도메인 온톨로지 구축을 위한 개념 자동 추출 및 클러스터링*

정현기○ 김유섭
{mayapple, yskim01}@hallym.ac.kr
한림대학교 컴퓨터공학과

Automatic Extraction and Clustering of Concepts for Domain Ontology Construction

Hyun-Ki Jung○ Yu-Seop Kim
Dept. of Computer Engineering, Hallym University

요 약

기존의 온톨로지 구축에 관한 연구들을 살펴보면 개념의 상·하위 관계와 관련한 연구 또는 기구축된 도메인에 특화된 지식베이스에 기반한 도메인 온톨로지 구축 연구가 주를 이룬다. 그러나 개념과 개념간의 관계는 상·하위 구조와 같은 단순한 계층적 구조로는 그 다양한 특성을 표현할 수 없으며, 도메인 온톨로지를 구축하는 경우에 기구축된 데이터베이스와 같은 개념간 관계가 잘 정의된 데이터는 반드시 필요하였다. 예를 들면, 다양한 지식이 구축되어 있는 데이터베이스나 특정 도메인에 관한 전문 사이트(예 : 의학정보, 약학정보 사이트) 등이 있어야 개념간의 다양한 관계가 표현되어 있는 온톨로지를 구축할 수 있었다. 본 연구에서는 도메인 온톨로지를 구축함에 있어서 이러한 제약을 극복하기 위하여 도메인에 특화된 문서들을 웹 검색을 통하여 수집하였고, 수집된 문서 데이터를 이용하여 자동으로 도메인에 특화된 개념들을 추출하고 이들 개념들을 클러스터링함으로써 개념들간의 다양한 관계를 표현할 수 있는 도메인 온톨로지의 자동 구축 가능성을 제시한다.

1. 서 론

차세대 웹 기술인 시맨틱 웹(Semantic web)은 최근 많은 연구가 활발히 진행되고 있다. 시맨틱 웹이 각광받으면서 시맨틱 웹을 구현하기 위한 핵심 기술인 온톨로지에 관한 연구 또한 활발하게 진행중이다.

Tom Gruber의 정의를 인용하면 '온톨로지는 개념의 명세화' 라고 정의할 수 있다[1]. 즉, 개념의 종류와 개념(concept)들 간의 관계(relation)를 명백하게 정의하는 것이 온톨로지의 주역할이라 볼 수 있다. 온톨로지의 구축 단계를 간단히 설명하면, 특정 도메인을 선정하고, 개념을 자동 또는 반자동으로 추출하며 추출된 개념들의 관계를 설정한다[2]. 이러한 온톨로지 구축 과정에서 어려운 점은 온톨로지 구축의 이론적인 체계와 원리가 아직 미흡하다는 것이다. 기존의 구축 사례를 보면 이미 만들어져 있는 시소러스나 의미 분류체계를 이용한 사례들이 많다. 그 예로 워드넷을 이용하여 온톨로지를 구축하는 방법[3]과 시소러스 기반 온톨로지 구축[4] 등이 있다. 하지만 이러한 연구에서는 개념들간의 다양한 관계을 찾아내기에는 한계를 보인다. 워드넷은 단어를 'synset'이라는 유의어 집단으로 분류하여 어휘목록 사이에 동의어, 상·하위어, 전체·부분 관계를 기술해 놓은 계층구조를 나타낸 개념 어휘망이다.

시소러스 또한 용어들 사이의 상위 개념, 하위 개념, 동의어 등의 관계에 대한 정보를 주는 어휘 도구이다. 워드넷이나 시소러스를 이용한 온톨로지 구축은 구축시간을 줄일 수 있는 장점을 가지고 있다. 하지만 개념과 개념들 사이의 다양한 관계를 표현하는 데에는 단순한 계층 구조는 한계를 가진다.

