한국어 위키피디아에 대한 RDF 기반 주석 정보 데이터베이스

원유성○, 함영균, 최기선
한국과학기술원, Semantic Web Research Center
{styner0305, hahmyg, kschoi}@kaist.ac.kr

RDF-based Annotation Database for Korean Wikipedia

Yousung Won○, YoungGyun Hahm, Key-sun Choi
KAIST, Semantic Web Research Center

요 약

본 논문은 한국어 위키피디아 전체 텍스트에 대하여, 최신 언어분석기를 통한 문법적 주석 정보를, 한국어 NLP2RDF 프레임워크를 이용하여 NIF(NLP Interchage Format)[1] 규격을 준수한 RDF(Resource Description Framework) 형태로 구축한 데이터베이스를 소개한다.

1. 서 론

개방된 데이터에 대한 공유 및 연결을 위해, 디지털 상태로 존재하는 방대한 텍스트가 가진 정보를 표현하고 이해하려는 노력이 계속 되어 왔다. 이를 위해서는 텍스트의 문법적인 이해와 의미적인 이해가 동시에 이루어 져야 하고, 이러한 이해 과정을 처리할 수 있는 여러 가지 방법론에 대한 연구가 활발하게 진행되고 있다. 본 논문은 기계가 텍스트를 이해하는 것에 앞서 기계가 텍스트 이해를 위한 문법적, 의미적인 정보의 접근을 가능하게 하는 기반을 제시하는 것을 목적으로 한다,

방대한 주석 정보를 위해서는 신뢰할만한 언어 분석기와 어휘 의미망 등과 같은 다양한 도구 및 언어 자원이 필요하지만, 이들의 결과물 또는 그 형태가 매우 다양해서 일관된 형태로의 변환이 반드시 필요하고, 변환된 데이터의 통합적인 관리와 접근의 용이성이 요구된다[1].

본 논문은 한국어 위키피디아 전체 텍스트에 대하여, 신뢰할만한 언어분석기를 통한 결과물 및 다양한 주석 정보를 한국어 NLP2RDF 프레임워크를 활용하여, NIF 규격을 준수한 RDF로 구축한 데이터베이스를 소개하고 이것의 활용 가치에 대해 알리고자 한다. 2장에서는 NLP2RDF의 국내외 적용 사례를, 3장에서는 한국어 NLP2RDF 프레임워크에 대한 간략한 소개를, 그리고 4장에서는 이를 이용한 한국어 위키피디아 전체 텍스트의 변환 현황을 공유할 것이며, 마지막으로 5장에서는 RDF 기반 주석 정보 데이터베이스의 향후 과제 및 활용 가능성에 대하여 논의하고자 한다.

2. 관련 연구

다양한 형태의 자연 언어 처리 결과물을 RDF(NIF 이용)를 이용하여 표현하려는 시도는 유럽의 LOD2 커뮤니티 과제 중 하나인 NLP2RDF[2] 프로젝트에서 찾아 볼 수 있다[2].

이를 적용한 대표적인 사례로서 국외에서는 Stanford Core-NLP가 형태소 분석, 개체명 인식, 의존구문 분석 등의 여러 자연 언어 처리 도구를 통합하여 주석 정보를 통합하였고, 국내에서는 KAIST에서 한국어 NLP2RDF 프레임워크를 개발하여 한국어 특성에 맞게 NIF 2.0 온톨로지를 보완 적용하는 등 다양한 한국어 자연 언어 처리 도구 및 한국어 언어 자원에 대한 주석 정보를 통합하는 시도를 하였다[3].

3 한국어 NLP2RDF 프레임워크

한국어 자연 언어 처리를 위한 프레임워크는 다음과 같은 구조를 하고 있다.

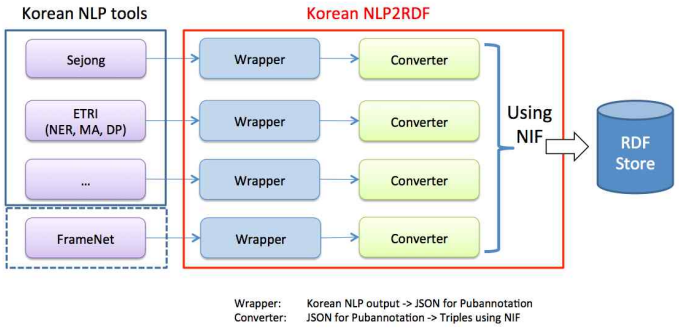

그림 1. 한국어 NLP2RDF 프레임워크 구조

자연 언어 처리 도구는 그 종류 만큼이나 다양한 포맷으로 결과물을 만들어 내는데 그 포맷의 일원화를 위해 NLP2RDF는 그 최종 결과물로 NIF 온톨로지를 적용한 RDF를 생성한다.

