Journal of Korea Multimedia Society Vol. 13, No. 9, September 2010(pp. 1263-1271)

서술어 온톨로지를 이용한 자연어 문장으로부터의 온톨로지 자동 생성

민영근[†], 이복주[††]

요 약

시맨틱 웹 구현의 중요한 수단인 온톨로지는 검색, 추론, 지식표현 등 다양한 분야에서 사용되고 있다. 그러나 잘 구성된 온톨로지를 개발하는 것은 시간적, 물질적으로 많은 자원이 소모된다. 이러한 문제를 극복하기 위해 온톨로지를 자동으로 구축하는 시도가 있었다. 본 연구에서는 자연어 문장으로부터 직접 온톨로지를 자동적으로 생성하기 위해 형태소와 문장의 구조를 분석하고 자연어 문장의 서술어를 찾아 해당 온톨로지 서술어로 변환되게 하기 위하여 '서술어 온톨로지(predicate ontology)'를 두어서 분석된 자연어 문장의 서술어가 적절한 온톨로지 서술어로 변환될 수 있도록 한다. 인간 온톨로지 구축가와 제안한 방법을 비교한 실험 결과 정확도에서 나은 결과를 보였다.

Automatic Ontology Generation from Natural Language Sentences Using Predicate Ontology

Youngkun Min[†], Bogju Lee[††]

ABSTRACT

Ontologies, the important implementation tools for semantic web, are widely used in various areas such as search, reasoning, and knowledge representation. Developing well-defined ontologies, however, requires a lot of resources in terms of time and materials. There have been efforts to construct ontologies automatically to overcome these problems. In this paper, ontologies are automatically constructed from the natural languages sentences directly. To do this, the analysis of morphemes and a sentence structure is performed at first. then, the program finds predicates inside the sentence and the predicates are transformed to the corresponding ontology predicates. For matching the corresponding ontology predicate from a predicate in the sentence, we develop the "predicate ontology". An experimental comparison between human ontology engineer and the program shows that the proposed system outperforms the human engineer in an accuracy.

Key words: ontology(온톨로지), automatic ontology generation(온톨로지 자동 생성), predicate ontology (서술어 온톨로지)

1. 서 론

시맨틱 웹[1]은 검색, 추론, 지식표현분야에서 광범위하게 사용되고 있다. 시맨틱 웹을 만들기 위한 중요한 요소인 온톨로지는 어떤 특정 관심영역에서 사용되는 개념들과 개념들 간의 관계를 정의해 놓은

※ 교신저자(Corresponding Author): 민영근, 주소: 경기도 용인시 수지구 죽전동 126번지 단국대학교 제2공학관 520호(448-701), 전화: 031)8005-3666, FAX: 031)8005-3666, E-mail: minyk@dankook.ac.kr
접수일 : 2009년 6월 23일, 수정일 : 2009년 11월 27일
완료일 : 2010년 6월 7일

[†] 준회원, 단국대학교 대학원 컴퓨터학과 박사과정
[††] 정회원, 단국대학교 컴퓨터학과 교수
(E-mail: blee@dankook.ac.kr)
※ 본 연구는 2008-2, 2009-1 년도 단국대학교 대학원 연구보조장학금의 지원으로 이루어진 것임.

것으로서, 이러한 온톨로지는 구축하려는 관심영역의 지식을 보유한 전문가와 온톨로지 구축가가 협업하여 개발하고 있다. 온톨로지를 구축하는 일반적인 과정은 다음 그림 1과 같다. 먼저, 구축할 온톨로지의 목적을 정하고 범위와 명세를 규정한 다음 정보를 획득하고 형식화한다. 정보의 형식화 단계에서 온톨로지를 구체적으로 구축하게 되며, 마지막으로 구축된 온톨로지에 대한 평가를 수행한다.

