대용량 OWL 온톨로지 자동구축을 위한세종전자사전 활용 방법론 연구

송도규*†

㈜씨컴테크

Do Gyu Song. 2005. A Study of Methodology for Automatic Construction of OWL Ontologies from Sejong Electronic Dictionary. *Language and Information 9.1*, 19–34. Ontology is an indispensable component in intelligent and semantic processing of knowledge and information, such as in semantic web. However, ontology construction requires vast amount of data collection and arduous efforts in processing these un-structured data. This study proposed a methodology to automatically construct and generate ontologies from Sejong Electronic Dictionary. As Sejong Electronic Dictionary is structured in XML format, it can be processed automatically by computer programmed tools into an OWL(Web Ontology Language)-based ontologies as specified in W3C. This paper presents the process and concrete application of this methodology. **(C-Come Tech Co., Ltd.)**

Key words: 온톨로지(ontology), 시맨틱 웹(Semantic Web), OWL(Web Ontology Language), RDF(Resource Description Framework), 세종전자사전(Sejong Electronic Dictionary), 언어지식공학(Language and Knowledge Engineering), 지능형 의미기반 지식/정보 처리(Intelligent and Semantic Processing of Knowledge and Information)

1. 머리말

컴퓨터에 의한 지능형 의미기반 지식/정보의 자동처리를 위해서는 온톨로지 구축과 활용의 중요성이 널리 공감되고 있다. 그러나 막상 온톨로지를 실제로 구축하려면 방대한 양의 자료 수집과 많은 숙련된 인력에 의한 장시간의 작업이 필요하다. 그것은 적합한 자료를 선별, 수집하는 데에도 많은 인력과 시간이 소요될 뿐만 아니라 이렇게 어렵게 모은 자료의 대부분이 무정형의 데이터이고 이 무정형의 데이터로부터 정보를 자

* 152-050 서울시 구로구 구로동 에이스테크노타워 5차 703호 (주)씨컴테크, E-mail: winwin@c-come.co.kr

† 귀중하고 적절한 지적을 해 주신 심사위원들께 감사 드립니다. 그리고 지적을 충분히 반영하지 못한 책임은 필자에게 있음을 분명히 밝힙니다.

동으로 추출하여 온톨로지를 구축하기가 매우 어렵기 때문이다. 이런 이유로 이미 구축되어 있는 정형의 데이터로부터 최종 온톨로지는 아니라 할지라도 사람이 수정, 보완만 하면 되는 온톨로지 초안을 자동으로 빠르게 또 대량으로 구축할 수 있다면 그 효용 가치는 매우 크다 하겠다. 본 논문에서는 기존에 시도된 바 없는 대용량 정형 사전으로부터 OWL 온톨로지를 자동으로 구축하는 방법론을 제시하고자 한다. 이 방법론은 자동구축을 위한 프로그램의 알고리듬으로 그대로 활용할 수 있고 추후 충분한 검증을 거쳐 실제의 프로그램으로 구현될 수 있다. 본 논문에서는 정형의 데이터로 세종전자사전을 활용했다. 세종전자사전은 정형의 SGML 그 중에서 주로 XML 포맷으로 되어 있어 컴퓨터에 의한 자동처리가 용이하다. 여기에서는 다른 품사의 표제어들도 온톨로지 구축의 대상이 되겠지만 논의의 편의를 위해 체언과 용언만 다루기로 한다. 온톨로지는 World Wide Web Consortium(http://www.w3.org, 이하 W3C)에서 차세대 인터넷 표준으로 제정하고 있는 OWL(Web Ontology Language) 포맷을 채택하기로 한다.

