접미사 패턴을 이용한 온톨러지의 구축방안

임수연[O] 구상옥 송무희 이상조
경북대학교 컴퓨터공학과
nadalsy@hotmail.com [O]

Ontology Construction methodology with Suffix pattern

Sooyeon Lim[0] Sangok Koo Muhee Song Sangjo Lee
Dept. Computer Engineering, Kyungpook National University, Korea

요 약

본 논문에서는 특정 도메인에서 사용되는 정보들과 그들의 관계를 정의해놓은 온톨로지를 반자동으로 구축하기 위하여 특정 도메인의 코퍼스에 있는 텍스트의 분석 결과를 이용하는 방안을 제시하고자 한다. 이를 위하여, 실험 도메인 내에서 빈번히 출현하는 전문용어들을 접미사와의 결합형태에 따라 추출하고 계층구조를 설정하는 알고리즘을 제안하고 약품매뉴얼을 대상으로 실험을 행하였다. 구축된 온톨로지는 자연스런 의미군을 형성하면서 풍부한 의미정보를 포함함으로써 정보검색 등의 전문적인 지식의 접근에 유용하게 쓰일 수 있으며, 검색의 성능을 향상시키기 위한 추론의 기반으로도 이용할 수 있다.

1. 서 론

온톨로지는 어떤 특정 도메인에서 사용되는 정보들과 그 정보들간의 관계를 정의해 놓은 것으로, 이에 관한 연구는 인공지능 분야의 시작과 더불어 지식 표현 분야의 핵심으로 활발히 연구되어져 왔다. 이들은 대부분 수작업으로 구축되어 왔으며 상당한 시간과 비용이 드는 수작업 대신 온톨로지를 반자동으로 구축하기 위한 방안이 연구되고 있다. 그 중의 하나는 기존의 시소러스나 사전 등과 같은 기존의 자원을 이용하는 경우로 개념이 부착된 대용량의 사전을 이미 확보함으로써 추가의 사전 작업 없이 바로 활용할 수 있는 지식베이스를 구축할 수 있다[2]. 다른 방법은 기존의 자원을 이용하지 않고 텍스트의 분석 결과로 얻어지는 단어들의 분포를 이용하여 온톨로지를 구축하고 확장하는 방법으로 개념의 확장이 용이하다[3]. 두 방법 모두 고품질의 의미관계패턴을 추출하는 것이 중요하다.

온톨로지의 학습은 온톨로지를 구축하고 갱신할 때의 시간과 비용을 줄일 수 있으며, 해당 도메인의 개념들과 그들 간의 의미관계를 추출하는 텍스트 마이닝 기술이 매우 중요하다[6].

본 논문에서는 보다 많은 의미관계패턴을 추출하기 위하여 특정 도메인에 있는 문서들에 출현하는 용어들의 형태를 분석하였다. 실험 도메인은 약학분야로 정하고, 약품의 매뉴얼(설명서)에 있는 텍스트들을 분석한 결과, 접미사와 결합한 전문 용어들이 많이 나타남에 이들에 대한 처리가 필요하였다. 따라서 결합한 접미사의 패턴이나 문맥을 고려하여 전문용어의 패턴들을 분류하고 이로부터 의미 군과 계층구조를 이끌어내어 온톨로지 내에서의 의미관계를 부여해주는 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 구축할 온톨로지의 구조와 표현, 온톨로지의 구축과정을 대략적으로 보여주며, 3장에서는 접미사 패턴을 이용한 전문용어의 처리방법을 보여준다 그리고 4장에서는 구축된 온톨로지와 그에 대한 실험결과를, 마지막 장에서 본 논문에 대한 결론을 맺는다.

2. 온톨러지의 구축과정

제안한 온톨로지 구축의 대략적인 과정은 그림1과 같이 네 단계의 과정으로 이루어진다. 첫 번째 단계에서는 관련도메인내의 웹 문서들을 구조화하여 코퍼스를 형성하고, 두 번째 단계에서는 개념들을 추출하기위한 간단한 자연어처리과정을 거친다. 세 번째 단계에서 전문용어를 추출하고 이로부터 계층구조를 구한 뒤, 마지막으로 온톨로지에 추출한 관계들을 추가한다. 각 단계에 대한 자세한 설명은 다음과 같다.

2.1 온톨러지의 구조와 표현

온톨로지는 관련 도메인 전문가들과의 협의에 의하여 개념들과 관계들의 구조를 정한뒤, 이들을 기반으로 구축된다. 즉, 실제의 응용시스템에서는 도메인마다의 특정적인 지식을 포함하는 온톨로지가 필요하다.