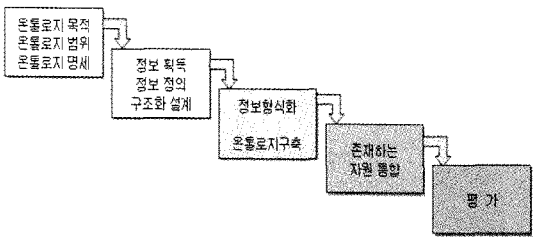

그림 1. 온톨로지 구축 방법

이러한 온톨로지 개발을 위해서는 다수의 관심영역 전문가와 다수의 온톨로지 구축가가 많은 시간과 비용을 들여 공동으로 개발하는 것이 일반적이며, 경우에 따라서 관심영역 전문가를 찾기 어려운 분야도 존재한다. 온톨로지를 자동으로 구축하고자 하는 시도는 기존에도 많이 연구되었다. 기존의 연구는 기구축된 자료를 사용하는 방법, 데이터마이닝 알고리즘을 사용하는 방법, 자연어 처리를 사용하는 방법으로 구분할 수 있다. 이러한 방법들 중 자연어 처리를 이용하는 방법이 가장 이상적으로 구축할 수 있는 방법이지만, 자연어 처리의 까다로움으로 인하여 개척되지 못한 것이 현실이다. 본 논문에서는 관심영역 전문가와 온톨로지 구축가의 많은 노력이 필요했던 온톨로지 구축을 자연어로부터 직접 자동으로 구축함으로써 온톨로지 구축을 매우 용이하게 할 수 있는 방법을 제안한다. 온톨로지 자동 구축의 다른 장점은 사람이 수동으로 구축함으로써 나타나는 지식의 불일치성, 편견을 줄이고 보편적인 지식에 기반을 둔 온톨로지를 구축할 수 있다는 것이다. 이를 달성하기 위하여 문장 서술어와 온톨로지 개념간의 매칭을 위하여 서술어 온톨로지(predicate ontology)를 제안하며, 자연어 문장을 온톨로지로 변환하기 위한 특별한 규칙들을 제시한다.

본 논문의 구성은 2절에서 기존의 온톨로지 구축 연구들에 대하여 기술하고 3절은 제안한 자연어 처리를 통한 온톨로지 자동 생성에 대해 기술한다. 4절은 실험 및 성능 평가, 5절은 향후 연구를 포함한 결론을 기술하였다.

2. 기존 연구

온톨로지 자동 구축에 관한 기존 연구에는 기존의 사전 또는 동의어 사전 등과 같이 기 구축되어 있는 자료를 확보함으로써 추가의 작업 없이 바로 활용할 수 있는 지식으로부터 구축하는 방법[3], 기존의 자료를 활용하지 않고 문서의 분석 결과로 얻어지는 단어들의 분포를 이용하여 온톨로지를 구축하고 확장하는 방법[4], 문서에 데이터마이닝 기법을 도입하여 문서로부터 개념을 추출하여 온톨로지를 구축하는 방법[5-7]이 있다. 접미사 패턴을 이용한 방법[8]은 관심영역의 전문용어들에서 빈번히 등장하는 접미사 20개의 상, 하위 개념을 미리 지정하여 개념의 접미사를 조사하여 상, 하위관계를 추출하고 미리 정의되어 있는 5개의 문맥 패턴에 의하여 관계를 생성하거나 새로운 전문용어를 추출하여 온톨로지를 생성한다. 이러한 방법은 미리 조사된 영역에서는 효과적일 수 있으나 다른 영역에 적용하기 위해서는 적용할 영역에서 사용되는 전문용어의 접미사 패턴과 문맥 패턴을 사전에 조사하여야 하는 어려움이 있다. 또한 대상 문을 분석하여 단어 종류, 품사, 분석된 구조 등을 사용하여 온톨로지를 구축하는 방법[9]이 있다. 이 방법은 자연어처리 결과를 단어, 단어-품사, 구문이 기술되어 있는 XML문서로 생성하고 미리 정의된 XSL 변환을 통하여 온톨로지를 생성한다. 이 방법 역시 사전에 정의된 XSL 변환 규칙이 필요하며 본 논문에서 대상으로 하는 한국어의 처리에 적용하기에는 너무 많은 변환 규칙이 XSL 문법으로 정의되어야 할 것이다. [10]에서는 표준화되어 있는 XML 문서로부터 데이터마이닝 기법의 연관 규칙 알고리즘을 적용하여 반자동으로 온톨로지를 구축하고 있다. [11]에서는 문서로부터 온톨로지를 구성하는 개념간의 관계를 정의하는 자동화된 방법을 제안하였다. 이러한 기존에 여러 연구들에서 구축되어 있는 온톨로지 자동, 반자동 구축 도구 등은 자연어 문장에서 서술어를 그대로 가져와서 온톨로지의 서술어로 사용한다. 영어의 경우 "WordNet"[12]이 잘 구축되어 있어서 이를 참조하여 서술어를 결정한다. 이런

경우, 구축된 온톨로지를 사용할 때 구축된 의도와 사용하는 의도가 다른 불일치문제가 야기될 수 있다. 또한 같은 자동, 반자동 구축 도구를 사용해도 자연어 문장에 따라 온톨로지의 서술어가 달라진다면 구축하는 의미가 없다고 할 수 있다. 표 1은 기존의 제안되었던 방법들을 원본, 변환기법, 대상 언어별로 정리한 것이다.

한편 온톨로지 자동구축에 필수적인 한글의 자연어 처리에 대한 연구는 국내 여러 대학과 연구기관에 의하여 진행되고 있으며 국책사업인 세종계획[13]에서도 진행 중이다. 한글의 자연어 처리는 형태소 분석기가 상당한 비중을 차지하며 여러 방법들이 제안되었다. 형태소 분석기는 문장을 어절로 분리한 후 단어의 품사와 품사에 따른 변형을 분석하여 그 결과를 반환한다. 이러한 결과를 사용하여 문장의 구조와 문장 내의 서술 구조를 파악 할 수 있다. 한편 영어권에서 자연어 처리 분야에서 동사 등을 개념에 대한 구조로 구축한 BerbNet[14], PropBank[15], NomBank [16], FrameNet[17] 같은 일들이 추진된 바 있다.

