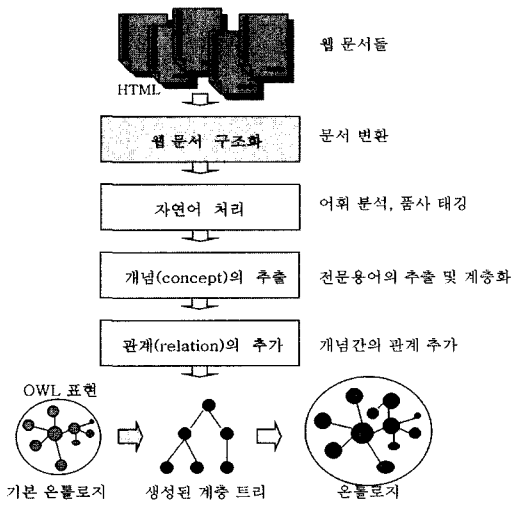


그림 1 온톨로지의 구축과정

2.2 온톨로지의 구조와 표현

본 논문에서는 구축할 온톨로지의 구조를 정하기 위하여, 웹상에 존재하는 약품과 관련이 있고 신뢰성이 있는 데이터베이스로 BITDruginfo(http://www.druginfo.co.kr)를 정하였다. 그 문서내의 데이터들을 분석한 결과를 이용하여 구축할 약품 온톨로지의 개념들과 이들을 연결시킬 관계들을 설정하였다. 그림 2는 설정된 개념들과 관계들로 이루어진 온톨로지의 구조를 개념 그래프로 나타낸 것이다. 약품명을 비롯한 각 개념들은 이 데이터베이스를 분석한 결과로부터 얻을 수 있으며, 약품명은 제조회사, 보험코드, 성분코드, 효능효과 등의 서브 카테고리로 나눌 수 있었다.

전문가들과 상의한 내용을 토대로 이 중 질의에 대한 응답을 위해 반드시 필요하다고 판단된 그림 2의 개념들과 관계들을 설정하였다. 병명이나 증세가 입력으로 들어오는 질의응답 시스템인 경우, 질의에 해당하는 약품명을 응답으로 돌려주기 위해서는 설정한 약품명과 관련이 있는 개념들 중 특히 <Effect: 효능효과>정보에 대부분 의존한다. 따라서 이 개념에 들어있는 정보를 집중적으로 분석하고 처리하고자 한다.

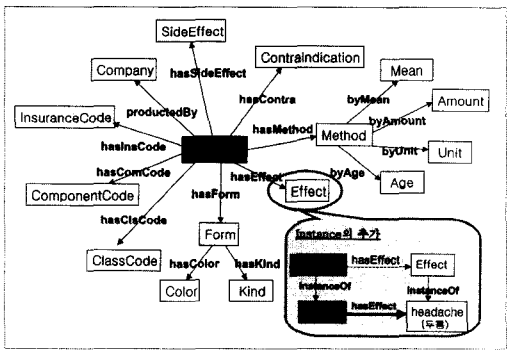


그림 2 설정된 개념들과 관계들로 이루어진 온톨로지의 구조

약품과 관계있는 수집 문서들은 반구조화된 문서로 학습(learning)을 위한 코퍼스를 형성하기 위해 설정한 그림2의 구조에 맞도록 변환과정을 거친다. 그리고 자연어처리 기술을 이용하여 텍스트 분석을 수행한다. 이때, 분석한 문서 내에서 특정 내용을 포함하는 구문 패턴이 존재하는 경우 그 패턴에 따라 텍스트들을 분류하고 태깅을 해준다. 예를 들어 약품에 관한 설명서를 대상문서로 할 경우, 문서들은 그림 3과 같이 패턴별로 태깅된 텍스트들로 이루어지며 태그들은 해당 약품에 대한 개념들을 각각 형성한다.