본 논문에서는 구축할 온톨로지의 구조를 정하기 위하여, 웹상의 약품과 관련된 신뢰성있는 데이터베이스의 구조를 분석하였다. 그 결과를 이용하여 구축할 약품 온톨

로지의 개념들과 이들을 연결시킬 관계들을 설정한다. 그림1는 설정된 개념들과 관계들로 이루어진 온톨로지의 구조를 개념 그래프로 나타낸 것이다.

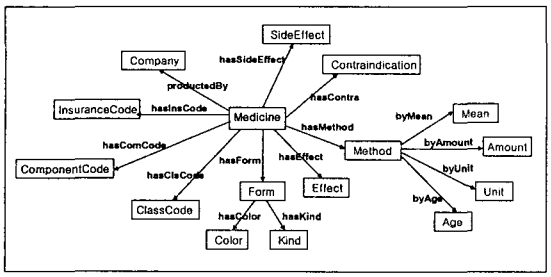

그림1. 설정된 개념들과 관계들로 이루어진 온톨로지의 구조

위의 그림에서 약품명은 제조회사, 보험코드, 성분코드, 효능효과 등 17개의 서브 카테고리로 나뉘어진다. 본 논문에서는 이 중 9개의 카테고리를 선정하고 그림1과 같은 개념들과 관계를 설정하였다. 그리고 가장 중요하다고 판단된 <Effect:효능효과> 태그에 있는 텍스트들을 집중적으로 분석하고 처리하고자 한다.

구축할 온톨로지에 존재하는 개념들과 그들의 관계는 OWL[5]을 이용하여 표현하고 한글과 영어를 혼용한다. 다음의 그림2은 개념 " Medicine:약품명" 을 OWL로 표현한 간단한 예를 보여준다.

```
<?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>
```

그림2. OWL로 표현된 온톨로지의 일부분

2.2 개념의 추출

해당 도메인과 관련이 있어 웹으로부터 수집한 문서들은 학습(learning)을 위한 코퍼스를 형성하기 위해 그림1의 구조에 맞도록 변환을 거친 뒤, 자연어 처리 기술을 이용하여 분석한 문서 내에서 특정 내용을 포함하는 구문패턴이 존재하는 경우에는 그 패턴에 따라 텍스트들을 분류하고 태깅해준다. 태깅된 텍스트들은 간단한 자연어 처리 파서를 거쳐 문서내의 모든 명사들을 추출한다.

약품 매뉴얼로부터 추출한 용어들은 주로 병명, 증세, 성분등을 나타내는 전문용어들로 출현하고 있었다. 그 원인은 전문용어를 많이 포함하는 약품 도메인 내의 문서들이란 특성 때문으로 추측된다. 어휘 특정성(specificity)이 큰 전문용어에 대한 처리는 풍부한 의미정보를 지니는 온톨로지의 구축이 가능하게 한다.

3. 전문용어의 처리-관계의 추가

본 논문에서는 기존 사전에 없는 단어들인 전문용어들을 자동으로 추출하기 위하여 그들의 출현형태를 분석하였다. 해당 도메인에 출현하는 전문용어들은 단일어절과 다중어절 두가지의 형태로 나타났다.

3.1 단일어절의 형태

약품도메인 내에서 복합명사를 이루고 있는 접미사들은 대략 20가지로 나누었다. 이 접미사들은 의미적으로 관련이 있는 전문용어들을 서로 연결시키며," 염, 증, 통, 균, 성, 질환, 속, 염증, 진, 감, 종, 병, 열, 궤양, 선, 백선, 증후군, 형, 환, 군" 이다. 약품도메인에서 단일어절의 형태로 출현하는 전문용어들은 " 방광염, 기관지염" 과 같이 접미사(감염증을 나타내는 " 염")의 하위 단어인 경우가 대부분이다. 따라서 이들은 "hyponymOf" 관계로 연결한다. 다음은 단일어절의 형태로 나타나는 전문용어 내에서 하위관계를 자동으로 추출하기 위한 알고리즘이다.

알고리즘. 단일어절 전문용어의 계층관계 추출

입력 : 접미사와 결합한 단일어절 전문 용어들
출력 : 전문 용어들간의 계층 트리

```
String Suffix[]={염,증,통,균,성,질환,속,염증, 진, 감, 종,병,열,궤양,
                 선,백선,증후군,형,환,균}
boolean matrix[][]=false;
int size=n;    //n=단어의 개수

// 계층관계 matrix
for (int i; i<size; i++) {
  for (int j; j<size; j++) {
     if ((i<>j) and (j번째 단어가 i번째 단어로 끝남))
          matrix[i][j]=true;
  }
}

// 온톨로지 노드에 서브트리를 추가
for (int i; i<size; i++) {
  for (int j; j<size; j++) {
     i번째 단어의 오른쪽 하위 단어를 찾아서
     i노드의 하위에 추가.
  }
}
```

3.2 다중어절의 형태

텍스트에 나타난 전문 용어들은 대부분 " 만성위염" 과 같이 수식어와 중심어의 관계를 가지며 중심어가 다시 단일 어절로 이루어진 전문 용어인 경우가 많이 출현하였다. 본 논문에서는 이 문맥 관계들을 다섯 개의 관계 패턴들로 설정하고 이에 따라 온톨로지 내에서의 의미관계를

추가하고 설정한다. 표1은 설정한 패턴들과 그에 따른 관계 설정방안을 보여준다.