2. 시맨틱 웹(Semantic Web)과 OWL(Web Ontology Language)

시맨틱 웹은 현재의 인터넷인 월드와이드웹의 한계를 극복하고자 Tim Berners-Lee 등이 제안하고 W3C에서 국제 표준으로 제정하고 있는 차세대 인터넷 표준이다. 현재의 인터넷은 무정형의 HTML(HyperText Markup Language)을 웹 문서 포맷으로 하고 있으며 이 포맷들은 파싱(parsing)만 가능해 패턴 매칭(pattern matching)에 의한 형태적인 처리만이 가능하다. 다시 말하면 인터넷 검색엔진이 검색어에 해당하는 문서들을 찾아 준다 하여도 그것은 검색어의 의미를 이해하지 못한 채 검색어의 형태만을 패턴 매칭하여 해당 형태의 검색어가 있는 문서의 리스트를 보여 주는 것에 불과하다. 이런 이유로 검색어로 "맛있는 배"라고 입력해도 먹는 배(梨) 뿐만 아니라 신체 부위인 배(腹), 타는 배(船), 곱절 배(倍) 심지어 "색이 배다"의 "배"가 들어 있는 문서도 모두 찾아 제시한다. "맛있는"이라는 어휘의 의미를 활용하여 "배"의 의미를 한정할 수 없기 때문이다. 이에 컴퓨터로 하여금 웹 상의 지식과 정보를 의미적으로 이해하게 해야 한다는 필요가 세계적으로 공감되게 되었다. 이런 이유로 W3C에서는 컴퓨터가 인터넷 상의 지식과 정보를 스스로 이해하고 추론, 판단까지 하는 새로운 패러다임을 추진하게 되었는데 이것이 바로 시맨틱 웹(Semantic Web)이다. 시맨틱 웹은 특정한 지식/정보 표현 포맷인 RDF(Resource Description Framework)와 일반적인 온톨로지 표현 포맷인 OWL(Web Ontology Language)로 구성된다. 온톨로지는 원래 철학에서 '존재론'을 의미하나 언어지식공학에서는 '사람이 가지고 있는 지식을 컴퓨터가 이해하기 쉽게 체계화한 것'을 의미한다.[1] 시맨틱 웹은 RDF와 OWL 포맷으로 구축된 지

[1] 보다 학문적인 정의는 Thomas Gruber (1993)가 자주 인용되는데 내용은 "An ontology is an explicit specification of a conceptualization.", Gruber (1993, p. 1), http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html.

식/정보를 사람의 개입없이 컴퓨터가 스스로 분석, 이해하고 추론, 판단까지 수행하여 사람이 원하는 최종 결과만을 제공해 줄 수 있는 차세대 인터넷 표준이다. 이것이 실현되면 지능형 의미기반 인터넷은 물론 기업, 관공서, 학교, 은행 등의 방대한 지식과 정보에 대한 자동이해에 기반한 자동처리가 가능해진다.

본고에서는 OWL 포맷의 온톨로지 구축을 목표로 한다. OWL은 Class와 Property로 구성되는데, Class는 Class-Property-Value의 트리플로 Property는 Property-Axiom-Value의 트리플로 각각 구성된다. 그림 1은 XML 형태로 표현된 OWL 문서의 한 예[2]이다.

```
<?xml version="1.0"?>
<!ENTITY vin "http://www.w3.org/TR/2003/CR-owl-guide-20030818/wine#" >
<!DOCTYPE rdf:RDF [
<!ENTITY food "http://www.w3.org/TR/2003/CR-owl-guide-20030818/food#" >
<!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
]>
<rdf:RDF
xmlns:food= "http://www.w3.org/TR/2003/CR-owl-guide-20030818/food#"
xmlns:vin = "http://www.w3.org/TR/2003/CR-owl-guide-20030818/wine#"
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#">

<owl:Class rdf:ID="ConsumableThing" />

<owl:Class rdf:ID="NonConsumableThing">
  <owl:complementOf rdf:resource="#ConsumableThing" />
</owl:Class>

<owl:Class rdf:ID="EdibleThing">
  <rdfs:subClassOf rdf:resource="#ConsumableThing" />
</owl:Class>

<owl:Class rdf:ID="PotableLiquid">
  <rdfs:subClassOf rdf:resource="#ConsumableThing" />
  <owl:disjointWith rdf:resource="#EdibleThing" />
</owl:Class>
```

[그림 1] food.owl의 일부분

OWL Class는 다음과 같은 Property를 가지는데, 각각의 의미와 설명은 다음 표 1과 같다.

