프레임넷을 통한 디비피디아 온톨로지 인스턴스 생성의 커버리지 개선

함영균○, 서지우, 황도삼†, 최기선
한국과학기술원, 영남대학교†
hahmyg@kaist.ac.kr, jiwoo35@kaist.ac.kr, dshwang@yu.ac.kr, kschoi@kaist.ac.kr

DBpedia Ontology Population Coverage Enhancement with FrameNet

Younggyun Hahm○, Jiwoo Seo, Dosam Hwang†, Key-Sun Choi
KAIST, Yeungnam University†

요 약

비구조 텍스트로부터 지식을 추출하여 온톨로지 기반 지식베이스를 구축하는 연구가 최근 국내외로 다양하게 진행되고 있다. 이러한 목적을 달성하기 위해서는 자연어 텍스트에서 나타난 지식요소들의 다양한 속성들을 표현할 수 있는 온톨로지를 필요로 한다. 디비피디아 역시 위키피디아의 지식들을 표현하기 위하여 디비피디아 온톨로지를 사용한다. 그러나 디비피디아 온톨로지는 위키피디아의 인포박스에 기반한 온톨로지로서, 요약된 정보를 설명하기에는 적합할 수 있으나 자연어 텍스트로 표현된 다양한 지식표현을 충분히 커버하는 것은 보증되지 않는다. 본 논문에서는 자연어 텍스트로 쓰여진 지식을 디비피디아 온톨로지가 충분히 표현할 수 있는지를 검토하고, 또한 그 불완전성을 프레임넷이 어느정도까지 보완할 수 있는지를 살핀다. 이를 통해 한국어 텍스트로부터 지식베이스를 자동구축하는 온톨로지 인스턴스 자동생성 연구의 방향으로서 디비피디아 온톨로지와 프레임넷의 효용성을 전망한다.

주제어: 디비피디아, 프레임넷, 온톨로지, 지식베이스

1. 서론

최근 시멘틱 웹과 빅데이터 및 질의응답 시스템 등에 대한 연구가 활발해지고 있으며, 이에 따라 온톨로지 기반 지식베이스를 구축하는 많은 연구들이 진행되고 있다[1][2]. 그러나 전통적으로 지식은 자연어로 작성되어 있으며, 특히나 몇몇 연구에 의하면 구조화된 데이터베이스보다 비구조 데이터에서 보다 많은 지식이 포함되어 있다고 알려져 있다[3][4]. 따라서 기존 지식베이스를 확장하기 위하여 자연어텍스트인 비구조 데이터로부터 온톨로지의 인스턴스들을 자동으로 생성하는 연구들이 최근 진행되고 있다.[5][6][7]. 특히, 시멘틱 웹의 관점에서 웹상의 지식을 구조화된 RDF로 표현하기 위해서는 지식요소들의 다양한 속성들을 충분히 설명할 수 있는 프로퍼티를 갖고 있는 온톨로지를 요구한다[8].

시멘틱 웹의 최신 기술인 디비피디아는 지식 코퍼스인 위키피디아로부터 자동구축된 지식베이스이며[9], 위키피디아의 지식을 표현하기 위하여 디비피디아 온톨로지를 사용한다. 그러나 디비피디아 온톨로지는 위키피디아의 인포박스에서 기원한 온톨로지로서, 위키피디아의 요약된 지식을 표현하기에는 충분할 것으로 간주할 수 있지만, 위키피디아 텍스트상의 모든 지식을 표현할 수 있는지는 보장되지 않는다.

따라서 본 논문에서는 웹상의 지식으로부터 온톨로지의 인스턴스들을 자동으로 생성하기 위하여서 디비피디아 온톨로지가 지식표현을 위해 충분한 것인가를 검토하고, 지식요소들인 인스턴스 개체들간의 관계를 설명할 수 있는 또다른 언어자원인 프레임넷[10]을 사용하여 디비피디아의 불완전성을 보완할 수 있는가를 검토한다.

이를 위하여, 2장에서는 본 논문의 문제를 정의하고, 3장에서는 실험대상으로서 한국어 위키피디아 텍스트에서 지식요소를 포함하고 있는 문장을 정의한다. 그리고 이에 대하여 4장과 5장에 걸쳐 디비피디아 온톨로지와 프레임의 지식표현 커버리지를 계산하고, 그 실험 결과를 6장에서 논한다.