본 논문에서는 구축할 온톨로지에 존재하는 개념들과

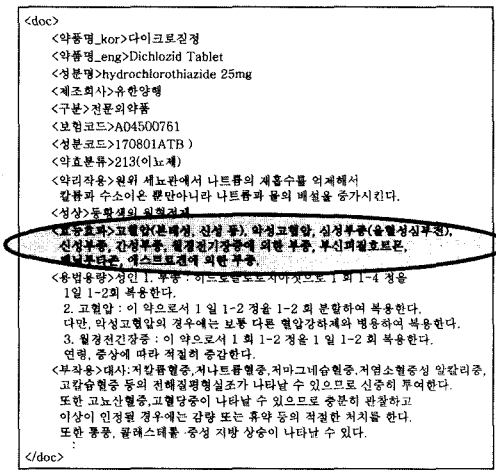


그림 3 형성된 코퍼스내의 태깅된 텍스트 구조의 예

```
<?xml version="1.0"?>
<!DOCTYPE owl [<!ENTITY owl  "http://www.w3.org/2002/07/owl#" >...]>
<rdf:RDF  xmlns:owl = "http://www.w3.org/2002/07/owl#"
            xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:rdfs= "http://www.w3.org/2000/01/rdf-schema#"
            xmlns:xsd = "http://www.w3.org/2000/10/XMLSchema#">
  <owl:Ontology rdf:about="">
        <rdfs:comment>Medicine OWL ontology</rdfs:comment>
        <rdfs:label> Medicine Ontology</rdfs:label>
  <owl:Class rdf:ID=" Medicine ">
        <rdfs:subClassOf rdf:resource="eating" />
        <rdfs:subClassOf>
          <owl:Restriction>
                <owl:onProperty rdf:resource="#productedBy" />
                <owl:allValuesFrom rdf:resource="#Company" />
          </owl:Restriction>
        </rdfs:subClassOf>
                :
        <rdfs:label xml:lang="eng"> Medicine </rdfs:label>
        <rdfs:label xml:lang="kor">약품명</rdfs:label>
  </owl:Class>
</rdf:RDF>
```

그림 4 OWL로 표현된 온톨로지(일부분)

그들의 관계를 OWL[11]을 이용하여 표현하고 한글과영어를 혼용한다. 다음의 그림 4는 개념 "Medicine: 약품명"을 OWL로 표현한 간단한 예를 보여준다.

2.3 개념의 추출

문서내의 태깅된 텍스트들은 간단한 자연어 처리 파서를 이용한 텍스트 분석 과정을 거치게 된다. 먼저, 텍스트내의 불용어(stop word)를 제거하고 스테밍을 거친 후, 문서내의 모든 명사들을 추출한다. 추출한 명사들은 온톨로지의 개념을 나타내고, 문서에 붙은 태그들과 동사들은 개념들 간의 관계를 나타내며 개념들을 연결짓는 링크로서의 역할을 한다. 즉, 온톨로지는 많은 어휘들로 구성된 네트워크의 일종인 것이다.

약품 매뉴얼의 <Effect: 효능효과>텍스트로부터 추출한 명사들을 살펴보면 병명, 증세, 성분 등을 나타내는 많은 고유명사나 복합명사들이 등장함을 알 수 있었다. 이는 각 도메인마다 고유한 특성을 나타내는 개체명이

존재하고 전문용어에 대한 별도의 처리방안이 요구됨을 의미한다. 전문용어를 추출하고 계층화하여 이를 온톨로지에 추가하는 알고리즘은 3장에서 설명한다.

2.4 관계의 추가

제안한 전문 용어의 처리방법(3장)에 따라 추출된 관계들은 주변에 나타난 명사들을 연결짓는 의미관계를 나타낸다. 다음의 그림 5는 예제 텍스트를 간단한 자연어처리 파서를 거치게 한 뒤, 그로부터 추출한 개념과 관계를 추가한 온톨로지의 일부분을 보여준다.

위의 그림에서 입력된 문장을 파싱한 결과, 해당약품(Medicine)에 대하여 개념인 급성 췌장염(Acute Pancreatitis)과 부작용(hasSideEffect), 감량(decrease)의 관계를 추출한다. 특히 개념 "Acute Pancreatitis"는 현재 온톨로지에 존재하지 않는 전문용어로 3장에서 제안한 전문용어 처리과정을 거치게 되고, 상위개념 Pancreatitis(췌장염)와 Inflammation(염)를 추출한다. 추출한 상위개념이 이미 해당 온톨로지에 존재하는 경우에는 "hyponymOf"관계만 이어주면 되지만, 그렇지 않은 경우에는 이들 상위개념을 추가하며 관계를 연결함으로써 온톨로지를 확장해간다.