OWL Property는 다음과 같은 Axiom이 있고 각각의 의미와 설명은 다음 표 2와 같다.

[2] 미국 스탠포드대학교 Knowledge Systems, AI Laboratory(http://ksl-web. stanford.edu)에서 만든 음식에 관한 영어 온톨로지 food.owl.

OWL Class의 Property	의 미	설 명
owl:complementOf	Complement	예) '남자'와 '여자' 간의 관계
owl:disjointWith	Disjunction	예) '해군'과 '공군' 간의 관계
owl:equivalentClass	Equivalence	동치 관계
owl:intersectionOf	Intersection	교집합
owl:oneOf	One among several classes	여러 클래스 중의 하나
rdfs:subClassOf	Subset	진부분집합, 예) '동물'에 대한 '사자'
owl:unionOf	Union	합집합

[표 1] OWL Class의 Property

OWL Property의 Axiom	의 미	설 명
rdfs:domain	Domain	정의역
owl:equivalentProperty	Equivalence	동치 관계
owl:FunctionalProperty	Only one value	오직 하나의 값만 가짐
owl:InverseFunctionalProperty	Unique domain	오직 하나의 정의역만 가짐
owl:inverseOf	Inverse	역(逆), 예) '팔다'와 '사다'의 관계
rdfs:range	Range	치역
rdfs:subPropertyOf	Subordinate property	하위 프라퍼티, 예) '행동하다'에 대한 '걷다'의 관계
owl:SymmetricProperty	Symmetry	대칭성, 예) 친구사이의 '친구이다' 관계 A ↔ B
owl:TransitiveProperty	Transitivity	전이성, 예) '~보다 크다'의 관계, A > B, B > C 이면 A > C

[표 2] OWL Property의 Axiom

OWL 온톨로지는 이상의 정해진 표준 키워드를 사용하여 온톨로지 구축을 위한 기제를 제공한다. 이를 기반으로 세종전자사전으로부터 OWL 온톨로지의 자동구축을 위한 몇가지 방법론을 다음 장에서 제안한다.

3. 세종전자사전을 활용한 OWL 온톨로지 자동구축

세종전자사전은 문화관광부에서 주관하여 1998년부터 10년간 3단계의 연구 과정으로 수행되고 있는 컴퓨터 가독형의 한국어 전자사전 개발 프로젝트의 결과물이다.[3] 세종전

[3] 본 논문에서 사용한 세종전자사전의 체계와 표기법은 2001년 11월에 발간된 '21세기 세종계획 전자사전

자사전은 표제어에 대한 풍부한 정보가 정형의 SGML 그 중에서 특히 XML 포맷으로 되어 있어 컴퓨터에 의한 분석과 자동처리가 용이하고 자동 툴의 개발과 사용이 가능하다. 2001년까지의 결과물은 단일어 명사와 복합어 명사를 합한 체언이 8만 표제어, 용언이 3만 표제어이다. 이들 표제어를 시맨틱 웹 표준 온톨로지 포맷인 OWL로 구축하는 데 있어서 먼저 대분류인 Class와 Property로 나누어서 살펴 보려 한다.