2. 문제정의

자연어텍스트로 쓰인 지식을 구조화된 지식베이스로 구축하기 위해서, 특히 RDF로 기술하기 위해서는 지식요소들의 다양한 속성들을 표현할 수 있는 온톨로지 프로퍼티가 요구된다. 디비피디아는 위키피디아의 지식을 표현하기 위하여 개체와 클래스, 그리고 프로퍼티로서 dataproperty 와 objectproperty를 사용한다. 그러나 디비피디아 온톨로지는 위키피디아의 인포박스에 기반한 온톨로지로서, 요약된 지식을 표현하기에는 충분할 수 있으나(예: 이름, 직업, 생년월일, 장소, 인구밀도 등) 위키피디아 텍스트의 모든 지식을 표현하는 것은 보증되지 않는다(예: 원인, 결과, 감정, 의견, 행동, 문제와 해결 등). 특히 질의응답의 목적으로서 다음과 같은 질문을 생각해 보자.

Q	이것은 바이러스에 감염된 동물 세포가 생성하는 당단백질이다. 바이러스의 감염과 증식을 저지하는 작용을 한다. 유전공학의 발달로 대량생산되며, B형 간염이나 헤르페스(포진) 따위의 바이러스 질병 치료에 쓰인다. 이것은 무엇인가?
A	인터페론

이 경우, 디비피디아 온톨로지는 정답 '인터페론'이 '당단백질'이라는 type을 RDF로 표현할 수 있지만, 보다 중요한 정보인 '감염된', '생성하는', '저지하는', '작용을 한다', '치료' 등을 기술하는 것은 디비피디아 3.9버전[1]에서는 불가능하다.

이에 본 논문에서는 논문 [13]의 방법론을 따라 디비피디아 온톨로지가 웹 텍스트로부터 지식베이스를 구축하는데 있어서 충분한 커버리지를 갖는가를 검토하고, 그 커버리지를 증가하기 위하여 프레임넷을 고려하여 이 역시 검토한다. [13]에서 사용된 데이터셋을 3장에서 다시 설명하고, 이를 위하여 추가적으로 진행된 작업인 디비피디아 온톨로지에 대한 어휘화 작업 및 디비피디아와 프레임넷에 대한 상호검토 작업을 4장 이후에 논한다.

1) 한국어 위키피디아에서 지식요소인 개체를 포함하는 문장셋(KS)을 정의한다.
2) 디비피디아 온톨로지 및 프레임넷이 KS의 동사들을 커버하는지 계산하기 위하여, 온톨로지를 어휘화(ontology lexicalization) 한다. 이를 위하여 [11]의 연구방법을 따른다.
3) KS에 대하여 디비피디아 온톨로지 및 프레임넷의 커버리지를 계산한다.
4) 이후, 디비피디아 및 프레임넷을 상호 검토하여 프레임넷의 기여도를 검토한다.

3. 데이터셋 정의

3.1 지식요소인 개체를 포함하는 문장셋

본 논문에서는 위키피디아 전체 텍스트를 고려하지 않고, 위키피디아 문장 중 디비피디아의 트리플의 Subject와 Object를 모두 포함하고 있는 문장들을 평가데이터셋으로 간주하였다. 이는, 디비피디아 온톨로지가 개체의 속성을 표현하기 위한 목적인데 개체가 포함되지 않은 문장들까지 모두 고려하게 된다면 지나치게 낮은 커버리지를 보여, 실험결과를 비교할 경우 잘못된 분석을 하는 것을 방지하기 위함이다.

위키피디아 텍스트에서 어떤 문장이 개체를 포함하고 있는가를 결정하는 가장 기본적인 방법은, 특정 문자열이 '링크'를 갖고 있다면 그것은 디비피디아의 URI를 가지므로 개체라고 간주하는 것이다. 그러나 위키피디아의 대부분의 개체들은 링크를 갖고 있지 않으므로, 이러한 링크를 자동으로 리태깅해주는 작업을 수행하였다. 본 방법은 [12]의 방법을 따라 수행되었다.