또한 특정 도메인의 온톨로지를 구축하는 경우, 그 도메인에 관한 지식이 필요하다. 기존의 연구를 보면 이미 잘 정리되어 있는 데이터베이스를 기반으로 온톨로지를 구축하는 경우가 많았다[5]. 또는 특정 도메인에 관한 전문 사이트에서 정의된 다양한 관계를 활용하여 새로운 도메인 온톨로지를 구축하였다[2]. 이렇게 잘 정리되어 있거나 특정 도메인에 관한 사이트를 이용하면 온톨로지를 구축하는데 있어 많은 이점이 있다. 즉 구축 시간이 단축되며, 구축에 들어가는 비용도 절감할 수 있다. 하지만 특정 도메인에 관한 기 구축된 데이터베이스 또는 전문 사이트의 데이터들이 부족하다면 구축에 어려움이 있다.

본 연구에서는 도메인에 적합한 온톨로지를 구축하기 위하여 먼저 웹 검색을 통하여 도메인을 대표하는 개념들을 추출하였다. 개념 추출을 위해서는 대표 키워드를 통해 검색질의를 생성하여 포털사이트의 검색엔진을 이용하여 웹 문서를 수집한다. 그리고 수집한 웹문서를 정

* 본 연구는 산업자원부의 지역혁신 인력양성사업 (KOTEF)의 지원으로 이루어졌습니다.

제하여 텍스트 코퍼스 데이터를 구축하고 이 데이터를 토대로 도메인을 대표하는 개념들을 추출한다. 추출된 개념들은 워드넷[7]을 사용하여 자동으로 클러스터링되는데, 이들 클러스터간의 다양한 관계는 추후 관련 텍스트 데이터를 분석함으로써 추출될 수 있다. 본 연구에서는 클러스터링까지를 연구의 주요 범위로 하였다. 따라서 본 논문에서의 클러스터링 이후의 과정은 그 가능성을 제시하는 것으로 제한하였다.

2장에서는 본 연구에서 제시된 전체 시스템의 구조에 대하여 간략하게 설명한다. 그리고 3장에서 5장까지는 2장에서 제시된 구조의 하위 구조에 대하여 상세하게 설명한다. 마지막으로 6장에서는 결론과 향후 연구에 대하여 설명하였다.

2. 전체 시스템 구조

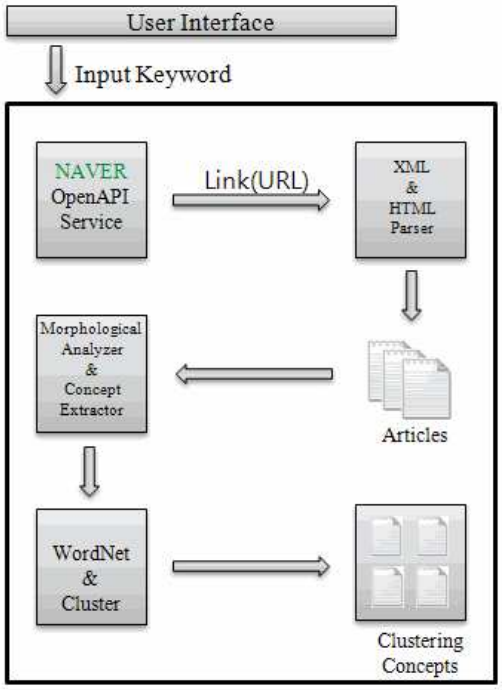

그림 1. 전체 시스템 구성도

온톨로지 구축을 위해 개념을 자동으로 추출하고 이들 개념들의 관계를 설정하기 위하여 클러스터링을 하는 개괄적인 과정은 그림 1에서 보인다. 본 연구에서는 Naver OpenAPI 서비스*를 통해 얻은 링크 정보를 이용하여 웹문서를 수집하고 파서를 통해 기사원문을 뽑아낸다. 기사원문을 형태소 분석 과정[6]을 거쳐 색인되 용어들의 빈도 정보를 분석 하여 도메인에 특화된 개념을 추출한다. 그리고 워드넷을 이용하여 개념들을 클러스터링한다.