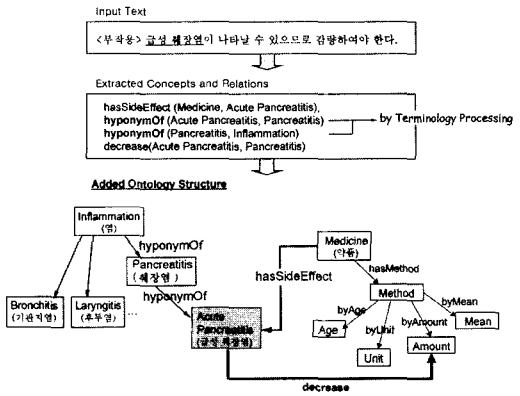


그림 5 추출한 개념과 관계를 추가한 온톨로지(일부분)

3. 전문용어의 추출

실험 대상 문서 내 <효능효과> 태그의 텍스트들을 분석한 후 추출한 용어들은 병명이나 증세를 나타내는 전문용어(terminology)들로 이루어져 있다. 그 원인은 전문적인 지식을 포함하는 약품 도메인내의 문서들이란 특성 때문인 것으로 추측된다. 전문용어란 주어진 도메인 안에서 의미를 가지고 있는 단어들의 집합을 의미한다. 낮은 모호성과 높은 특정성 때문에 이들 단어들은 지식도메인을 개념화하거나 도메인 온톨로지를 만들 때 매우 유용하다. 어휘특정성(specificity)이 큰 전문용어에 대한 처리는 풍부한 의미정보를 지니는 온톨로지의 구축을 가능하게 한다. 기존 사전에 없는 단어들인 전문용어들을 처리하기 위하여, 현재 관련분야의 전문가들이 수작업으로 전문용어를 추출하는 경우가 많다. 이런 경우 객관적인 용어의 추출이 어렵고 많은 시간과 노력이 소모된다.

본 논문에서는 전문용어들을 자동으로 추출하기 위하여 그들의 출현형태를 분석하였다. 해당 도메인에 출현하는 대부분의 전문용어들은 복합명사의 형태로 출현하였으며, 크게 두 가지의 결합형태로 나눌 수 있다. 하나는 단일어절(singleton term) 즉, 띄어쓰기가 없는 한 어절로 나타나는 단순한 결합형태이고, 다른 하나는 다중어절(multi_word term) 즉, 띄어쓰기가 나타나며 앞의 어절성분과 의미적으로 관련이 있는 두 어절이상으로 이루어진 복합명사이다.

3.1 단일어절의 형태

약품 도메인 내에서 전문용어를 이루고 있는 복합명사들은 한자어로부터 파생된 경우가 많으며 두 가지의 형태로 나뉜다. 하나는 명사와 명사가 결합한 형태이고, 다른 하나는 명사와 접미사가 결합한 형태이다. 본 논문에서는 복합명사를 구성하기 위해 결합되는 명사나 접미사를 20가지로 분류하였다. 이들은 "염, 증, 통, 균, 성, 질환, 속, 염증, 진, 감, 종, 병, 열, 궤양, 선, 백선, 증후군, 형, 환, 군"이며 또한 의미적으로 관련이 있는 전문용어들을 서로 연결시킨다. 표 1은 특정 명사나 접미사와 결합된 전문용어목록의 일부분을 보여준다.

약품도메인에서 단일어절의 형태로 출현하는 위 목록의 전문용어들은 "방광염, 기관지염"과 같이 특정명사(감염증을 나타내는 "염")의 하위단어인 경우가 대부분이다. 따라서 이들은 그림 6과 같이 "hyponymOf" 관계로 연결한다.

표 1 단일어절형태 전문용어(일부분)

염	방광염, (비만성범세, 급성, 만성)기관지염, 편도염, 중이염, ...
통	두통, 요통, 근육통, 관절통, 신경통, ...
균	인플루엔자균, 폐렴구균, 포도구균, (화농성)연쇄구균, ...
성	(퇴행성)관절질환, 만성호흡기질환, 비관절류마티스질환, ...
속	엘라속, 살모넬라속, ...
염증	요로감염증, 피부감염증, 칸디다감염증, ...
진	습진, 담마진, 단순포진, ...
증	패혈증, 기관지확장증, 협심증, ...
종	간혈성, 확장성, 급성, 뇌성, ...
병	불쾌감, 팽만감, 독감, ...