지금까지의 온톨로지 구축 작업은 특정 온톨로지를 구축해서 개별 시스템에 사용하고 있다면 추후에는 온톨로지가 재 사용성이 높아지도록 해야 한다. 본 연구에서는 '서술어 온톨로지'를 통하여 문장 서술어를 정형화된(well-formatted) 온톨로지로 변환한다. 구축된 온톨로지와 서술어 온톨로지가 공개된다면 어떠한 시스템에도 범용 적으로 활용될 수 있다.

3. 서술어 온톨로지를 이용한 온톨로지 자동 구축

본 연구에서는 자연어 문장으로부터 온톨로지로 구축하기 위하여 그림 2의 온톨로지 자동구축 구조를 제안한다. 제안된 구조는 형태소 분석기, 문장 구조 분석기, 온톨로지 구축기의 세 가지 모듈과 서술어 온톨로지로 구성되어 있다. 형태소 분석기는 입력된 문장을 분석하여 문장을 형태소 단위로 나누어주며 형태소의 품사를 결정하고 이렇게 결정된 형태소와 품사는 문장 구조 분석기의 입력으로 사용된다.

표 2는 온톨로지 자동 구축의 전체적인 알고리즘을 나타내고 있다. 먼저 사용자로부터 분석해야 할 텍스트를 입력 받는다. 텍스트 내의 각 문장에 대해 형태소 분석을 수행하고 유한 상태 기계(FSM: finite state machine)를 이용하여 문장을 분석하며 서술어 온톨로지를 이용하여 온톨로지를 생성한다. 그리고 이를 온톨로지 저장소에 저장한다. 이를 모든 문장에 대해 반복한다. 각각의 단계는 본 절에서 상세히 기술된다.

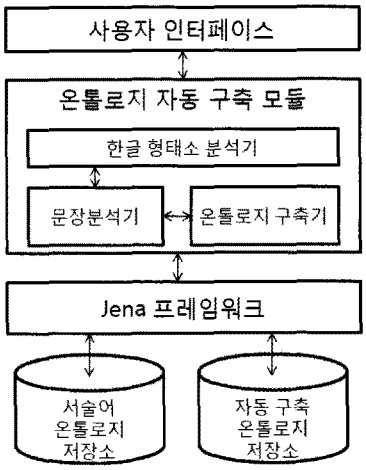

그림 2. 온톨로지 자동 구축의 시스템 구조

표 1. 기 제안된 방법들의 비교

제안된 방법	원 본	변환기법	대상언어
S.J. Kang, J.H. Lee[3]	동의어/유의어 사전		한국어
S.Y. Lim 등[4]	자연어 문서	단어의 분포	한국어
P. Clerkin 등[5] A. Wrobel 등[6] M. Cannataro 등[7]	자연어 문서	데이터 마이닝	영 어
임수연 등[8]	자연어 문서	접미사 패턴	한국어
J. Saias, P. Quaresma[9]	자연어 문서	자연어 처리, XSLT	영 어
구미숙 등[10]	XML 문서	데이터 마이닝	한국어
김희수 등[11]	자연어 문서	통계	한국어

표 2. 온톨로지 자동 구축 과정

1. 사용자 입력으로부터 분석해야 할 텍스트 받음
2. 텍스트를 형태소 분석
3. 형태소 분석된 텍스트의 각 문장에 대해
3.1 FSM을 이용한 문장 분석
3.2 서술어 온톨로지를 이용한 온톨로지 생성
3.3 온톨로지 저장소에 저장

표 3은 문장의 예시와 분석된 형태소를 나타내고 있다. "온톨로지 자동 생성기 구조는 형태소 분석기, 문장 구조 분석기, 온톨로지 생성기로 구성된다."라는 예시 문이 '온톨로지 자동 생성기 구조는', '형태소 분석기,', '문장 구조 분석기,', '온톨로지 생성기로', '구성된다.'의 네 부분으로 나뉘고, 각각의 단어에 품사가 지정된 모습이다.

표 3. 형태소 분석 예

문장	온톨로지 자동 생성기 구조는 형태소 분석기, 문장 구조 분석기, 온톨로지 생성기로 구성된다.	
	어절	결정된 품사
분석	온톨로지 자동 생성기 구조는	온톨로지 / NN 자동 / NN 생성기 / NN 구조 / NN 는 / JO, SB
	형태소 분석기,	형태소 / NN 분석기 / NN ,/SY
	문장 구조 분석기,	문장 / NN 구조 / NN 분석기 / NN ,/SY
	온톨로지 생성기로	온톨로지 / NN 생성기 / NN 로 / JO, AD
	구성된다.	구성되 / VV ㄴ다 / EM, NM ./SY