본 연구에서는 위와 같은 시스템을 구축하였고 그 결과로 얻어진 클러스터링된 개념들을 이용하여 실제 온톨로지 구축의 가능성을 보이고자 한다.

* http://openapi.naver.com/

3. 키워드를 이용한 도메인 코퍼스 수집

본 연구에서는 실험 도메인을 '스키'로 정하고 개념 추출을 위한 코퍼스는 스키관련 신문기사로 한정하였다. 신문기사를 수집하기 위해 Naver에서 제공하는 Open API 뉴스검색 서비스를 이용하였다. Naver Open API 서비스는 검색결과를 얻기 위해 표 1과 같은 요청변수를 조합하여 다음과 같은 URL을 생성한다.

http://openapi.naver.com/search?key=등록key&query=스키&target=news&start=1&display=100'

위와 같이 생성된 URL을 통해 그림 2와 같은 검색 결과를 XML 문서형태로 얻을 수 있다. 위의 URL을 통해서는 100개의 기사만을 얻을 수 있으나 본 연구에서는 start변수를 100씩 증가시켜 최대 1000개의 신문기사의 URL을 얻을 수 있었다.

표 1. Naver Open API 요청 변수

요청변수	설명
key	이용 등록 key
target	news : 뉴스검색
query	검색 질의(keyword)
display	출력건수(최대 100)
start	검색 시작위치(최대 1000)
sort	정렬 옵션(date:날짜순, sim:유사도순)

그림 2. 검색결과로 얻어진 XML 문서

그림 2와 같은 검색 결과로 얻어진 XML문서에서 <link>와 </link>의 사이의 정보를 추출하여 기사원문의 URL을 XML파서를 통해 모두 추출한다. 추출된 URL 정보를 통해 HTML형태의 웹문서를 얻고 HTML파서를 통해 기사원문을 추출하여 전체 문서에서 원문 추출이

가능한 407개의 기사원문을 텍스트 파일로 변환시켰다. 그림 3은 이러한 텍스트 파일의 추출 과정을 보여준다.

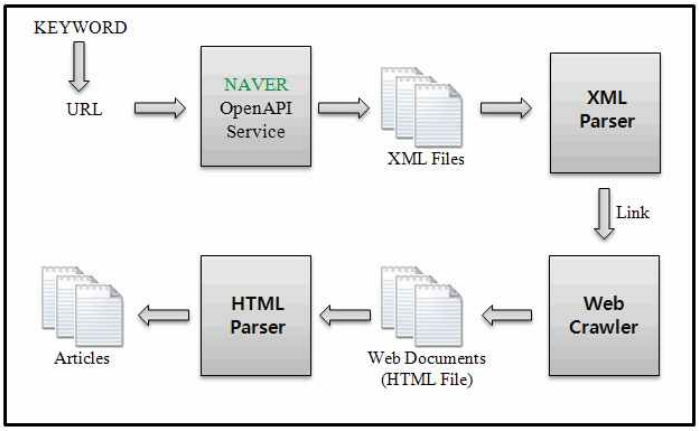

그림 3. 키워드를 이용한 문서 수집과정

4. 개념 자동 추출

3장의 과정을 통해 얻어진 도메인 코퍼스에서 형태소 분석기를 이용해 주요 용어 및 용어의 빈도 정보를 추출한다. 표 2는 코퍼스의 주요 용어 및 빈도 정보를 보여주고 있는데 좌측 표는 3장의 과정을 통하여 새로이 구축된 '스키' 관련 코퍼스에서 추출된 정보이고, 우측 표는 기존에 보유하고 있는 도메인에 특화하지 않은 일반 신문 기사 코퍼스로부터 추출한 정보이다.