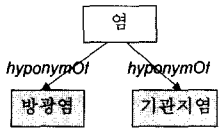


그림 6 단일어절형태 전문용어의 관련

관련이 있는 대상(Object)이나 행위자(Agent)등과 같은 수평적인 관계는 일반적인 텍스트 분석에 의해 얻어질 수 있으나 상위(Hypernym)나 하위(Hyponym)등과 같은 수직적인 관계는 코퍼스로부터 추출하기가 어렵다. 다음은 단일어절의 형태로 나타나는 전문용어 내에서 하위관계를 자동으로 추출하기 위한 알고리즘을 자바 언어 형식으로 표현한 것이다.

```
입력 : 접미사와 결합한 단일어절 전문 용어들 (word[1..n])
출력 : 전문 용어들간의 계층 트리

String Suffix[]={염,증,통,균,성,질환,속,염증, 진, 감, 종,병,열,궤양,
                 선,백선,중후군,형,환,균}
boolean matrix[][];     // 계층관계 matrix

// 모든 단어에 대해서 그 단어가 다른 단어에 포함되는지 조사
MakeSubTree {
   for (int i; i<n; i++)
      for (int j; j<n; j++)
         if (i!=j && (word[j].endsWith(word[i]))
            matrix[i][j]=true;
         if (word[i].length()==1)    // 1음절 어휘는 제외
            matrix[i][j]=false;
}

// 온톨로지 노드에 서브트리를 추가
AppendSubWords {
   for (int i; i<n; i++)
      if (processed[i]=false)  // 단어가  다른 단어의 Substring인지 조사
         boolean isSuper = true;  // 최상위 단어일 때만 하위 추가
         for (int j; j<n; j++)
            if (matrix[j][i]=1)
               isSuper=false;
            // i번째 단어의 오른쪽 하위 단어를 찾아서
            // i노드의 하위에 추가.
            if (isSuper=true && appendSubWords(i)=null)
               root.appendChild;
            processed[i]=true;
}
```

그림 7 단일 어절 형태의 전문 용어로부터 계층 관계를 추출하는 알고리즘

추출된 전문용어들은 결합하는 명사나 접미사의 종류에 따라 의미군을 형성한다. 그러나 제안된 전문용어 인식방법을 일반 도메인으로 확장했을 때에는 문제가 발생할 수 있다. 예를 들면 "천일제염"과 "후두염"을 같이 인식하고 하나의 군으로 형성하는 것이다. 그러나 이들은 아무 의미관계가 없는 단어들이므로 온톨로지에서 하나의 군으로 연결되는 것은 바람직한 일이 아니다. 따라서 관계를 이어줄 때 제약이 필요했다. 예를 들면, "염"이란 명사는 "기관지, 후두,..." 등의 인체기관과 관계있는 어휘와 결합했을 때만 병명으로 하위관계를 가지도록 하는 것이다. 이로써 구축된 온톨로지의 관계들은 풍부한 의미관계를 가질 수 있을 것이다.

3.2 다중어절의 형태

실험 텍스트에서 나타난 전문용어들은 대부분 "만성 위염"과 같이 수식어와 중심어의 관계를 가지는 다중어절의 형태로 나타났다. 실험 도메인에서 나타난 이 다중어절 형태의 전문용어들은 중심어가 다시 단일어절로 이루어진 전문용어로 이루어진 경우가 많았다. 우리는 이 관계들을 두 어절 이상으로 이루어진 전문용어의 처리에 이용하며, 다섯 개의 관계 패턴들을 설정하고 이에 따라 온톨로지 내에서의 의미관계를 추가하고 설정할 것이다.

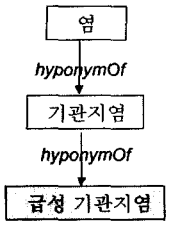
예를 들어, 이중어절의 형태로 출현하는 전문용어(N1 + N2)의 경우 중심어인 N2는 앞서 출현하는 어휘인 N1의 접미사나 조사와의 결합형태에 따라 특정 관계를 가진다. 그림 8은 설정한 패턴의 종류와 그에 따른 관계 설정 방안을 나타낸다.