3.1 서술어 온톨로지

서술어 온톨로지는 본 연구에서 제안된 서술어 간의 관계를 정의한 온톨로지로서 문장 서술어와 온톨로지 서술어 간의 매칭을 가능하게 한다. 문장 서술어의 원형과 품사로 온톨로지 서술어에 해당하는 서술어를 검색한다. 서술어의 형태는 "rdfs:subClassOf", "owl:oneOf" 등의 이미 표준으로 정의되어 있는 형태이거나, "color", "partOf"와 같이 본 서술어 온톨로지에서 특별히 정의되어 있는 형태이다. 형태소 분석을 통하여 나온 형태소들 중 온톨로지의 3항-구조(triple) 요소 중 하나인 서술어에 적합한 형태소를 판단하여 적절한 서술어에 맞춰 주기 위하여 사용한다. 그림 3는 본 연구에서 개발된 서술어 온톨로지의 개념 구조를 보여주고 있다. 최상위 노드에 모든 서술어의 최상위인 'Predicate'가 위치하고 그 아래에 'Adjective', 'Adverb', 'Verb'가 위치한다. 형용사에 해당하는 'Adjective'는 상태를 의미하는 'AdjState'를 가지며 'AdjState'는 다시 'Color', 'Shape'으로 나누어져 있다. 동사에 해당하는 'Verb'는 주어의 상태를 나타내는 'VerbState'와 어떠한 동작을 의미하는 'Action'으로 구성되어 있다.

이러한 서술어 온톨로지의 구체적 개체는 'primitive', 'predicate', 그리고 'inversepredicate' 특성을 가지며 'primitive' 특성은 한글의 원형을 나타내고, 'predicate' 특성은 사용될 온톨로지 서술어를 나타내며, 'inversepredicate' 특성은 역 관계를 서술한다. 온톨로지를 생성할 때 정 관계만이 아니라 역관계도 같이 생성되어야 생성된 온톨로지의 활용도를 더욱 높일 수 있으며 'inversepredicate' 특성은 이 경우를 대비하기 위해서이다. 표 4는 서술어 온톨로지 개체의 예를 보여주고 있다. 첫 번째 개체인 'adj_002'의 개념은 'AdjState'이므로 그림 2에서 'adjective'의 하위 클래스로서 형용사임을 알 수 있고, 원형을 나타내는 'primitive' 특성의 값은 '이'이며 사용될 서술어는 'rdfs:subClassOf'와 'state' 두 가지가 있으

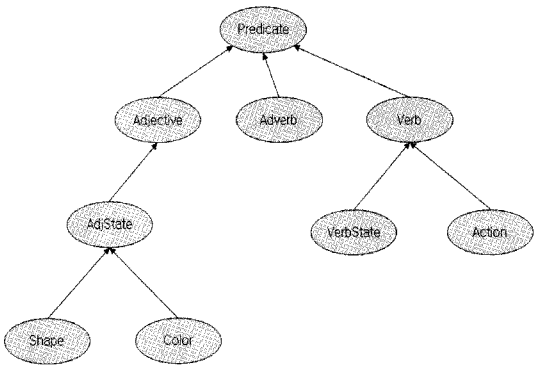

그림 3. 서술어 온톨로지의 개념 구조도

표 4. 서술어 온톨로지 개체의 예

```
<AdjState rdf:ID="adj_002">
  <primitive rdf:datatype="http://www.w3.org/2001/XMLSchema#string">이</primitive>
  <predicate rdf:datatype="http://www.w3.org/2001/XMLSchema#string">state</predicate>
  <predicate rdf:datatype="http://www.w3.org/2001/XMLSchema#string">rdfs:subClassOf</predicate>
  <inversepredicate rdf:datatype="http://www.w3.org/2001/XMLSchema#string">superClassOf</predicate>
  <inversepredicate rdf:datatype="http://www.w3.org/2001/XMLSchema#string">stateOf</predicate>
</AdjState>
<VerbState rdf:ID="verb_0011">
  <predicate rdf:datatype="http://www.w3.org/2001/XMLSchema#string">assembledWith</predicate>
  <inversepredicate rdf:datatype="http://www.w3.org/2001/XMLSchema#string">partOf</inversepredicate>
  <class rdf:datatype="http://www.w3.org/2001/XMLSchema#string">vv</class>
  <primitive rdf:datatype="http://www.w3.org/2001/XMLSchema#string">구성되</primitive>
</VerbState>
```

며 역관계 서술어로 'superClassOf'와 'stateOf' 두 가지가 지정되어 있다. 두 번째 개체인 'verb_0011'은 원형이 '구성되'이며 동사이고 서술어로 'assembledWith'가, 역관계 서술어로 'partOf'가 지정되어 있다.

본 연구에서 역관계 서술어로서 OWL의 기본 특성 중 하나인 "inverseOf"를 사용하지 않고 새로운 특성인 "inversepredicate"를 사용한 이유로는 "inverseOf" 특성은 실제 데이터 값을 가지는 특성이 아니라 다른 클래스를 가리키는 특성으로서 "inverseOf"를 사용하게 되면 "rdfs:subClassOf"의 역 관계 특성을 서술어 온톨로지가 아닌 W3C에서 제공하는 "rdfs" 온톨로지의 표준으로 추가해야 하는 부담이 있다. 서술어 온톨로지에서는 역 관계 특성을 자체 정의함으로서 이러한 부담을 줄이고자 하였으며, 제안된 시스템을 사용하여 구축된 온톨로지에는 역 관계가 표준 특성인 "inverseOf" 특성으로 구축된다.