한국어 NLP2RDF도 마찬가지로 다양한 자연 언어 처리 결과물을 일관된 포맷으로 통일시킨다는 목적은 동일하지만, 위 그림에서도 볼 수 있듯이 Wrapper와 Converter 두 차례의 가공 과정을 거치는 특징을 가진다.

첫 번째로 Wrapper는 각 언어 분석기를 통해 나온 결과물을 Pubannotation JSON이라는 특정 포맷을 중간 생성물로 만들어 낸다. Pubannotation[3]은 텍스트 및 주석 정보 관리 및 통합을 위한 오픈 소스 프로젝트로서, 기본 데이터 포맷으로 JSON을

1) http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html#d4e952
2) http://aksw.org/Projects/NLP2RDF.html
3) http://www.pubannotation.org/docs/about/

사용하기 때문에, 사용자가 NLP2RDF의 최종 결과물인 RDF 포맷 뿐 만 아니라 기계적 가공이 쉬운 JSON형태의 결과물을 이용할 수 있는 기회를 열어 둔다.

두 번째로 Converter는 앞선 과정의 결과물인 Pubannotation JSON을 입력 데이터로 하여 NIF 규격을 준수한 RDF를 생성하는 역할을 한다. Wrapper라는 중간 처리 과정을 거쳐 가는 Converter는 아직은 한국어의 특성을 모두 반영하기에는 부족한 부분이 있는 NIF 온톨로지의 개선 사항을 유연하게 대응할 수 있는 구조적 특성을 가지고 있다[4].

4. 한국어 위키피디아의 RDF 기반 주석 정보

한국어 NLP2RDF의 입력 자원으로서, 본 논문에서는 Linked Open Data로서 이용 가치가 높은 한국어 위키피디아 전체 텍스트와 이를 위한 자연언어처리 도구로 ETRI 언어분석기를 이용하였다. 입력 데이터의 상세 사항은 다음과 같다.

한국어 위키피디아:
- 2014년 9월 11일자 덤프
- 총 287,963개 문서
- 총 2,758,207개 문장

위의 덤프를 통해 한국어 위키피디아의 전체 텍스트를 추출한 뒤, 이에 대한 ETRI 언어 분석기 처리 결과물을 확보하였다.

그림 2.1 형태소 분석 결과

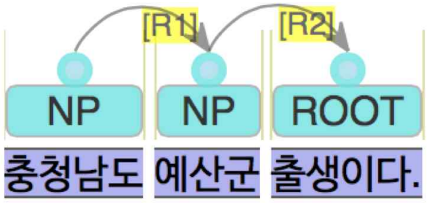

그림 2.2 의존구조 분석 결과

그림 2.3 개체명 인식 분석 결과

그림 2. ETRI 언어분석기 결과
(한국어 NLP2RDF를 통한 Visualization)

그림 2에서 볼 수 있듯이 한국어 위키피디아 전체 텍스트에 대하여 ETRI 언어분석기의 형태소, 의존 구조, 개체명 인식에 대한 주석 정보를 이용 하였고, 이 결과물을 한국어 NLP2RDF의 입력 데이터로 하여 언어 분석 결과에 대한 주석 정보의 Pubannotation JSON(중간 생성물) 및 RDF(NIF 적용, 최종 결과물) 생성을 완료 하였다.

```
KOWIKI:458=88_90#2014-08-05     rdf:type                nif:String
KOWIKI:458=88_90#2014-08-05     rdf:type                nif:Phrase
KOWIKI:458=88_90#2014-08-05     nif:anchorOf            "영화"
KOWIKI:458=88_90#2014-08-05     nif:beginIndex          "88"
KOWIKI:458=88_90#2014-08-05     nif:endIndex            "90"
KOWIKI:458=88_90#2014-08-05     nif:nerTag              "ETRI:FD_ART"
KOWIKI:458=88_90#2014-08-05     nif:phraseTag           "ETRI:NP"
KOWIKI:458=88_90#2014-08-05     nif:dependency          KOWIKI:458=95_99#2014-08-05
KOWIKI:458=88_90#2014-08-05     nif:sentence            KOWIKI:458=44_149#2014-08-05
KOWIKI:458=88_90#2014-08-05     nif:referenceContext    KOWIKI:458=0_298#2014-08-05

...

KOWIKI:458=44_149#2014-08-05     rdf:type               nif:Sentence
KOWIKI:458=44_149#2014-08-05     nif:anchorOf
        "1998년 연극 《강거루군》으로 대학하고 연극무대를 통해 연기경험을 다져왔으며 영화 《왕의 남자》의 원작인 ..."
KOWIKI:458=44_149#2014-08-05     nif:beginIndex         "44"
KOWIKI:458=44_149#2014-08-05     nif:endIndex           "149"

...
```