이를 통하여 한국어 위키피디아(2014년 1월 26일)[2]의 2,862,181 문장에 대하여 502,674개의 KS셋을 추출하였다. 그 결과는 아래 표 1과 같다.

- Subject only: 디비피디아 트리플 중 Subject 개체만을 포함하는 문장
- Object only: 디비피디아 트리플 중 Object 개체들만을 포함하는 문장
- Both S&O: 디비피디아 트리플 중 Subject와 Object를 모두 포함하는 문장

KS 유형	문장 개수	동사 개수	동사/문장
Subject only	364,899	337,195	0.92
Object only	1,261,259	3,128,629	2.48
Both S&O	502,674	1,022,492	2.03
sum	2,128,832	4,488,316	2.11

표 1. KS셋 결과

표 1에서, Subject only 유형의 KS의 경우, 디비피디아 트리플 중 Subject에 해당하는 것만이 포함된 문장인데, 이 경우에는 해당 Subject의 상태만을 기술하는 동사가 등장하는 경향이 있고, 이에 따라 동사가 문장에서 약 1번정도 등장하는 것을 볼 수 있다. 그러나 Object only 및 Both S&O 유형의 경우 문장에서 동사가 평균 2번 이상 등장하는 것을 확인할 수 있는데, 이는 하나의 문장에서 개체들간의 관계를 표현하는 동사들이 2개 이상임을 알 수 있다. 특히 Object only의 경우에는 주어가 없거나 대명사인 문장들인 경우가 많은데, 이러한 상호참조문제는 본 논문의 범위에서 고려하지 않았다. 따라서 본 논문에서는 지식을 포함하는 문장셋으로서의 KS를 Subject 와 Object를 모두 포함하고 있는 문장, 즉 Both S&O 유형의 KS셋 만을 고려하였다.

3.2 문장에서의 한국어 동사 정의

3.1장에서 정의된 KS셋으로부터, 디비피디아 온톨로지 및 프레임넷의 커버리지를 계산하기 위해서는, 디비피디아 온톨로지 및 프레임넷이 해당 KS 문장의 동사를 커버하는지 여부를 검토하여야 한다. 본 논문에서는 본 연구팀이 구축한 NLPHub[3] 프레임워크를 사용하여 KS 문장에 대하여 형태소분석 및 품사태깅을 수행하였다. 여기서 고려된 한국어 동사의 품사태그 패턴은 아래와 같다.

1) nc+xsv+ '다' (예:출생+하+다)
2) pv+ '다' (예:들어가+다)

이러한 패턴을 고려한 이유로는, 실제 한국어 텍스트에서는 '출생하다'라는 동사가 '출생하는', '출생하여', '출생한' 등으로 다양하게 나타나는데 이러한 모든 동사를 커버하기 위하여, 위의 태그 조합에 대해 '다'를 붙임으로서 자연스러운 형태로 변환하였다.

1) http://blog.dbpedia.org/2014/09/09/dbpedia-version-2014-released/
2) http://dumps.wikimedia.org/kowiki/
3) http://nlphub.kaist.ac.kr

4. 디비피디아 온톨로지 커버리지

본 장에서는, 3장에서 정의된 KS셋의 한국어 동사들에 대하여 디비피디아 온톨로지가 어느정도의 커버리지를 갖는가를 검토한다. 이를 위하여서는 영어의 단어들로 이루어진 디비피디아 온톨로지를 어휘화(ontology lexicalization) 하는 작업이 필요하다.

4.1 온톨로지 어휘화 작업 개요

온톨로지 어휘화에 대한 예는 다음과 같다:

- 한국어 문장: “반기문은 1944년 6월 13일에 충청북도 음성군 원남면 상당리에서 태어났다.”
- 디비피디아 온톨로지:
  dbo:birthPlace, dbo:birthDate

위 예시에서, 디비피디아 온톨로지의 dbo:birthPlace 및 dbo:birthDate 는 개체 ‘반기문’의 출생지 및 출생연도를 표현할 수 있는 프로퍼티이다. 그리고 이러한 의미는 실제 한국어 문장에서 ‘태어났다’라는 어휘와 의미를 같이한다. 따라서 온톨로지 어휘화 작업에서 이상적인 결과로는 dbo:birthPlace 및 dbo:birthDate 의 한국어 어휘화 결과가 ‘태어났다’ 로서 나타난다. 이를 위하여 [11]의 연구에서는 영어 텍스트를 대상으로 온톨로지 어휘화 작업을 수행하였는데, 이는 크게 1) 레이블 기반 방법 및 2) 의존관계 기반 방법 2가지로 나뉜다.