3.2 문장 구조 분석기 및 온톨로지 구축기

문장 구조 분석기는 자연어 문장의 형태와 단어의 품사를 토대로 문장의 구조를 분석하여 문장에 사용된 단어들로부터 주어, 서술어 목적어 쌍으로 구성하여 온톨로지 구축기에 전달한다. 이를 위하여 문장 구조 분석기는 유한 상태 기계(finite state machine)로 설계하여 현재 상태와 다음 단어의 품사에 따라 문장의 구조를 분석한다.

온톨로지 구축기는 문장 구조 분석기의 분석 결과로부터 온톨로지 구문들을 생성하는데, 이때 문장의 서술어에 적합한 온톨로지 서술어를 결정하기 위하여 3.1절에서 서술한 서술어 온톨로지를 참조한다. 온톨로지 구축기는 문장 구조 분석기의 결과 중 서술어에 저장되어 있는 형태소를 사용하여 서술어 온톨로지를 검색하여 생성될 온톨로지에 적합한 서술어를 결정한다. 예시문인 "온톨로지 자동 생성기는...로 구성된다."라는 문장에서 문장 구조 분석기는 '구성되'가 서술어임을 결정하였다. 온톨로지 구축기는 다음의 순서를 통하여 온톨로지에 사용될 서술어를 결정한다. 첫째, 서술어 온톨로지에서 predicate가 'primitive'이고, object가 '구성되'인 subject가 존재하는지 검색한다. 둘째, 존재한다면 해당하는 subject와 predicate에 'predicate'를 사용하여 해당 개체가 저장하고 있는 서술어를 찾아서 배열 형태로 저장하여 반환한다. 셋째, 존재하지 않는다면 '구성되'에 해당하는 개체가 없는 것이다. 이 경우에는 '구성되'의 품사인 'Verb'를 토대로 서술어를 결정하기 위하여 서술어 온톨로지에서 subject가 'Verb', predicate가 'default'인 개체를 검색하여 기본 서술어를 반환한다. 이러한 순서를 통하여 반환된 서술어를 사용하여 온톨로지를 생성한다. 이 때 생성될 온톨로지의 주어와 목적어인 명사들을 기존의 생성되어 있는 온톨로지에서 검색하여 동일한 주어와 목적어가 있다면 이미 생성되어 있는 온톨로지 개념을 사용하여 생성하며, 존재하지 않는다면 먼저 주어와 목적어에 해당하는 명사로부터 새로운 온톨로지 개념을 생성하고 생성된 온톨로지 개념을 사용하여 완전한 온톨로지 문장을 생성한다. 이러한 절차를 거침으로써 생성되는

표 5. 온톨로지 구축기 알고리즘

1. 형태소 분석 결과 문장 서술어 P 얻음
2. 서술어 온톨로지에서 predicate = 'primitive' 이고 object = P인 subject S 얻음
3. subject S가 존재하면
   3.1 subject = S 이고 predicate = 'predicate' 인 모든 object들을 서술어 배열로 반환
4. subject가 존재하지 않으면
   4.1 P의 품사 Pr 얻음
   4.2 서술어 온톨로지에서 subject = Pr 이고 predicate = 'default' 인 개체 얻음
   4.3 얻어온 개체의 기본 서술어 반환
5. 온톨로지에 추가

온톨로지가 계속하여 누적될 수 있는 효과가 있으며, 주어와 목적어를 단순한 문자열 형태가 아니라 하나의 온톨로지 개념으로 생성하여 복잡한 문장에 적응할 수 있도록 한다. 이는 알고리즘은 표 5에 기술되어 있다.

다음 그림 4은 전체 알고리즘의 흐름도이다.

먼저 자연어 문장으로부터 형태소 분석을 통하여 각 형태소의 품사와 기본형을 결정하고 구문분석을 통하여 주어, 서술어, 목적어 관계를 결정한다. 그 후 서술어의 품사와 기본형으로부터 서술어 온톨로지를 추론하여 구축할 온톨로지 구문에 사용할 서술어를 결정하고 결정된 서술어를 사용하여 온톨로지를 구축한다.