3.1 Class

Class는 체언으로 분류한 표제어들에 해당하고 Property는 대개 용언으로 분류한 표제어에 해당한다. 먼저 OWL의 Class를 논의한다. 세종전자사전에 작성된 체언의 실례를 '음악'이라는 표제어로 보면 다음 그림 2와 같다.[4]

여기에서 주목해야 할 부분은 의미 구획('<sense>'와 '</sense>'로 싸인 부분, 다의어와 동음이의어의 경우에 '<sense' 태그 뒤에 'n=1', 'n=2' 식으로 일련번호를 붙여 구분한다)과 어휘관계 구획('<lr>'과 '</lr>'로 싸인 부분)이며 의미 구획 안의 전문분야('@domain'의 값)와 의미하위부류('@sem'의 값) 그리고 어휘관계 구획 안의 동의어('@syn'의 값), 반의어('@ant'의 값), 상위어('@hyper'의 값), 하위어('@hypo'의 값), 동위어('@coord'의 값), 부분어('@mero'의 값) 그리고 관련어('@rel'의 값)의 데이터를 주로 활용할 것이다.

1. 전문분야 분류

   (a) 각 표제어가 '@domain'의 값으로 전문분야[5](domain)가 분류되어 각 전문분야별로 온톨로지를 구축하기가 용이하다. 특정한 전문분야의 온톨로지를 구축하려면 먼저 해당 전문분야의 표제어들을 선별, 취합한 후 온톨로지를 구축한다.

   (b) @form=[개체] @domain=[생물]
   @form=[기소] @domain=[법률]
   @form=[노출] @domain=[사진]
   위의 예들에서 '개체'는 '생물'의 전문분야로, '기소'는 '법률'의 전문분야로, 또 '노출'은 '사진'의 전문분야로 분류하여 온톨로지를 구축한다. 2001년 결과물까지는 전문분야의 분류가 충실히 되어 있지 않으나 21세기 세종계획 사업이 진행되면서 전문분야의 분류가 정확하고 충분하게 이루어질 것으로 기대한다.

개발분과 연구보고서'에 근거한다. (21세기 세종계획 홈페이지 http://www.sejong.or.kr)

[4] 내용이 많아 본 논의에 불필요한 일부 데이터는 글쓴이 임의로 삭제하였다.