4.2 레이블 기반 디비피디아 온톨로지 어휘화

디비피디아 온톨로지를 한국어로 어휘화하기 위하여서 수행한 첫 번째 방법은 레이블 기반 접근방법이다. 이는 실제 디비피디아 온톨로지의 프로퍼티를 한국어로 변환하는 방법인데, 본 논문에서는 아래 2가지 방법으로 이를 수행하였다.

1) 한국어 디비피디아 프로퍼티 (Korean Label) 사용
- 디비피디아 온톨로지가 영어로 작성되어 있지만, 한국어 디비피디아에서 사용되는 많은 프로퍼티들이 온톨로지와 매핑되어 있지 않고, 한국어 위키피디아의 인포박스의 프로퍼티를 그대로 가져와서 사용된다. 예컨대 디비피디아 온톨로지에서는 영어로 dbo:birthplace 및 dbo:area, dbo:founder 등으로 되어 있지만, 이에 대한 한국어 레이블은 prop-ko:출생지, prop-ko:면적, prop-ko:창립자 등으로 되어 있다. 따라서 본 논문에서는 디비피디아 온톨로지의 프로퍼티가 인포박스에 기반하여 위키피디아 전체 텍스트의 지식을 충분히 표현하지 못할것이라는 가정에서 출발하였기 때문에, 이러한 한국어 디비피디아 프로퍼티를 모두 사용하였다. 여기서는 총 9,080개의 한국어 프로퍼티를 모두 사용하였다.

2) 디비피디아 온톨로지 프로퍼티의 번역
- 디비피디아 온톨로지는 모두 영어로 작성되어 있기 때문에 이를 한글로 변환하는 가장 기본적인 접근법으로서 기계번역 작업을 수행하였다. 이로부터 디비피디아 온톨로지의 2,215개의 모든 프로퍼티를 한글로 번역하였다.
- 예: dbo:birthData -> 출생장소,
  dbo:occupation -> 직업,
  dbo:populationDensity -> 인구밀도

4.3 의존관계 기반 디비피디아 온톨로지 어휘화

의존관계 기반 온톨로지 어휘화 작업은, 두 개체들 사이의 의미적 관계는 자연어 텍스트 문장에서 가장 가까운 ‘동사’로서 연결되어 있다는 가정을 바탕으로 한다. 이에 대한 예시는 다음과 같다.

- 한국어 문장: “반기문은 충청북도에서 태어났다.”

이 문장의 경우 두 개의 개체, 즉 ‘반기문’ 과 ‘충청북도’ 라는 개체 사이의 관계를 표현하고 있는데 이는 동사 ‘태어났다’ 로 파악할 수 있다. 그러나 [7] 연구에서는 영어의 경우 두 개의 명사 사이에 동사가 위치한다는 특성이 있기 때문에, 두 개체 사이에 존재하는 동사 및 어휘의 패턴을 사용하였지만, 한국어의 경우에는 주요 동사가 문장의 맨 뒤에 나타나거나 위치가 특정하지 않아 이러한 방법을 그대로 적용하기 어렵다. 이는 영어에서도 유사한 문제를 발생하기 때문에, [11]의 연구에서도 두 개체 사이의 동사가 아닌, 의존문법 구조에서의 두 개체와 가장 가까운 동사 노드를 사용하였다. 이에 대한 예시는 그림 1과 같다.

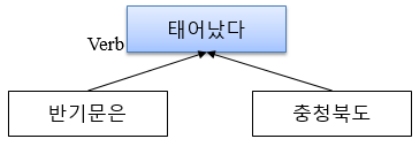

그림 1. 문장 “반기문은 충청북도에서 태어났다”의 의존구조의 예

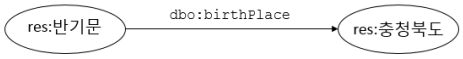

그림 2. 디비피디아 트리플의 예

그림 1에서, 동사 노드인 ‘태어났다’ 의 경우, 두 개체 ‘반기문’ 및 ‘충청북도’ 에 대한 관계를 표현하고 있다. 따라서 의존관계 기반 온톨로지 어휘화에서는 만약 ‘반기문’과 ‘충청북도’ 라는 개체가 특정한 프로퍼티로 관계를 갖고 있다면, 그 프로퍼티에 대한 어휘화 결과는 ‘태어났다’ 가 된다. 실제로 디비피디아에서는 그림 2와 같은 트리플이 존재한다.