그림 3. NIF 온톨로지를 적용한 RDF 표현 예

한국어 위키피디아 텍스트의 각 형태소, 어절, 문장, 문서 등은 위 그림 3의 예에서 볼 수 있듯이

<KOWIKI=0=88_90#2014-08-05>

위와 같은 URI를 통해 표현한다. URI는 임의의 개체가 가지는 위치 및 시간 정보(위키피디아 문서 번호, 문서 내 Offset 정보, 문서의 최종 수정 날짜)를 이용한다. 이러한 URI는 각 개체를 대표하고, NIF 2.0 온톨로지의 클래스와 프로퍼티를 통해, 이 개체가 어느 문서에, 어느 문장에, 또는 어느 어절에 해당되는지를 설명하고 각 개체와 개체와의 의존 관계 및 기타 문법적인 주석 정보를 설명한다[5]. 이렇게 생성된 한국어 위키피디아 전체 텍스트에 대한 Triple은 오픈 소스 RDF 저장소인 Virtuoso에 업로드하여, 자연 언어 처리 주석 정보를 통합적으로 관리하고 이용한다.

최종적으로 생성된 한국어 위키피디아 전체 텍스트의 RDF 기반 주석 정보 데이터의 집계는 다음과 같다.

RDF 기반 주석 정보 데이터베이스(한국어 위키피디아):
- 총 287,963개 Context 클래스 (문서 단위)
- 총 2,758,207개 Sentence 클래스 (문장 단위)
- 총 36,091,298개 Phrase 클래스 (어절, 개체명 단위)
- 총 73,550,613개 Word 클래스 (형태소, 개체명 단위)
- 총 1,060,743,087개 Triple 생성

(*) Phrase 클래스에 해당하는 개체의 수는 언어분석기가 의존 관계를 가진다고 간주하거나 개체명으로 인식 된 어절의 개수를 의미
(*) Word 클래스에 해당하는 것은 언어분석기가 형태소 또는 개체명으로 인식하는 개체의 수를 의미
(*) 위의 수치는 NIF 2.0 온토롤지 중에서 주석정보를 표현하기 위해 기본적으로 필요한 클래스와 프로퍼티를 이용
(*) 적용한 주요 클래스 및 프로퍼티 예

클래스
nif:Context, nif:Sentence, nif:Phrase, nif:Word, nif:String
프로퍼티
nif:referenceContex, nif:sentence, nif:phrase, nif:word, nif:subString
nif:posTag, nif:nerTag, nif:phraseTag, nif:anchorOf, nif:beginIndex, nif:endIndex

한국어 NLP2RDF 프레임워크를 이용하여 언어 분석 결과물을 RDF 포맷으로 통일시킨 것은 다음의 장점을 가진다.

현실적으로 자연 언어 처리에 소요되는 과정이 상당히 복잡한데 반해 그 작업이 일회성에 그치는 경우가 많다. 예를 들면 누군가 자연 언어 처리 도구를 통해 방대한 텍스트에 대한 주석 정보를 생성해 내었을 때, 불특정 다수가 그 데이터에 접근하여 이용하는 것은, 사실상 불가능하거나 또는 상당히 많은 자원을 소모하게 되는 것이 현실이다. 기계적으로 생성된 언어 자원 뿐만 아니라 현재는 사람이 직접 많은 시간과 돈을 들여 수동적으로 주석 작업을 하여 생성된 양질의 언어 자원도 존재하고 있지만 이것을 이용하는 것에도 똑같은 한계가 있다.