패턴	패턴의 형태 관계 설정 방안 (예)
패턴1	N1(~성, ~형)+N2 N1N2를 N2로부터 확장된 전문용어로 보고 N2의 하위개념으로 연결 (예) 급성 기관지염, 만성 피로
패턴2	N1(~에 의한, ~(으)로 인한, ~(으)로 인해 유발된)+N2 N1(~에 따른)+N2 N1(~시(의), ~상태에서, ~후(의))+N2 N2는 관계 causeTo에 의해 N1과 연결 N2는 관계 accompanyWith에 의해 N1과 연결 N2는 관계 stateOf에 의해 N1과 연결 (예) 농부에 의한 출혈, 수술시 국소마취
패턴3	N1+"의"+N2, N1+N2 N1N2를 전문용어로 추출 (예) 근이완의 유지 --> 근이완유지
패턴4	N1+"및"+N2, N1+"·"+N2, N1+"또는"+N2 N1과 N2 각각을 전문용어로 추출 (예) 소화효소결핍 및 담즙분비촉진-->소화효소분비, 담즙분비촉진
패턴5	N1 (suffix_ 1)+"· ,및"+N2(suffix_ 2)+N3 (if suffix_ 1=suffix_ 2) N1N3와 N2N3 를 전문용어로 추출 (예) 지연형·활동성 만성간염-->지연형 만성간염, 활동성 만성간염

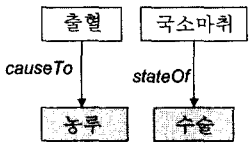
그림 8 다중어절형태 전문 용어 패턴들과 관계 설정 방안

위의 패턴들 외에도 "1차피부감염(농가진, 농창, 심상성모창, 조갑주위염)"의 예에서 볼 수 있듯이 [N1 + "(" + (N2, N3, ...) + ")"]와 같은 패턴이 존재한다. 일반적인 도메인에서 괄호안의 구들은 대개 대역어구를 나타내지만 실험대상 문서 내에 나타나는 괄호안의 구들은 특별히 하위관계를 나타낸다. 따라서 이와 같은 패턴인 경우 실험도메인에 한해 예외적으로 처리해주며 N1을 관계 hyponymOf에 의해 N2, N3, ...와 연결한다. 괄호안의 구들인 "농가진, 농창, 심상성모창, 조갑주위염" 각각의 개념들은 "1차피부감염" 개념과 hyponymOf관계로 연결된다. 그림 9는 앞에서 보여준 패턴별 처리방안의 일부를 그래프로 표현한 것이다. 패턴 3, 4, 5는 관계를 가지지 않고 다만 개념만 분리되어 생성되는 경우

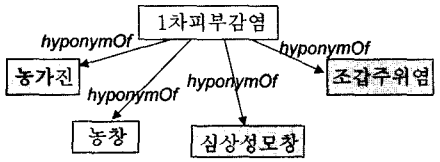
이므로 아래의 그림에서 생략하였다.



[패턴1] 급성 기관지염



[패턴2] 농루에 의한 출혈 수술시 국소마취



[기타] 1차피부감염(농가진, 농창, ...)

그림 9 그래프로 나타낸 관계설정의 예

4. 실험 및 평가

제안된 전문용어 처리방법을 약품도메인에 적용하였다. 실험 문서 수는 21,113개이고, 구문분석에 의하여 추출된 전체 명사수는 총 78,902개이다. 이 중 추출된 전문용어들의 수는 55,870개로 전체 명사수의 약 70.8%를 차지한다. 이들에 대한 분포가 아래의 표2에 나타나 있으며, 전문용어의 출현형태에 따른 분석을 행하고자 한다.