이때, 디비피디아 온톨로지인 dbo:birthPlace 의 경우에는 한국어 ‘태어났다’로 어휘화 된다. 이러한 방법으로, 다음과 같이 의존관계 기반 디비피디아 온톨로지 어휘화 작업을 수행하였다.

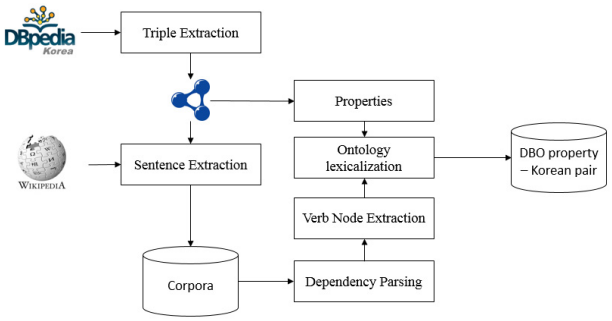

그림 3. 의존관계 기반 디비피디아 온톨로지 어휘화 작업 워크플로우

1) Triple Extraction 모듈: 디비피디아 온톨로지를 프로퍼티로 사용한 모든 한국어 디비피디아 트리플 추출
2) Sentence Extraction 모듈: 해당 트리플의 Subject와 Object를 모두 포함하고 있는 문장 추출
3) Dependency Parsing 모듈: 2의 결과 문장들을 의존관계 구문분석 수행
4) Verb Node Extraction: 3의 결과로부터, 두 개체들 사이에서 가장 가까운 거리에 있는 동사 노드 추출
5) Ontology Lexicalization: 두 개체의 관계를 갖는 프로퍼티와, 4의 결과로부터 추출된 동사 노드를 매칭하는 작업을 수행

이러한 레이블 기반 및 의존관계 기반 디비피디아 온톨로지 어휘화 작업의 결과는 아래 그림과 같다.

DBO property	Korean Lexicalization
http://dbpedia.org/ontology /birthPlace	태어나
	출생하
	출생지
	출생 장소
	...

그림 4. 디비피디아 온톨로지 어휘화 작업결과의 예

4.4 디비피디아 온톨로지 커버리지 계산

디비피디아 온톨로지의 지식표현으로서의 커버리지를 계산하기 위하여, 3장에서 지식을 표현하는 문장으로 정의된 KS셋의 한국어 동사에 대하여, 온톨로지 어휘화 된 디비피디아 온톨로지의 프로퍼티들의 커버리지를 계산하였다. 사용된 수식은 다음과 같다:

$$Coverage = \frac{n(KS \cap DBO)}{n(KS)}$$

이때, n(KS)는 전체 KS문장의 개수가 되고, n(KS∩DBO)는 디비피디아 온톨로지의 어휘화 결과 리스트를 포함하고 있는 KS문장의 개수가 된다. 이에 대한 실험 결과는 표 2와 같다.

	방법	n(KS∩DBO)	커버리지
1	한국어 디비피디아 프로퍼티	245,597	48.86%
2	디비피디아 온톨로지 번역	126,398	25.15%
3	레이블 기반 방법 (1+2)	253,553	51.24%
4	의존관계 기반 방법	72,125	14.35%
5	전체 (3+4)	280,455	55.79%

표 2. 한국어 위키피디아 텍스트에 대한 디비피디아 온톨로지의 커버리지

5. 프레임넷 커버리지

5.1 프레임넷 개요

프레임넷이란 실제 문장에서 어휘들이 어떻게 사용되는가를 시멘틱 프레임(Semantic-Frame)의 형태로 어노테이션 하여 구축된 언어자원이다[10]. 이에 대한 예시는 그림 5와 같다.