표 2. 코퍼스 형태소분석 결과 일부분

도메인 코퍼스

용어	Frequency
스키	1403
것	884
이	850
등	778
그	444
스키장	374
대회	355
한국	334
기자	295
러시아	274
때	258
하	257
세계	242
미국	238
눈	231
리조트	220
선수	199
슬로프	182

일반 코퍼스

용어	Frequency
것	29725
이	28565
등	28045
수	20351
한	15804
하	13687
대한	12886
한국	12674
때문	12505
때	11764
서울	11547
뒤	10678
중	10467
나	10115
원	10024
데	9457
미국	9446
저	9263

본 연구에서는 407개의 문서로부터 25,175개의 용어를 추출할 수 있었다. 이들 용어 중에서 도메인에 특화된 개념으로 사용할 수 있는 것을 선별하기 위해서 다음의 2가지 과정을 적용하였다.

첫 번째 과정은 '것', '이', '등'과 같은 불용어를 제거하고, 어느 문서에나 자주 나오는 용어 즉, 특정 도메인과는 상관없는 용어를 제거하는 것이다. 이 과정을 위해서 일반 코퍼스(신문기사 39,331개)에서 용어와 그 빈도를 뽑아 도메인 코퍼스의 용어의 빈도와 일반 코퍼스의 용어의 빈도를 비교하였다. 일반 코퍼스에서 빈도가 1,000 이상인 용어는 일반적으로 자주 쓰이는 것이라고 할 수 있고 빈도가 10 이하인 용어는 일반적으로 쓰이지 않는 용어이라고 할 수 있다. 도메인 코퍼스에서 뽑은 용어들 중에서 일반 코퍼스에서 빈도가 1000이상이고 10이하인 용어들을 제거한다.

두 번째 과정은 첫 번째 과정에서 제거된 용어 중에서 비록 자주 쓰이거나 혹은 자주 쓰이지 않는 용어라도 특정 도메인에 개념의 일부가 될 수 있는 용어를 추가 시키는 과정이다. 두 코퍼스에서 각 용어의 빈도 순위를 백분율로 계산하여 일반 코퍼스에서 보다 도메인 코퍼스에서 그 위치가 상위에 있다면 그 용어를 추가 시킨다. 예를 들어 'a'라는 용어의 빈도가 일반 코퍼스에서는 상위 90%에 있고 도메인 코퍼스에서는 상위 10%에 있다면 'a'는 도메인에 특화된 개념의 일부로 인정하는 것이다. 이와 같은 과정을 거쳐 6,255개의 용어를 추출하였고 그 용어의 빈도가 5이하인 용어를 제거하고 2,711개의 개념을 추출하였다. 그림 4는 이러한 개념의 추출 과정을 보여준다.

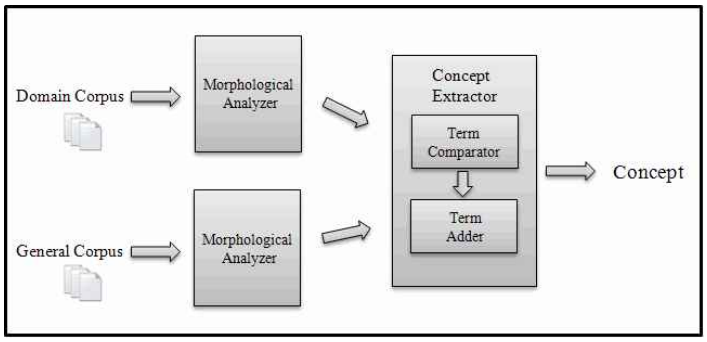

그림 4. 개념 추출 과정

표 3은 이 장에서 제시된 추출 과정을 거쳐서 실제 추출된 '스키'와 관련한 주요 개념들이다.