[5] 세종전자사전은 다음과 같은 전문분야로 표제어들이 분류되어 있다.
'가톨릭, 건설, 경제, 고적, 고유명사, 공업, 광업, 교육, 교통, 군사, 기계, 기독교, 논리, 농업, 문학, 물리, 미술, 민속, 법률, 불교, 사회, 생물, 수학, 수산, 수공, 식물, 심리, 약학, 언론, 언어, 역사, 연영, 예술, 운동·음악, 인명, 정치, 지리, 지명, 책명, 천문, 철학, 출판, 통신, 컴퓨터, 한의학, 항공, 해양, 화학'

```
<superEntry>
●음악
<entry n=1>%dic=[체언/단일어]
    <toplevel> @form=[음악]@pos=[nn]@see=[]
      <morph_a> @var[xs=; xd=; xx=]@abb=[]@lng=[]@str=[]@org=[si(音樂)]@symb=[] </morph_a>
      <morph_b> @hom[]@der=[(n)~가|인|제|회|관|당|사|성|실|적|학;(v);(a);(av)]@comp=[(n)~감상
       실|교육|대학|선생|교사사|애호가|평론가|이론;(v);(a);(av)]@metc=[(n);(v);(a);(av)]@img=[xs={};
       xd={}; xx={}; xp={}] </morph_b>
      <froz> @idnp=[] @idna=[] @idnv=[] @idda=[] @prv=[] @idetc=[] </froz> </toplevel>
    <sense n=1> @eg=[그녀는 ~을 전공하기 위해 예술학교에 진학을 할 예정이다]
       @trans=[music]@domain=[]@reg=[]@con=[]@curs=[U]@sem=[교과목/학문]@nm_sub=[]
       @cl_sub=[] @np_sub=[sem=;num=;ref=]@rel_n=[]
      <lr> @syn=[]@ant=[]@hyper=[학문|교과목]@hypo=[서양음악|아악|국악|고전음악|전통음악|실내
       악|교]@coord=[미술|문학|체육]@holo=[]@mero=[]@rel=[예술|교육|학문] </lr>
      <synt_a> @cl=[uni=;grp=;div=;qnt=]@prt=[]@av=[]@ds=[] </synt_a>
      <synt_b> @comb_aj=[~이 어렵다|쉽다|인기있다] @magn=[] @comb_v=[~을 배우다|공부하다|전
       공하다|가르치다|전수하다] @comb_ida=[] @comb_n=[] @supv=[X-가 ~을 하다; cor_v={}]
       @max_n=[] @sel_res=[X=인물] </synt_b>
      <synt_c> @pre_d=[]@pre_n=[]@pre_s=[] </synt_c>
      <synt_d> @flt=[]@carord=[car=;ord=]@etc=[] @etc=[] </synt_d> </sense>
    <sense n=2> @eg=[그 ~은 부드러운 선율로 사람들의 마음을 사로잡는다]
       @trans=[music]@domain=[]@reg=[]@con=[]@curs=[C]@sem=[예술품(소리)]@nm_sub=[]
       @cl_sub=[] @np_sub=[sem=;num=;ref=]@rel_n=[]
      <lr> @syn=[곡|노래]@ant=[]@hyper=[악곡]@hypo=[실내악|피아노 음악|교향악|독주곡]
       @coord= [시|소설|춤|그림]@holo=[]@mero=[]@rel=[음악회|연주|청중|애호가] </lr>
      <synt_a> @cl=[uni=곡|노래;grp=;div=;qnt=]@prt=[]@av=[]@ds=[] </synt_a>
      <synt_b> @comb_aj=[~이 부드럽다|강하다|역동적이다|감동적이다|매혹적이다] @magn=[]
       @comb_v=[~을 듣다|작곡하다|편곡하다|연주하다] @comb_ida=[] @comb_n=[] @supv=[;
       cor_v={}] @max_n=[] @sel_res=[] </synt_b>
      <synt_c> @pre_d=[]@pre_n=[]@pre_s=[] </synt_c>
      <synt_d> @flt=[]@carord=[car=;ord=]@etc=[] @etc=[] </synt_d> </sense>
 </entry>
</superEntry>
```

[그림 2] 세종전자사전 표제어 '음악'의 예

(c) 관련어(related terms)를 망라하여 전문분야 온톨로지를 구축할 수 있다.

@form=[위반] @rel=[범칙금|법률|질서|규칙|규약]
@form=[유랑] @rel=[집시|떠돌이]
@form=[위협] @rel=[관계|분쟁|해결책|타협]
@form=[위장] @rel=[의학|신체부위]
@form=[위장] @rel=[은폐|피신|범죄|범인수배]

위의 표제어 '위반'의 경우 관련어로 기입되어 있는 '범칙금', '법률', '질서', '규칙', '규약'까지도 동일한 분야의 온톨로지에 속하는 것으로 간주하여 세종전자사전에서 각각의 표제어를 찾아 해당 정보를 해당 전문분야 온톨로지로 구축한다. 나머지 예도 같은 방식으로 각각의 전문분야 온톨로지를 구축한다.

2. 각 표제어가 체계적으로 분류된 의미부류[6]에 따라 하위 분류되어 있어 'rdfs:subClassOf'[7] 관계를 자동구축한다.

(a) 의미하위부류('@sem'의 값) 정보를 이용한다.

i. @form=[음악] @sem=[예술품(소리)]

⇓

```
<owl:Class rdf:ID="음악">
    <rdfs:subClassOf rdf:resource="#예술품(소리)">
</owl:Class>
```

(b) 의미하위부류에 위계가 있는 경우에는 이의 온톨로지를 순차적으로 구축한다.

i. @form=[어머니] @sem=[인물/친족][8]

⇓

```
<owl:Class rdf:ID="어머니">
    <rdfs:subClassOf rdf:resource="#친족">
</owl:Class>
```

[6] 세종전자사전의 2001년의 연구 결과로 구축된 의미부류의 총 개수는 1층위부류(최상위부류) 9개, 2층위부류 83개, 3층위부류 142개, 4층위부류 59개, 5층위부류 21개, 6층위부류 12개를 포함하여 326개이다.