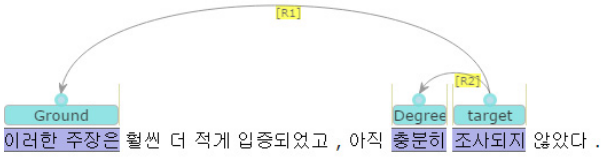

그림 5 프레임넷 어노테이션의 예

위 그림에서, 문장 “이러한 주장은 훨씬 더 적게 입증되었고, 아직 충분히 조사되지 않았다.”에서, 사용된 프레임 인덱스는 Scrutiny 이며, 이에 대한 한국어 LU(Lexical Unit)는 ‘조사되지’이다. 이 때, 프레임 논항(Frame Argument)으로서, Ground 및 Degree 등을 갖는데, 이에 해당하는 한국어 어휘들은 ‘이러한 주장은’ 및 ‘충분히’ 이다.

현재 프레임넷은 1,179개의 프레임 인덱스를 갖고 있으며(2014년 9월 12일)[4], 프레임넷의 커버리지를 계산하기 위하여 4장에서 수행했던 것과 유사한 방식으로 프레임 인덱스에 대한 한글화 작업을 수행하였다.

5.2 프레임넷 한국어 동사화

프레임넷의 프레임 인덱스에 해당하는 어휘들은 LU로서 이미 어휘화 되어 있는 상태이지만, 현재 시점에서는 프레임넷 한국어판이 공개되어 있지 않아 한국어 LU들을 사용할 수 없다. 이에, 본 논문에서는 세종용언사전을 사용하여, 영어 프레임넷의 LU의 번역에 해당하는 한국어 동사들로 매핑하였다. 세종용언사전의 경우 21,390개의 한국어 동사로 구성되어 있고, 각 한국어 동사에 대해서 영어 번역이 쌍으로 존재한다. 따라서 프레임 LU에서 특정 영어가 사용되었다면, 그에 대한 번역은 세종용언사전에서의 한국어 동사가 된다. 이러한 작업을 통하여 17,251개의 한국어 동사를 889개의 프레임 인덱스와 매핑하는 작업을 수행하였다.

4) https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=frameIndex

5.3 프레임넷 커버리지 계산

5.2장에서 구축된 한국어 동사 리스트를 사용하여, 4.4장에서 사용한 커버리지 계산 방법을 사용하였다:

$$Coverage = \frac{n(KS \cap FNindex)}{n(KS)}$$

이때, n(KS)는 전체 KS문장의 개수가 되고, n(KS∩FN.index)의 경우는, 프레임 인덱스에 매핑된 세종용언사전을 포함하는 KS문장의 개수가 된다.이에 대한 실험결과는 다음 표 3과 같다.

	방법	n(KS∩FN.index)	커버리지
1	프레임넷	359,407	73.85%

표 3. 한국어 위키피디아 텍스트에 대한 프레임넷의 커버리지

6. 결과 분석

6.1 위키피디아 텍스트의 한국어 동사에 대한 디비피디아 온톨로지와 프레임넷 커버리지

표 2와 3에서, 디비피디아 온톨로지와 프레임넷의 커버리지는 각각 55.79% 및 73.85%를 보였다. 이 결과는 한국어 위키피디아에서 지식요소인 디비피디아 개체의 Subject와 Object를 모두 포함하고 있는 문장인 KS셋에 대한 커버리지를 의미한다. 그러나 표 1에서 보인 것처럼, 각 문장들에는 평균 2개의 동사가 포함되어 있다. 그 동사들에 대한 커버리지는 표 4와 같다.

	대상	covered KS.verb	커버리지
1	디비피디아 온톨로지	359,407	42.62%
2	프레임넷	875,743	95.42%

표 4. 위키피디아 텍스트 한국어 동사에 대한 디비피디아 온톨로지와 프레임넷의 커버리지

표 4에서 보이듯, KS문장셋에서 사용된 동사들에 대한 커버리지는 디비피디아 온톨로지가 42.62%, 프레임넷이 95.42%를 보여, 프레임넷은 한국어 동사의 대부분을 커버하고 있음을 볼 수 있다. 이는, 자연어 텍스트에서 나타난 개체간의 관계인 동사를 디비피디아 온톨로지는 충분히 설명하지 못하지만, 프레임넷으로는 충분히 설명한다는 것을 의미한다고 보여진다.