표 2 추출한 명사들의 분포

추출한 명사	수	점유율
일상 용어(단일명사)	23,032	29.19%
전문 용어(단일어절 형태)	24,896	31.55%
전문 용어(다중어절 형태)	30,974	39.26%
전체명사수	78,902	100.00%

다음 페이지에 있는 표 3은 단일어절의 형태로 출현하는 전문용어들의 결합형태에 따른 분포와 이들에 대한 정확도를 보여주는데, 제안된 알고리즘을 적용한 결과 출현한 전문용어들의 인식과 함께 2,864개의 하위개념이 추가된 것을 알 수 있다. 추출한 전문용어에 대한 평가는 두 명의 전문가에 의해서 수동으로 조사하였다. 결과는 추출된 전문용어 중에 올바른 관계로 연결된 전문용어의 비율을 나타내는 정확도로 평가되며 식은 다음과 같다.

$$\text{정확도} = \frac{\text{올바른 관계로 연결된 전문용어의 갯수}}{\text{추출된 전문용어의 갯수}}$$

예를 들어 명사 "진"의 경우, "촉진, 증진,..." 과 같은 개념들이 "습진, 농가진..." 개념들과 같은 군으로 형성된다. 이와 같이 잘못된 군이 많이 형성되는 경우에는 오류로 인식하여 정확도(71.87%)가 낮게 나온다.

그리고 표 4는 다중 어절의 형태로 출현하는 전문용어들의 패턴별 분포와 정확도를 나타낸다. 표에서 패턴의 수는 추가된 개념의 수를 나타낸다.

실험결과, 제안된 전문용어 처리방법에 의해 인식하지 못하는 오류형태는 다음과 같았다.

단일어절: **변형**, **합병**, 옷**감**, 전**환**, 고**통**, 증**진**, 촉**진**, 감**염**, 배**열**, 지**속**, 신**속**, ...

다중어절: 균**형** 유지, 각종 원인**에 의한**, 아래**의** 질환, 급 만성 질**환**, ...

표 3 단일어절형태 전문용어들의 분포와 정확도

결합형태	출현빈도	점유율	하위개념수	정확도
염	5,827	23.41%	506	98.50%
증(염증 제외)	4,306	17.30%	721	97.50%
통	3,220	12.93%	140	98.57%
균	2,238	8.99%	217	98.15%
성	2,156	8.66%	267	94.00%
질환	989	3.97%	175	93.14%
속	976	3.92%	115	92.17%
염증	748	3.00%	60	99.99%
진	705	2.83%	96	71.87%
감	648	2.60%	77	96.10%
종	596	2.39%	123	97.56%
병	574	2.31%	107	93.46%
열	562	2.26%	46	93.47%
궤양	454	1.82%	38	99.99%
선(백선 제외)	341	1.37%	50	78.00%
백선	191	0.77%	22	99.99%
증후군	163	0.65%	40	99.99%
형	114	0.46%	34	79.41%
환(질환 제외)	47	0.19%	18	77.78%
군(증후군 제외)	41	0.16%	12	91.67%
합계	24,896	100.00%	2,864	-
평균 정확도				92.57%

표 4 다중어절형태 전문용어들의 분포와 정확도

패턴유형	패턴수	출현빈도	점유율	정확도
패턴1	1,853	3,888	12.55%	90.69%
패턴2	975	1,327	4.28%	83.91%
패턴3	2,456	4,258	13.75%	81.79%
패턴4	1,361	2,379	7.68%	66.76%
패턴5	287	1,110	3.58%	76.67%
기타	-	18,012	58.15%	-
합계		30,974	100.00%	
평균정확도				66.64%

단일 어절 형태의 경우에는 결합하는 명사나 접미사가 동일한 단일명사와 전문용어를 구분하지 못하였다. 이는 명사사전의 불충분으로 인한 문제이므로 특정접미사나 명사로 종결되는 단일명사사전을 보강하고 이를 우선적으로 검색함으로써 간단히 해결될 수 있다 하지만 다중어절인 경우에는 "다음의... 아래에 의한..."과 같은 특정어휘에 대한 별도의 정교한 처리가 필요함을 알게 되었다.

5. 결 론

본 논문에서는 특정 도메인에 해당하는 문서들을 수집하여 코퍼스를 만들고, 코퍼스에 있는 텍스트의 분석결과를 이용하여 반자동으로 온톨로지를 구축하는 방법을 제안하였다. 이 때 웹으로부터 수집한 약품에 관련된 문서들을 실험 대상으로 삼았으며, 온톨로지의 구축에 필요한 개념과 관계들을 추출하기 위하여 결합하는 특정 명사나 접미사를 이용한 전문용어의 처리방안을 제시하였다. 구축된 온톨로지는 도메인에 의존적이었으며 접미사의 형태에 따른 의미군으로 분류되는 양상을 보여주었다. 접미사와 결합한 단일어절로 나타나는 전문용어를 인식한 결과 2,864개의 하위개념을 추가하고 평균 92.57%의 정확도를 보였으며, 다중어절로 나타나는 전문용어의 경우에는 평균 66.64%의 정확도를 보였다.