표 3. 추출된 개념의 일부분

용어	용어	용어
스키	산악	겨울
스키장	코스	할인
리조트	여행	스노보드
만원	강원	올림픽
슬로프	리프트	스키어

5. 워드넷을 이용한 클러스터링

5.1 워드넷(WordNet)

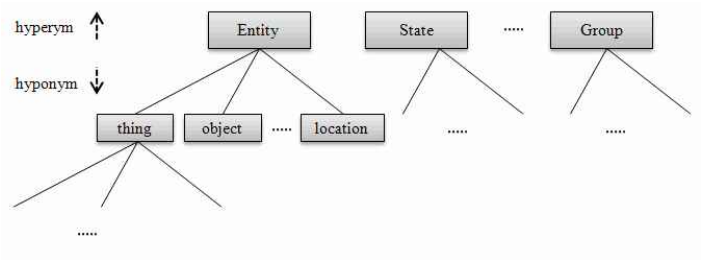

그림 5. 워드넷 계층구조

표 4. 워드넷의 데이터베이스 일부분

synsetOffset	word	hypernym	hyponym
00001740	entity 0		00002056_n_...
00002056	thing 0	00001740_n_...	00002342_n_...
00002342	anything 0	00002056_n...	
00002452	something 0	00002056_n...	
00002560	nothing 0, nonentity 0	00002056_n...	

워드넷[7]은 단어를 'synset'이라는 유의어 집단으로 분류하여 어휘목록 사이에 동의어, 상·하위어, 전체·부분 관계를 기술해 놓은 계층구조를 나타낸 개념 어휘망이다. 본 연구에서 사용한 워드넷은 9개의 최상위 레벨의 개념을 가지며 각 노드와 노드는 상·하위 의미관계로 연결되어있다. 그림 5는 워드넷의 계층구조를 보여주고 있으며 표 4는 워드넷을 데이터베이스화한 결과를 보여준다. 그림 5는 워드넷의 9개 최상위 계층의 개념들과 'entity' 개념의 하위 개념들을 보여준다.

최상위 레벨 어휘	→ hyponym	entity 하위어
entity		thing_0
psychological_feather		object
abstraction		substance
state		location
event		thing_1
act, human_act, human_activity		anticipation
group		enclosure
procession		expanse
phenomenon		sky

그림 5. 최상위 레벨 어휘 및 entity의 하위어

5.2 클러스터링

본 연구에서 추출된 개념들을 클러스터링 하기 위해 워드넷의 최상위 레벨 개념 9개를 이용한다. 즉, 추출된 개념이 어느 최상위 개념의 하위 노드에 있느냐에 따라 개념의 그룹이 정해지게 되는 것이다. 개념들을 클러스터링 한 결과 대부분이 'entity' 그룹으로 클러스터링 되었다. 그래서 본 연구에서는 'entity'의 바로 아래 레벨의 개념을 포함시켜 그림 5의 17개의 그룹으로 세분화 하였다.

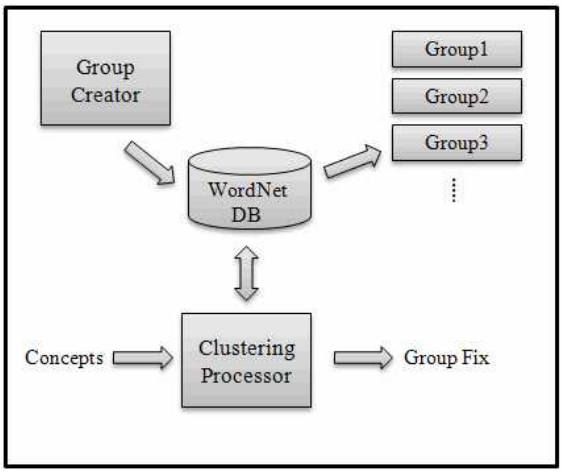

그림 6. 클러스터링 과정

클러스터링 과정 중 첫 번째는 그룹을 생성하는 것이다. 이를 위해서 워드넷 데이터베이스의 상위어 필드 값이 'NULL' 인 개념을 찾는다. 상위어 필드가 'NULL' 값을 갖는다는 의미는 더 이상의 상위어가 없다는 의미로 최상위 레벨의 개념인 것이다. 이러한 방식으로 9개의 그룹을 생성하고 추가로 'entity'의 한 단계 아래의 개념들을 포함 시켜 총 17개의 그룹을 생성하였다. 그 다음 과정은 추출된 개념들을 위 과정에서 생성된 그룹에 지정해 주는 것이다. 본 연구에서 사용된 워드넷 데이터베이스에는 3개의 테이블이 있다. 한글 단어 테이블(StdDic)과, 영어 단어(synset), 그리고 한글과 영어를 번역해 주는 테이블(Tran)이 존재한다. 예를 들어 스키라는 한글 단어는 StdDic 테이블에서 인덱스 243578의 값을 갖고 그 인덱스 값을 Tran 테이블에서 stddicIdx 필드에서 매칭시켜 wordnetIdx 값을 얻어 Synset 필드에서 한글 단어를 영어 단어로 바꿔준다.