6.2 디비피디아 온톨로지와 프레임넷 비교

디비피디아 온톨로지와 프레임넷을 비교하기 위하여, 아래와 같은 3가지 유형으로 분류를 해 보았다.

1) DBO∩FN.index: 디비피디아 온톨로지와 프레임넷의 프레임 인덱스가 유사한 개념인 경우
2) Only DBO: 프레임넷으로는 설명할 수 없지만, 디비피디아 온톨로지로는 설명가능한 개념
3) Only FN.index: 디비피디아 온톨로지로는 설명할 수 없지만, 프레임넷으로는 설명가능한 개념

이러한 분류를 하기 위하여, 문자열이 일치하는 경우와, 디비피디아 온톨로지의 도메인 및 레인지와 프레임 인덱스의 논항을 고려하여 매핑 작업을 수행하였고, 수작업에 의해 검증작업 및 그루핑 작업을 거쳤다.

유형	# DBO	# FN.index
시작, 끝 관련	85	12
장소, 지역, 빌딩 관련	101	11
이름 관련	66	3
사람, 직업, 관계 관련	201	14
주제, 분류 관련	62	6
수, 랭킹, 시간 관련	135	21
조직, 기관 관련	76	6
기타	233	210
계	961	283

표 5 DBO∩FN.index

표 5는 디비피디아 온톨로지와 프레임넷에 공통적으로 존재하는 개념을 의미한다. 예컨대 '시작' 관련하여서는 FN.index에는 Activity_Start 가 존재하며, 디비피디아 온톨로지에는 dbo:active_Year, dbo:launchDate 등이 존재한다. 이 결과는, 프레임넷도 위키피디아의 인포박스에 해당하는 요약된 정보를 설명하는데 있어서 어느정도는 커버리지를 갖는다는 것을 의미한다.

유형	# DBO	# FN.index
항공, 노선, 우주 관련	46	-
ID, Code, 링크 등 관련	121	-
수 (인구수, 비율 등)	143	-
측정 (무게, 높이, 거리 등)	60	-
특정 날짜, 연도 등	84	-
방향, 지역 등	116	-
수상, 경기, 스포츠 관련	78	-
기타	603	-
계	1253	-

표 6 Only DBO

표 6은 프레임넷으로는 설명할 수 없지만, 디비피디아 온톨로지를 통해 설명할 수 있는 개념을 의미한다. 결과를 볼 때, 백과사전인 위키피디아의 특성상 '숫자'에 관련된 항목이 많으며(예:dbo:population, dbo:distance, dbo:electionDate) 또한 스포츠에 관련된 프로퍼티가 많음을 알 수 있다. 이는 스포츠에 관해서는 텍스트로 쓰여진 정보의 비중보다는 요약된 통계치나 결과에 대한 정보가 인포박스에서 중요하게 다루어지기 때문이라고 보여진다.

표 7은 디비피디아 온톨로지로는 표현할 수 없지만, 프레임넷으로는 표현 가능한 개념이다. 이를 통해 프레임넷을 사용할 경우, 디비피디아 온톨로지로 할 수 없는 특정 개체에 대한 원인 및 결과, 방법, 의견 및 행동, 상태 등에 대해 기술할 수 있음을 의미한다.

유형	# DBO	# FN.index
원인, 결과	-	63
직업, 방법 관련	-	51
대회, 감정, 의견, 결정 등 관련	-	65
행동, 경험, 이벤트, 사건 관련	-	47
움직임, 제스쳐 관련	-	177
상태 (동사, 형용사)	-	123
기타 일반명사	-	90
기타	-	277
계	-	283

표 7 Only FN.index

즉, 본 논문에서 디비피디아 온톨로지가 위키피디아의 인포박스에서 기원한 온톨로지이기 때문에 요약된 지식을 표현하기에는 적합할 수 있으나, 표 7의 항목들에 대해서는 아직 불완전함을 알 수 있다. 따라서 위와 같은 지식 항목을 표현하기 위해서는 프레임넷을 통한 커버리지 증대가 이루어져야 함을 의미한다고 보여진다.