이와 같이 특정 도메인내의 텍스트를 분석하여 구축된 온톨로지는 자동으로 개념과 관계를 추가함으로써 좀 더 풍부한 정보를 가지게 되어 다양한 질의에 응답할 수 있다. 이는 온톨로지에 정의된 개념과 규칙들이 검색을 향상시키기 위한 추론의 기반으로 이용될 수 있다는 것을 의미한다. 앞으로 구축된 온톨로지를 일반 도메인에 적용하도록 확장하는 방안에 관해 계속 연구해 나가야 할 것이다.

참 고 문 헌

[1] Guarino, N.: Formal Ontology and Information Systems. In Proceeding of the 1st International Conference, Trento, Italy, IOS Press, 1998.

[2] Michele M., Paola V. and Paolo F., "Text Mining Techniques to Automatically Enrich a Domain Ontology," Applied Intelligence 18, 322-340, 2003.

[3] Kang, S. J. and Lee, J. H.: Semi-Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries, and Large Corpora. ACL 2001 Workshop on Human Language Technology and Knowledge Management, Toulouse, France, 2001.

[4] Lim, S. Y., Koo, S. O., Song, M. H., Lee, S, J., "Hub_word based on Ontology Construction for Document Retrieval," IC-AI'03, Las Vegas, USA, 2003.

[5] 이현민, 박혁로, "복합명사의 역방향 분해 알고리즘", 정보처리학회 논문지(B), 제8-B권 4호, pp. 357-364, 2003.

[6] 오종훈, 이경순, 최기선 "분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출", 정보과학회 논문지, 제29권 4호, pp. 258-269, 2002.

[7] 이경희, 이주호, 최명석, 김길창, "한국어 문서에서 개체명 인식에 관한 연구", 제12회 한글 및 한국어 정보처리 학술대회 학술발표논문집, pp. 292-299, 2000.

[8] 황이규, 윤보현, "HMM에 기반한 한국어 개체명 인식", 정보처리학회 논문지(B), 제-10권 2호, pp. 229-236, 2003.

[9] Vossen P., "Extending, trimming and fusing WordNet for technical documents," NAACL-2001 Workshop on WordNet and Other Lexical Resources:Applications, Extensions and Customizations, 2001.

[10] Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R.G.:WordNet: An On-line Lexical Database. International Journal of Lexicography, 1990.

[11] Michael K. Smith, Chris Welty, Deborah L. McGuinness, "OWL Web Ontology Language Guide," World Wide Web Consortium, http://www.w3.org/TR/owl-guide, 2003.

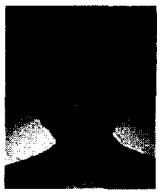
[12] Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishers, Boston, 2002.

[13] Baeza-Yates, R. and Robeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York, NY, USA, 1999.

[14] Bettina, B., Andreas, H., Gerd, S.: Towards Semantic Web Mining. International Semantic Web Conference, 2002.

임 수 연

1988년 경북대학교 전자공학과(공학사). 1991년 경북대학교 컴퓨터공학과(공학석사). 1996년~경북대학교 컴퓨터공학과 박사과정. 관심분야는 정보검색, 시멘틱 웹, 자연어처리



송 무 희

1984년 경북대학교 식품공학과(학사). 1998년 경북대학교 컴퓨터공학과(석사). 2002년~ 경북대학교 컴퓨터공학과 박사과정. 1995년~현재 경북대학교 정보전산원 근무. 관심분야는 정보검색, 문서분류, 지식베이스

이 상 조

1974년 경북대학교 수학교육과(이학사). 1976년 한국과학기술원(이학석사). 1994년 서울대학교 컴퓨터공학과(공학박사). 1976년~현재 경북대학교 컴퓨터공학과 교수. 관심분야는 자연어 처리, 정보검색, 기계학습, 운영체제, 프로그래밍언어, 시