[7] 'rdfs:'는 'rdfs'라는 namespace에 해당하는 문서 'http://www.w3.org/2000/01/rdf-schema'에 정의된 내용을 참조한다는 의미이다. 같은 방법으로 'rdf:'는 'rdf'라는 namespace에 해당하는 문서 'http://www.w3.org/1999/02/22-rdf-syntax-ns', 'owl:'은 'owl'이라는 namespace에 해당하는 문서 'http://www.w3.org/2002/07/owl'에 정의된 내용을 참조한다.

[8] '인물/친족'처럼 '/'으로 구분된 경우는 위계를 나타내며 '도서|서적'과 같이 '|'로 구분된 경우에는 단순한 나열이다.

```
<owl:Class rdf:ID="친족">
    <rdfs:subClassOf rdf:resource="#인물">
</owl:Class>
```

ii. @form=[부엉이] @sem=[동물/새]

⇓

```
<owl:Class rdf:ID="부엉이">
    <rdfs:subClassOf rdf:resource="#새">
</owl:Class>
<owl:Class rdf:ID="새">
    <rdfs:subClassOf rdf:resource="#동물">
</owl:Class>
```

3. 동의어(synonym), 반의어(antonym), 상위어(hyperonym), 하위어(hyponym) 동위어(coordinate terms), 부분어(meronym) 데이터에서 동의어는 'equivalentClass', 반의어는 'complementOf', 상위어와 하위어는 'subClassOf', 동위어는 'disjointWith', 부분어는 'unionOf' 관계로 활용한다.

(a) 동의어 (synonym)

i. 동의어 '@syn'의 값을 활용한다.

@form=[죽음] @syn=[사망]

⇓

```
<owl:Class rdf:ID="죽음">
    <owl:equivalentClass rdf:about="#사망">
</owl:Class>
```

ii. 동의어의 값이 위계가 아닌 나열인 경우에는 이의 온톨로지를 반복적으로 구축한다.

@form=[책] @syn=[도서|서적]

⇓

```
<owl:Class rdf:ID="책">
    <owl:equivalentClass rdf:about="#도서">
</owl:Class>
<owl:Class rdf:ID="도서">
    <owl:equivalentClass rdf:about="#서적">
</owl:Class>
```

이 경우에 시맨틱 웹의 추론 기능에 의해서 [책 owl:equivalentClass 서적]도 자동으로 추론된다.

(b) 반의어 (antonym)

i. 반의어 '@ant'의 값을 활용한다.

@form=[죽음] @ant=[삶]

⇓

```
<owl:Class rdf:ID="죽음">
    <owl:complementOf rdf:about="#삶">
</owl:Class>
```

(c) 상위어와 하위어 (hypernym & hyponym)

i. 상위어는 표제어와 상위어 간의 'rdfs:subClassOf' 관계를 제시한다.

@form=[근시] @hyper=[시력]

⇓

```
<owl:Class rdf:ID="근시">
    <rdfs:subClassOf rdf:about="#시력">
</owl:Class>
```

ii. 하위어도 하위어와 표제어 간의 'rdfs:subClassOf' 관계를 나타낸다. 그러나 상위어의 경우와는 포섭 방향이 반대이다.