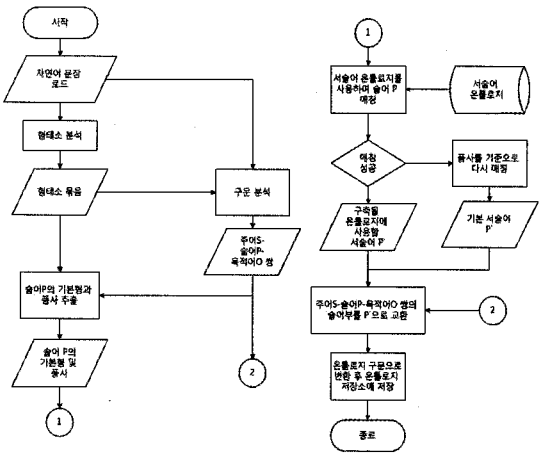

그림 4. 전체 알고리즘 흐름도

4. 실험 및 성능 평가

본 연구에 사용된 자연어 처리 모듈 중 형태소 분석기는 공개된 모듈[17] 을 수정하여 본 논문에서 제안한 XML양식으로 출력할 수 있도록 수정하여 사용하였다. 표 6은 이 형태소 분석기를 사용한 결과 XML 파일을 보이고 있다. <sentence> 태그는 하나의 문장 당 하나씩 생성되며 하위 요소로 문장에서 나누어진 어절을 나타내는 노드를 가지고 있다. <eojel> 태그는 문장에서 나누어진 어절에 대하여 기술하고 있으며 하위 요소로 형태소 분석의 가장 작은 단위인 형태소 노드를 가진다. <morphe> 요소는 형태소의 원형을 'data' 특성으로, 형태소 분석을 통하여 결정된 품사를 'class', 'func', 'type' 특성으로 나타내고 있다. 'class' 특성은 형태소의 품사를 나타내며 'func' 특성은 조사의 격, 어말어미 등 품사의 기능에 대한 부분을 추가적으로 나타낸다. 'type' 특성은 조사의 구분, 어말어미 구분 등 품사의 구분에 대하여 추가적인 정보를 제공한다.

표 6과 같은 형태소 분석결과를 문장구조 분석기와 온톨로지 구축기를 거치면 표 7과 같은 온톨로지가 생성되게 된다. 예시 문장 "온톨로지 자동 생성기는..."의 명사인 '온톨로지 자동 생성기', '형태소 분석기', '문장 구조 분석기'와 '온톨로지 구축기'는 온톨로지 개념으로서 생성되었으며 자연어 서술어인 '구성되'가 'po:assembledWith'로 변환되어 '온톨로지 자동 생성기'와 형태소 분석기', '문장 구조 분석기'와 '온톨로지 구축기' 사이의 관계를 설명하고 있다. 본 연구에서 제안한 서술어 온톨로지의 URI prefix 이름은 'po:'로 지정하였으며 자동적으로 생성된 온톨로지의 접두사(URI prefix)는 'ABo:'로 지정하였다.

그림 5는 생성된 온톨로지를 시맨틱 네트워크 형태로 보인 것이다.

본 연구에서 제안한 온톨로지 자동 생성 방법의 성능 측정을 위해 비교적 큰 크기의 예문을 입력으로 하여 실험하였다. 인간 온톨로지 구축가와 제안된

표 6. 형태소 분석 결과의 XML 파일

```
<result>
<sentence ID="0">
 <eojeol ID="0" data="온톨로지 자동 생성기는">
  <morphe ID="0" class="NN" data="온톨로지" func="" type=""/>
  <morphe ID="1" class="NN" data="자동" func="" type=""/>
  <morphe ID="2" class="NN" data="생성기" func="" type=""/>
  <morphe ID="3" class="JO" data="는" func="SB" type="CL"/>
 </eojeol>
 <eojeol ID="1" data="형태소 분석기">
  <morphe ID="0" class="NN" data="형태소" func="" type=""/>
  <morphe ID="1" class="NN" data="분석기" func="" type=""/>
 </eojeol>
 <eojeol ID="2" data=",">
  <morphe ID="0" class="SY" data="," func="" type=""/>
 </eojeol>
 <eojeol ID="3" data="문장 구조 분석기">
  <morphe ID="0" class="NN" data="문장" func="" type=""/>
  <morphe ID="1" class="NN" data="구조" func="" type=""/>
  <morphe ID="2" class="NN" data="분석기" func="" type=""/>
 </eojeol>
 <eojeol ID="4" data=",">
  <morphe ID="0" class="SY" data="," func="" type=""/>
 </eojeol>
 <eojeol ID="5" data="온톨로지 생성기로">
  <morphe ID="0" class="NR" data="온톨로지" func="" type=""/>
  <morphe ID="1" class="NN" data="생성기" func="" type=""/>
  <morphe ID="2" class="JO" data="로" func="AD" type="CL"/>
 </eojeol>
 <eojeol ID="6" data="구성된다">
  <morphe ID="0" class="VV" data="구성되" func="" type=""/>
  <morphe ID="1" class="EM" data="ㄴ다" func="NM" type="ED"/>
 </eojeol>
 <eojeol ID="7" data=".">
  <morphe ID="0" class="SY" data="." func="" type=""/>
 </eojeol>
</sentence>
</result>
```