특히 한국어 언어 자원은 영어나 기타 외국어에 비해 가용한

자원이 많이 부족한 실정이기 때문에 다양한 한국어 언어 자원 정보를 일회성으로 생산하고 소비하는 것 보다는 이에 대한 통합을 통해 자원을 효율적으로 이용하는 방안이 반드시 필요하다. 이것이 정제되지 않고 흩어져 있는 한국어 언어 분석 정보를 W3C 시멘틱웹 권장 포맷인 RDF로 생성하여 이용하는 당위성이라고 할 수 있다.

RDF로 접근 가능한 방대한 텍스트(한국어 위키피디아)의 주석 정보는 SPARQL 쿼리에 대한 기본적인 지식이 있는 사람이면 누구나 접근 및 가공이 용이하다.

다음은 한국어 위키피디아의 RDF 기반 데이터베이스[4]에서 사용 할 수 있는 SPARQL 쿼리의 예시 이다.

```
PREFIX nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>

select ?string ?stringLabel ?stringTag
where {
?sentence nif:anchorOf "1976년 스티브 워즈니악, 로널드 웨인과 함께 애플을 공동 창업하고, 애플 2를 통해 개인용 컴퓨터를 대중화했다." .
?string nif:sentence ?sentence .
?string nif:anchorOf ?stringLabel .
?string nif:nerTag ?stringTag .
}
```

string	stringLabel	stringTag
http://ko.wikipedia.org/wiki?curid=34415=105_111#2014-09-10	"로널드 웨인"	"ETRI:PS_NAME"
http://ko.wikipedia.org/wiki?curid=34415=116_118#2014-09-10	"애플"	"ETRI:OGG_BUSINESS"
http://ko.wikipedia.org/wiki?curid=34415=132_133#2014-09-10	"2"	"ETRI:QT_ORDER"
http://ko.wikipedia.org/wiki?curid=34415=89_94#2014-09-10	"1976년"	"ETRI:DT_YEAR"
http://ko.wikipedia.org/wiki?curid=34415=95_103#2014-09-10	"스티브 워즈니악"	"ETRI:PS_NAME"

그림 4. RDF 기반 주석 정보 데이터베이스 이용 예시

위 그림 4의 쿼리는 한국어 위키피디아에 있는 실제 문장인 “1976년 스티브 워즈니악, 로널드 웨인과 함께 애플을 공동 창업하고, 애플 2를 통해 개인용 컴퓨터를 대중화했다.” 에서 언어 분석기가 찾아 낸 개체명을 조회하는 기능을 한다.

	구문 Tag	Frequency	Propotion
1	NP	22,091,346	68.05%
2	VP	7,471,971	23.02%
3	VNP	1,152,122	3.55%
4	AP	1,002,837	3.09%
5	DP	622,687	1.92%
6	X	112,920	0.35%
7	IP	4,851	0.01%
8	R	2,148	0.01%
9	L	1,227	0.00%
10	Q	1,042	0.00%
	TOTAL	32,463,151	

그림 5.1 형태소 별 주석 빈도

	POS Tag	Frequency	Propotion
1	NNG	20,729,545	27.63%
2	JKB	3,856,605	5.14%
3	EC	3,416,267	4.55%
4	VV	3,373,973	4.50%
5	NNP	3,322,523	4.43%
6	NNB	3,077,836	4.10%
7	SN	3,062,031	4.08%
8	ETM	2,790,010	3.72%
9	SS	2,767,237	3.69%
10	JX	2,614,528	3.49%
	TOTAL	75,014,811	

그림 5.2 구문 별 주석 빈도

	Dependency Relation		Frequency	Propotion
1	NP	NP	10,423,194	35.02%
2	NP	VP	9,729,384	32.69%
3	VP	VP	3,119,191	10.48%
4	VP	NP	2,135,708	7.18%
5	NP	VNP	1,328,968	4.47%
6	AP	VP	794,352	2.67%
7	DP	NP	600,628	2.02%
8	VP	VNP	436,128	1.47%
9	VNP	NP	370,676	1.25%
10	VNP	VP	211,999	0.71%
	TOTAL		29,760,751	