표1. 문맥의 패턴에 따른 관계설정방안

패턴1	*N1*(~성, ~형)+ *N2*
	*N1N2*가 *N2*로부터 확장된 전문용어로 보고 *N2*의 하위개념으로 연결 (예) 급성 기관지염, 만성 파로
패턴2	*N1*(~에 의한, ~(으)로 인한, ~(으)로 인해 유발된)+ *N2* *N1*(~에 따른)+ *N2* *N1*(~시(의), ~상태에서, ~후(의))+ *N2*
	*N2*는 관계 causeTo에 의해 *N1*과 연결 *N2*는 관계 accompanyWith에 의해 *N1*과 연결 *N2*는 관계 stateOf에 의해 *N1*과 연결 (예) 농무에 의한 운전, 수술시 국소마취
패턴3	*N1*+ "의"+ *N2*, *N1*+ *N2*
	*N1N2*가 전문용어로 추출 예) 근이완의 유지 --> 근이완유지
패턴4	*N1*+ "및"+ *N2*, *N1*+ "."+ *N2*, *N1*+ "또는"+ *N2*
	*N1*과 *N2* 각각을 전문용어로 추출 (예) 소화효소결핍 및 담즙분비촉진 --> 소화효소분비, 담즙분비촉진
패턴5	*N1* (*suffix_ 1*)+ "· ,및"+ *N2*(*suffix_ 2*)+ *N3* (이 때, *suffix_ 1=suffix_ 2*)
	*N1N3*와 *N2N3* 가 전문용어로 추출 (예) 지연형·활동성 만성간염 --> 지연형 만성간염, 활동성 만성간염

설정한 복합명사의 패턴에 따라 추출된 관계들은 주변에 나타난 명사들을 연결짓는 의미관계를 나타낸다. 다음의 그림3은 예제 텍스트로부터 추출한 개념과 관계를 추가한 온톨로지의 일부분을 보여준다.

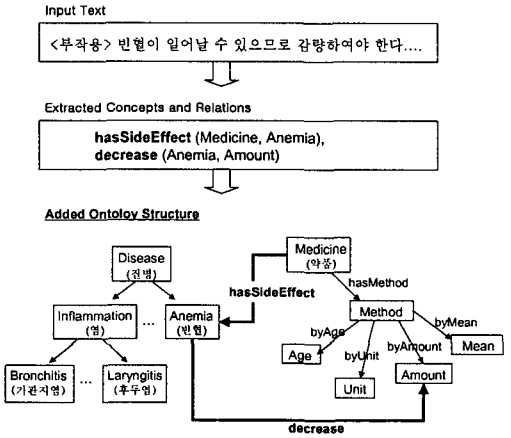

그림3. 추출한 개념과 관계를 추가한 온톨로지의 일부분

4. 실험 및 평가

제안된 전문용어 처리방법을 약품 도메인에 적용하였다. 실험 문서 수는 21,113개이고, 구문분석에 의하여 추출된 전체 명사수는 총 76,782개이다. 이 중 전문용어의 수는 55,870개이고, 특히 접미사가 부착된 단일어절 형태의 전문용어수가 24,896개이다. 표2는 단일어절의 형태로 출현하는 전문용어들의 접미사형에 따른 분포와 이들에 대한 정확율을 나타내고 표3은 다중일어절의 형태로 출현하는 전문용어들의 패턴별 분포와 정확율을 나타낸다.

표2. 단일어절형태 전문용어들의 분포와 정확율

접미사	출현빈도	점유율	하위개념수	정확율
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%
전환	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%
충후군	163	0.65	40	99.99%
혈	114	0.46	34	79.41%
환(전환 제외)	47	0.19	18	77.78%
군(충후군 제외)	41	0.16	12	91.67%
합계	24,896	100.00	2864	
평균 정확율				92.57%

표3. 다중어절형태 전문용어들의 분포와 정확율

	패턴1	패턴2	패턴3	패턴4	패턴5
패턴수	1,853	975	2,456	1,361	287
빈도	3,888	1,327	4,258	2,379	1,110
정확율	90.69%	83.91%	81.79%	66.76%	76.67%

5. 결론

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

Reference

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

[2] 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.

[3] 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.

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

[5] 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.

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