표 7 온톨로지 변환 결과 예

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:ABo="http://ai.dankook.ac.kr/ABontology.owl#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:po="http://ai.dankook.ac.kr/PredicateOntology.owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<owl:Class rdf:about="http://ai.dankook.ac.kr/ABontology.owl#온톨로지 생성기">
  <po:partOf rdf:resource="http://ai.dankook.ac.kr/ABontology.owl#온톨로지 자동 생성기"/>
</owl:Class>
<owl:Class rdf:about="http://ai.dankook.ac.kr/ABontology.owl#형태소 분석기">
  <po:partOf rdf:resource="http://ai.dankook.ac.kr/ABontology.owl#온톨로지 자동 생성기"/>
</owl:Class>
<owl:Class rdf:about="http://ai.dankook.ac.kr/ABontology.owl#문장 구조 분석기">
  <po:partOf rdf:resource="http://ai.dankook.ac.kr/ABontology.owl#온톨로지 자동 생성기"/>
</owl:Class>
  <owl:Class rdf:about="http://ai.dankook.ac.kr/ABontology.owl#온톨로지 자동 생성기">
    <po:assembledWith rdf:resource="http://ai.dankook.ac.kr/ABontology.owl#온톨로지 생성기"/>
    <po:assembledWith rdf:resource="http://ai.dankook.ac.kr/ABontology.owl#문장 구조 분석기"/>
    <po:assembledWith rdf:resource="http://ai.dankook.ac.kr/ABontology.owl#형태소 분석기"/>
  </owl:Class>
  <owl:ObjectProperty rdf:about="http://ai.dankook.ac.kr/PredicateOntology.owl#assembledWith">
    <owl:inverseOf rdf:resource="http://ai.dankook.ac.kr/PredicateOntology.owl#partOf"/>
  </owl:ObjectProperty>
</rdf:RDF>
```

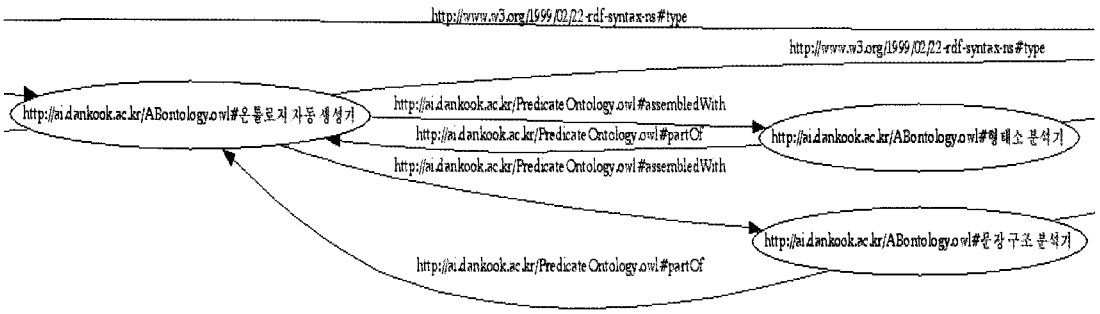

그림 5. 생성된 온톨로지 (일부)

표 8. 온톨로지 구축가와 제안된 시스템 비교

예문 크기		주어진 예시문	온톨로지 구축가	제안된 시스템
500 문장	개념	856 개	847개 (9개 못 찾음)	886개 (30개 더 찾음)
	개념간의 관계	1944 개	1923개 (11개 못 찾음)	2152개 (108개 더 찾음)

방법을 실행하여 그 결과를 비교하였다. 성능 측정 방법으로서 개념과 관계가 미리 결정되어 있는 주어진 문장들로부터 제안한 방법을 통하여 추출한 개념과 관계의 개수를 측정하였다. 결과는 표 8과 같다.

제안한 온톨로지 자동 생성 시스템의 결과로부터 성공적으로 찾아낸 개념이 총 886개로서 예시 문에 비하여 30개를 초과하여 찾아내었다. 초과 달성된 이유를 분석하기 위해 실험 진행 기록을 분석한 결과, 자연어 처리기가 잘못 동작되어 발생한 것으로 판명되었는데, 예를 들어 주어진 문장이 "온톨로지 생성기는 온톨로지를 생성한다."일 때 주어, 서술어, 목적어 쌍이 "온톨 로지 생성기, 지르, 온톨"과 같이 자연어 처리기 형태소 분석이 잘못된 결과를 출력한 결과이다. 개념간의 관계의 경우에도 온톨로지 구축가는 1923개로 부족하게 찾은 반면, 제안된 시스템은 2152개로 108개가 초과되어 생성되었다. 초과되어 생성된 관계들은 예시 문에서 주어진 개념간의 관계 이외에 온톨로지를 다루기 위해 사용된 Jena[19] 프레임워크에서 자체적으로 생성된 관계들로 판별되었다.

5. 결 론

본 연구는 많은 비용과 시간이 소모되는 온톨로지 구축을 자동화하기 위해서 형태소 분석을 통하여 자연어 문장으로부터 각각의 형태소를 추출하고, 구문 분석을 통하여 자연어 문장의 서술구조를 분석하여 자연어 문장을 온톨로지로 구축하는 방법을 제안하였다. 그리고 자연어문장의 서술어를 온톨로지 구문의 서술어로 변환 할 수 있는 서술어 온톨로지를 제안하고 이를 사용하여 시스템을 개발하였으며, 실험을 통하여 본 연구에서 제안한 방법을 사용하면 빠른 시간에 대규모의 문장을 분석하고 온톨로지로 변환하여 저장할 수 있음을 증명하였다.

실험을 통하여 얻은 결과로부터 본 논문에서 제안한 온톨로지 자동 구축방법은 개념이 856개, 관계가 1944개 포함된 예시 문들로부터 개념과 관계를 각각 886개, 2152개 찾았으며 잘못 찾은 경우는 3.34%인 30개에 불과하였다. 인간 온톨로지 구축가는 주어진 문장들로부터 개념을 847개, 관계를 1923개 추출하여 개념과 관계를 더 적게 추출하였다.