그림 5.3 의존관계 패턴 별 주석 빈도

	NER Tag	Frequency	Propotion
1	PS_NAME	1,117,900	13.46%
2	CV_POSITION	607,058	7.31%
3	DT_YEAR	593,411	7.15%
4	LCP_COUNTRY	556,767	6.71%
5	DT_OTHERS	341,768	4.12%
6	QT_COUNT	298,758	3.60%
7	CV_OCCUPATION	252,841	3.05%
8	QT_ORDER	245,934	2.96%
9	LCP_CITY	185,493	2.23%
10	OGG_BUSINESS	184,054	2.22%
	TOTAL	8,302,638	

그림 5.1 개체명 별 주석 빈도

그림 5. 한국어 위키피디아 주석 정보 집계 표
(주석 정보 빈도에 따른 상위 10개 결과)

5. RDF 기반 주석 정보 데이터베이스의 향후 과제 및 활용 가능성

한국어 NLP2RDF 프레임워크는 한국어 위키피디아 전체 텍스트에 대한 주석 정보로 부터 약 10억개 정도의 Triple을 생성하여, RDF 기반의 주석 정보 데이터베이스 구축을 가능하게 하였다. 한국어 위키피디아는, 보유하고 있는 양질의 텍스트 정보 자체로서의 자원일 뿐만 아니라 디비피디아(DBpedia)를 통하여 위키피디아의 리소스(Resource)들 간의 관계 및 속성을 통해 구조화된 정보를 제공해주는 원천이기도 하다. 디비피디아는 그것이 보유하고 있는 RDF 만으로도 상당히 많은 지식을 Linked Data로서 활용이 가능하다. 현재 한국어 NLP2RDF를 통한 한국어 위키피디아의 주석 정보는 위키피디아 텍스트의 문법적인(Syntatic) 요소에 대한 대량의 메타데이터를 제공할 수 있다는 측면과 디비피디아 자체가 잠재적으로 제공해 줄 수 있는 의미적인(Semantic) 정보, 그리고 여기에 WordNet과 같은 어휘 의미망과의 맵핑을 통해 디비피디아가 보유하고 있는 또는 향후 보유할 방대한 의미 정보와의 연계가 이루어 질 수 있다는 가능성을 열어두고 있다.

가까운 계획으로서, 위키피디아 텍스트의 Entity Linking을 가능하게 하는 한국어판 DBpedia Spotlight[5][6]와의 연계를 통하여 각 텍스트의 개체가 가질 수 있는 URI정보를 RDF로 표현하여 데이터베이스에 추가하는 것이 향후 과제이다. 이를 통해 NIF를 이용한 RDF 기반 주석 정보를 이용하여 가공되지 않은 텍스트로부터 RDF의 Population에 활용될 방안도 연구 과제로 남아 있다.

RDF 기반 데이터베이스의 잠재성은 RDF로 표현된 그래프의 활용에 있다. 디비피디아나 프리베이스(Freebase)와 같이 Linked Data로 표현된 거대한 구조화된 정보와 더불어 자연 언어의 문법적 의미적 요소의 결합, 그리고 개체간의 연결을 통한 그래프의 확장은 텍스트의 이해를 보다 향상 시키는 역할을 할 것이다. 이것이 자연 언어의 이해의 고도화에 앞서 주석 정보의 RDF 기반 데이터베이스로 구축하는 의의라고 할 수 있다.

4) http://143.248.135.60:8501/sparql

사사

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

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발]

참고 문헌

[1] Sebastian Hellmann, Jens Lehmann, Sören Auer, "NIF: An ontology-based and linked-data-aware NLP Interchange Format", 2012

[2] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer1, Integrating NLP using Linked Data, ISWC, 2013

[3] 원유성, 서지우, 김정욱, 함영균, 최기선, "한국어 NLP2RDF 프레임워크", 제26회 한글 및 한국어 정보처리 학술대회, 2014

[4] 서지우, 원유성, 김정욱, 함영균, 최기선, "한국어 자연언어처리의 NIF 적용에 관한 연구", 제26회 한글 및 한국어 정보처리 학술대회, 2014

[5] 김영식, 함영균, 김지성, 최기선 “한국어 텍스트의 개체 URI 탐지: 품사 태깅 독립적 개체명 인식과 중의성 해소” 제26회 한글 및 한국어 정보처리 학술대회, 2014

[6] 김지성, 김영식, 함영균, 최기선 “URI 중의성 해소 및 오류 감소를 위한 LDA 기반 접근법” 제26회 한글 및 한국어 정보처리 학술대회, 2014