@form=[길] @hypo[도로|보도|인도|차도]

⇓

```
<owl:Class rdf:ID="도로">
    <rdfs:subClassOf rdf:about="#길">
</owl:Class>
<owl:Class rdf:ID="보도">
    <rdfs:subClassOf rdf:about="#길">
</owl:Class>
<owl:Class rdf:ID="인도">
    <rdfs:subClassOf rdf:about="#길">
</owl:Class>
<owl:Class rdf:ID="차도">
    <rdfs:subClassOf rdf:about="#길">
</owl:Class>
```

(d) 동위어 (coordinate terms)

i. 표제어와 동위어의 관계가 자칫 'equivalentClass'로 추출되기 쉬우나 대부분의 경우 'disjointWith'로 짝 지어지는 것이 적합하다.

@form=[근시] @coord=[난시|약시|원시]

⇓

```
<owl:Class rdf:ID="근시">
    <owl:disjointWith rdf:about="#난시">
    <owl:disjointWith rdf:about="#약시">
    <owl:disjointWith rdf:about="#원시">
</owl:Class>
```

(e) 부분어 (meronym)

부분어는 다음 예들에서 보는 바와 같이 표제어와 'unionOf' 관계가 설정된다.

@form=[꽃] @mero=[꽃잎|꽃받침|꽃술|잎사귀|줄기|향기]

⇓

```
<owl:Class rdf:ID="꽃">
    <owl:unionOf rdf:parseType=Collection>
        <owl:Class rdf:about="#꽃잎">
        <owl:Class rdf:about="#꽃받침">
        <owl:Class rdf:about="#꽃술">
        <owl:Class rdf:about="#잎사귀">
        <owl:Class rdf:about="#줄기">
        <owl:Class rdf:about="#향기">
    </owl:unionOf>
</owl:Class>
```

3.2 Property

세종전자사전에서 용언의 실례를 '언도하다'라는 표제어로 보면 다음 그림 3과 같다.[9]

여기에서 OWL 온톨로지 구축에 직접 활용할 부분은 통사정보그룹의 통사/의미 관계(<synSem>과 </synSem>로 싸인 부분, 다의어와 동음이의어의 경우에는 여러 개가 있다)이며 그 안의 선택제약정보(<selRst>과 </selRst>로 싸인 부분)와 의미역 정보(<thtRol>과 </thtRol>로 싸인 부분) 그리고 의미정보그룹(<semGrp>와 </semGrp>로 싸인 부분)의 의미하위부류정보(<semClass>와 </semClass>로 싸인 부분)와 전문영역(<domain>과 </domain>으로 싸인 부분) 그리고 의미관계정보(<semRel>과 </semRel>로 싸인 부분)의 동의어(<syn>과 </syn>으로 싸인 부분), 반의어(<ant>와 </ant>로 싸인 부분), 양태상위어(<trohpr>와 </trohpr>로 싸인 부분), 양태하위어(<trohpo>와 </trohpo>로 싸인 부분), 양태연쇄어(<troser>와 </troser>로 싸인 부분)를 주로 활용하게 된다.

용언은 대체로 'Property'(물론 후에 'ObjectProperty', 'DatatypeProperty' 등으로 세분해야 한다)에 해당한다.[10]

---

[9] 내용이 많아 본 논의에 불필요한 일부 데이터는 글쓴이 임의로 생략하였다.