마지막으로, 생성된 온톨로지의 정확도를 높이기 위하여 자동으로 생성된 온톨로지를 검증하고, 검증된 결과에 따라 정제하기 위한 방법에 대한 연구도 진행되어야 할 것이며, 또한 생성된 온톨로지를 다른 응용 프로그램에서 활용하기 위해서는 OpenAPI와 같이 손쉽게 사용할 수 있는 방법도 연구되어야 할 것이다.

감사의 글

자바로 구현되어있는 형태소 분석기를 제공해주신 서울대학교 이동주 박사과정님께 감사의 말씀을 전합니다.

참 고 문 헌

[1] T. Berners-Lee, J. Handler, and O. Lassila, "The Semantic Web," Scientific American, May 2001.

[2] 김수경, 안기홍, "시맨틱 웹 응용을 위한 웹 온톨로지 구축기법," 한국정보처리학회 정보처리학회 논문지 D, 제15-D권, 제01호, pp. 47-60, 2008. 2.

[3] S.J. Kang, and J.H. Lee, "Semi-Automatic Practical Ontology Construction by Using a Thesaurus," Workshop on Human Language Technology and Knowledge Management ACL2001, Toulouse France, July 2001.

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

[5] P. Clerkin, P. Cunningham, and C. Hayes, "Ontology Discovery for the Semantic Web Using Hierarchical Clustering," Trinity College Dublin Computer Science Dept., Technical Reports, 2001.

[6] A. Wrobel and O. Wurmli, "Data Mining for Ontology Building," Diploma Thesis-Dept. of Computer Science WS 2002/2003.

[7] M. Cannataro and C. Comito, "A Data Mining Ontology for Grid Programming," 1st Workshop on Semantic in Peer-to-Peer and Grid Computing at the Twelfth International World Wide Web Conference, May 2003.

[8] 임수연, 구상옥, 송무희, 이상조, "접미사 패턴을 이용한 온톨로지의 구축 방안," 한국정보과학회 2003년 추계학술대회, Vol. 30, No. 2-1, pp. 547-549, 2003. 10.

[9] J. Saias and P. Quaresma, "Using NLP Techniques to Create Legal Ontologies in Logic Programming Based Web Information Retrieval System," In Proceedings of the International Conference on Artificial Intelligence and Law, June 2003.

[10] 구미숙, 황정희, 류근호, 홍장의, "데이터마이닝 기법을 이용한 XML 문서의 온톨로지 반자동 생성," 한국정보처리학회논문지 D, 제13-D권, 제3호, pp. 299-308, 2006. 6.

[11] 김희수, 최익규, 김민구, "개념간 관계의 추출과 명명을 위한 통계적 접근 방법," 한국정보처리학회논문지 B, 제12-B권, 제4호, pp. 479-486, 2005. 8.

[12] WordNet, http://wordnet.princeton.edu/

[13] 세종계획, http://www.sejong.or.kr.

[14] VerbNet, http://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

[15] PropBank, http://verbs.colorado.edu/~mpalmer/projects/ace.html.

[16] NomBank, http://nlp.cs.nyu.edu/meyers/NomBank.html.

[17] FrameNet, http://framenet.icsi.berkeley.edu/.

[18] Cheoli, http://ids.snu.ac.kr/wiki/철이.

[19] Jena - A Semantic Web Framework for Java, http://jena.sourceforge.net/

민 영 근

2005년 단국대학교 전기전자컴퓨터 공학과(학사)
2007년 단국대학교 대학원 전자컴퓨터학과 졸업(공학석사)
2007년~현재 단국대학교 대학원 컴퓨터학과 박사과정 재학

관심분야 : 시맨틱 웹, 기계 학습

이 복 주

1986년 서울대학교 컴퓨터공학과 졸업(공학사)
1992년 South Carolina 대학교 컴퓨터학과 졸업(공학석사)
1996년 Texas A&M 대학교 컴퓨터학과 졸업(공학박사)
1997년~1999년 AT&T Senior Member of Tecnical Staff
2000년~2001년 ICU 조교수
2001년~현재 단국대학교 컴퓨터학과 교수
관심분야 : 인공지능, 웹 정보처리, 시맨틱 웹