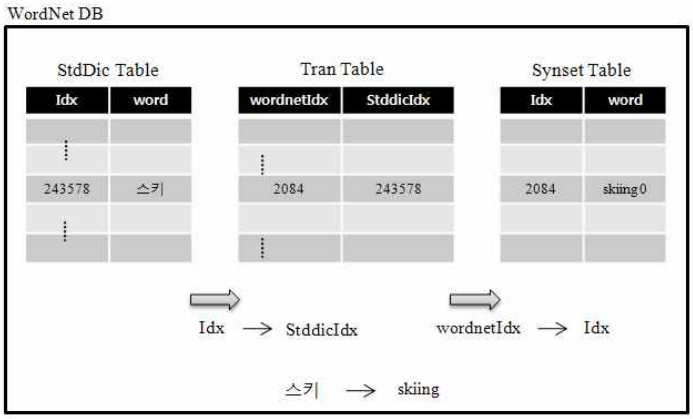

그림 7. 단어 변역과정(스키→skiing)

추출된 개념들은 위와 같은 번역 과정을 통해 Synset 테이블에서 그룹이 지정되게 된다. 예를 들어 표 4에서

영어로 번역된 단어가 'nothing' 이라고 할 때 'nothing'의 그룹을 찾아 가는 과정은 hypernym 필드에서 상위어를 찾는다. 'nothing'의 상위어는 00002056 이라는 값을 가진다. 그 값을 synsetOffset 필드에서 매칭되는 값을 찾는다. 0002056의 word는 'thing' 이고 다시 이 과정을 반복하여 hypernym 필드가 'NULL' 값을 찾을 때 까지 반복을 시킨다. 위의 경우는 'thing'이 그룹이기 때문에 반복 과정을 거치지 않고 'nothing'는 그룹 'thing'로 지정되게 된다. 그림 8에서는 이러한 과정을 보여준다.

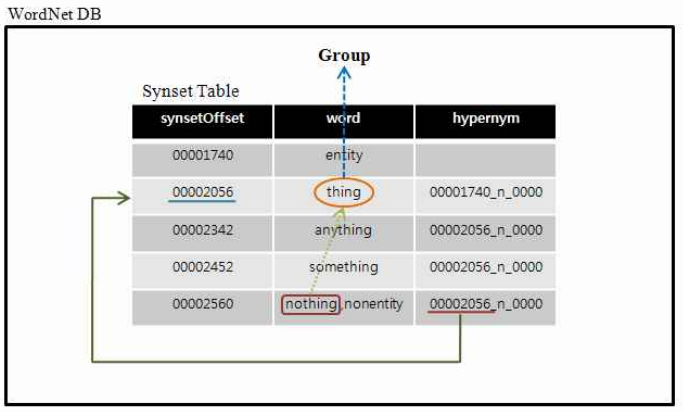

그림 8. 그룹 지정 과정('nothing' → Group 'thing')

이러한 과정을 거쳐 그림 9와 같은 클러스터링의 결과를 얻을 수 있다. 여기서 구축된 클러스터를 토대로 기본 도메인 온톨로지를 구축할 수 있다.