2장에서 보여진 질문셋에 경우, '생성하다'는 FN.index:Creating, '감염된'의 경우는 FN.index:Influence_of_event_on_cognizer, '저지하다'의 경우 FN.index:Intercepting, '치료하다'의 경우 FN.index:Cure 등의 프레임넷으로 표현이 가능하다.

7. 결론 및 향후 연구

본 연구팀은 한국어 자연어 텍스트로부터 지식베이스를 자동구축하기 위한 방법의 일환으로 온톨로지 인스턴스 자동생성 방법을 연구하고 있고, 본 논문에서는 이를 위하여 링크드 데이터의 중심인 디비피디아 지식베이스의 온톨로지의 커버리지를 계산하여 충분한 지식을 텍스트로부터 자동추출할 수 있는가를 검토하였다. 특히나 질의응답 시스템과 같은 복잡한 응용시스템에 적합한 지식베이스를 구축하기 위하여서는 다양한 의미적 지식이 요구되는데, 위키피디아 인포박스에 기반한 디비피디아 온톨로지는 이에 대해 어느정도 한계를 가지며, 이에 대한 대안으로서 프레임넷의 유용성을 확인하였다. 향후 본 연구팀은 프레임넷의 한국어버전을 구축하고 이를 기반으로 한 한국어 프레임 파서를 개발하여, 텍스트마이닝 및 지식베이스 구축 등의 연구에 활용할 계획이다.

사사

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신•방송 연구개발사업의 일환으로 수행하였음 [10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발]

참고문헌

[1] Kurt Bollacker, Coling Evans, Praveen Paritosh, Tim Sturge, Jamie Taylr. "Freebase: a collaboratively created graph database for structuring human knowledge." In Proceedings of:SIGMOD '08, Pages 1247-1250, 2008

[2] HJohannes Hoffart, Fabian M. Suchanek Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, Gerhard Weikum. "YAGO2: exploring and querying world knowledge in time, space, context, and many languages." In Proceeding of the WWW '11, Pages 229-232, 2011

[3] Blumberg, Robert, and Shaku Atre. "The problem with unstructured data." DM REVIEW 13: 42-49. 2003

[4] Gaag, Andreas, Andreas Kohn, and Udo Lindemann. "Function-based solution retrieval and semantic search in mechanical engineering." In Proceedings of the 17th International Conference on Engineering Design. 2009.

[5] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., Tom M. Mitchell. "Toward an Architecture for Never-Ending Language Learning." AAAI. Vol. 5. 2010.

[6] Ferrucci, David. "Build Watson: an overview of DeepQA for the Jeopardy! challenge." Proceedings of the 19th international conference on Parallel architectures and compilation techniques. ACM, 2010.

[7] Gerber, Daniel, and A-C. Ngonga Ngomo. "Bootstrapping the linked data web." 1st Workshop on Web Scale Knowledge Extraction@ ISWC. Vol. 2011.

[8] Heath, Tom, and Christian Bizer. "Linked data: Evolving the web into a global data space." Synthesis lectures on the semantic web: theory and technology 1.1, 1-136. 2011.

[9] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. "DBpedia-A crystallization point for the Web of Data." Web Semantics: Science, Services and Agents on the World Wide Web 7.3, 154-165. 2009.

[10] Baker, Collin F., Charles J. Fillmore, and John B. Lowe. "The berkeley framenet project." Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, 1998.

[11] Paul Buitelaar, Philipp Cimiano, J. McCrae, Elena Montiel-Ponsoda, Thierry Declerck. "Ontology lexicalisation: The lemon perspective." 2011.

[12] Younggyun Hahm, Jungyeul Park, Kyungtae Lim, Youngsik Kim, Dosam Hwang, Key-Sun Choi. "Named Entity Corpus Construction using Wikipedia and DBpedia Ontology." In Proceedings of The 9th edition of the Language Resources and Evaluation Conference (LREC), 2014.

[13] Younggyun Hahm, Youngsik Kim, Yousung Won, Jongsung Woo, Jiwoo Seo, Jiseong Kim, Seongbae Park, Dosam Hwang, Key-Sun Choi. "Toward Matching the Relation Instantiation from DBpedia Ontology to Wikipedia text: Fusing FrameNet to Korean." In Proceedings of The SEMANTiCS 2014, 2014.