표 5. 클러스터링된 개념의 일부분

Group Name	Concepts
abstraction	겨울, 금메달, 여름 ...
act	스키, 여행, 할인, 관광, 강습, 출발 ...
event	올림픽, 동계, 뉴스...
group	연맹
location	러시아, 모스크바, 캐나다, 말레이시아, 우주...
object	스키장, 코스, 리프트, 온천, 산, 보드...
psychological feature	자세, 원리, 테마 ...
thing	어깨, 무릎

5.3 그룹간의 관계설정 사례

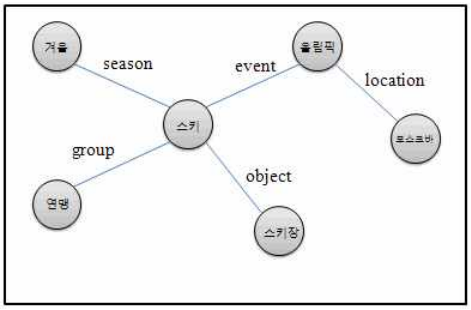

그림 9. 관계설정 사례

각 그룹간의 관계를 설정하기 위해서 본 연구에서는 추출되어진 그룹들의 개념을 보며 직접 개념간의 관계를 그림 9와 같이 설정하여 보았다. 개념간의 관계는 워드넷으로 클러스터링 된 각 그룹의 관계들이다. 이러한 관계들은 도메인 코퍼스를 구문/의미 분석함으로써 자동적으로 유추될 수 있을 것이다.

6. 결론 및 향후 연구

본 연구에서는 도메인 코퍼스를 자동으로 수집하여 수집된 코퍼스에서 개념을 자동으로 추출하고 클러스터링 하였다. 여기서 구축된 클러스터를 토대로 기본 도메인 온톨로지의 구축 가능성을 제시하였다.

그러나 더욱 도메인에 적합한 개념들을 추출하기 위해서는 더 많은 양의 코퍼스가 필요하고 코퍼스를 정교하게 정제하는 과정이 필요하다. 예를 들면 추출된 개념 중에 '우주', '이소연' 과 같은 '스키'와는 관련이 적은 개념이 추출되었다. 그 이유는 최근 기사에 '한국 최초우주인 이소연씨'에 관한 기사가 많이 실렸기 때문이다.

그리고 개념간의 관계 설정에서도 단순한 관계 밖에 설정하지 못하였다. 워드넷을 이용하여 개념간의 관계를 설정하기에는 부족함이 있다.

다양한 코퍼스 수집과 도메인에 적합한 코퍼스를 정제하기 위한 연구가 앞으로 더 진행 되어야 할 것이고 온톨로지를 자동으로 구축하기 위해 개념과 개념사이의 관계를 다양하게 설정하기 위해서 문장 분석과정을 통해 구문관계, 의미관계 등의 다양한 관계를 확장해 나아가는 연구가 활발하게 진행되어야 할 것이다.

참고문헌

[1] 서휘, "온톨로지 자동구축을 위한 OWL의 어휘와 구문 사용법에 대한 이론적 연구", 정보과학회지 제 37권, 제2호, pp. 191-216, 2006.
[2] 임수현, 박성배, 이상조, "의미관계 정보를 이용한 약품 온톨로지의 구축과 활용", 정보과학회 논문지, 제 32권, 제 5호, pp. 428-437, 2005.
[3] 공현장, 황명권, 김원필, 김판구, "특정 도메인에 대한 자동 온톨로지 구축 방법에 관한 연구", 한국정보과학회 추계학술발표 논문집, 제32권, 제2호, pp. 595-597, 2005.
[4] 한국영, 이현실, "시소러스를 활용한 온톨로지 구축 방안 연구", 한국비블리아학회지, 제 17권, 제 1호, pp. 285-303, 2006.
[5] 송도규, "대용량 OWL 온톨로지 자동구축을 위한 세종전자사전 활용 방법론 연구", 한국언어정보학회지, 제 9권, 제 1호, pp. 19-34, 2005.
[6] 한국어 형태소 분석기 KTL version 2.1.0f
http://nlp.kookmin.ac.kr
[7] KoreanWord-Net ver.2.0. Korean Language Processing Laboratory, Pusan National University.