[10] 시맨틱 웹의 RDF(Resource Description Framework) 포맷에서 하나로 다루던 'rdf:Property'가 OWL에서는 'ObjectProperty', 'DatatypeProperty', 'FunctionalProperty', 'InverseFunctionalProperty', 'SymmetricProperty' 그리고 'TransitiveProperty'로 세분되었다.

```
<superEntry>
  <entry>
    <headGrp>
      <orth>언도하다</orth>
      <org>si+ ko(言渡+ 하)</org>
      <var type=X></var>
    </headGrp>
    <synGrp>
      <frmClass type=FTR>
        <caseFrame>
          <frame>N0-이 N1-에게 N2-을 V</frame>
          <synSem>
            <selRst>N0=인물(관사)|단체(법원|가정법원) N1=인물 N2=(사형|형|징역)</selRst>
            <argRst></argRst>
            <ordRst></ordRst>
            <thtRol>N0=DON N1=GOL N2=THM</thtRol>
            <synPtr>idval=01</synPtr>
            <conCor type=X></conCor>
            <vaAlt type=itr:con>언도받다</vaAlt>
            <eg>법원은 그에게 국가보안법 위반으로 사형을 언도하였다.</eg>
          </synSem>
        </caseFrame>
      </frmClass>
    </synGrp>
    <semGrp>
      <sem idval=01>
        <trans>pronounce a sentence</trans>
        <semClass></semClass>
        <domain>법률</domain>
        <semDef>재판의 판결을 일반 사람들에게 알리다</semDef>
        <reg></reg>
        <semRel>
          <syn>선고하다</syn>
          <ant>언도받다</ant>
          <trohpr>판결하다</trohpr>
          <trohpo></trohpo>
          <troser></troser>
        </semRel>
      </sem>
    </semGrp>
    ...
  </entry>
</superEntry>
```

[그림 3] 세종전자사전 표제어 '언도하다'의 예

1. 용언의 경우도 각 표제어의 전문분야(domain)가 분류되어 각 분야별로 온톨로지를 구축하기가 용이하다.

   (a) 세종전자사전은 용언도 전문분야로 표제어들이 분류되어 있어[11] 전문분야별로 용언 표제어를 자동 분류하여 분야별 온톨로지 구축이 가능하다. 특정한 전문분야의 온톨로지를 구축하기 위해 먼저 해당 전문분야의 표제어들을 선별, 취합한다. 전문분야는 <domain>과 </domain>이라는 XML 형태의 태그 사이에 기입된다.

      i. <orth>언도하다</orth> <domain>법률</domain>
      ii. <orth>접하다</orth> <domain>수학</domain>

      위의 예에서 '언도하다'는 '법률'의 전문분야로, '접하다'는 '수학'의 전문분야로 분류하여 온톨로지를 구축한다.

   (b) 양태연쇄어(troser)를 망라하여 전문분야 온톨로지를 구축한다.

      iii. 양태연쇄어는 체언의 동위어에 해당하고 여기에 나타나는 어휘들을 수집, 망라하여 해당 분야의 온톨로지를 구축한다.

2. 각 표제어가 체계적으로 분류된 의미부류에 따라 하위 분류되어 있다.

   (a) 의미하위부류 정보를 이용하여 'rdfs:subPropertyOf' 관계를 툴로서 자동화할 수 있다.[12]

3. 선택제약정보[13]와 의미역[14]을 활용하여 'rdfs:domain'과 'rdfs:range' 정보를 추출한다.

   (a) 의미역 정보가 'AGT'(행위주, Agent)[15]인 것에 해당하는 것을 선택제약정보에서 찾아서 그 값을 rdfs:domain의 값(Value)으로 넣어주고 의미역 정보가 'THM'(대상, Theme)인 것에 해당하는 것을 선택제약정보에서 찾아서 그 값을 'rdfs:range'의 값으로 넣어준다.

---

[11] 전문분야는 체언의 경우와 같으므로 각주 5 참조. 용언의 경우 2001년 결과물까지는 '법률'(언도받다), '수학'(접하다)만 분류되어 있음.

[12] 2001년 결과물까지 의미하위부류의 내용은 기입되지 않았지만, 내용이 채워 넣어지면 용언의 'rdfs:subPropertyOf' 관계를 이 정보를 이용해 자동 분류할 수 있다.

[13] 선택제약정보는 주어진 논항 구조(격틀)의 논항(명사와 보문)의 분포적 속성이다. 이 정보는 논항 구조(격틀) 바로 아래 위치하면서 각 논항에 분포하는 명사 부류에 대한 통사 의미적 정보를 명시적으로 나타낸다.

[14] 용언의 논항에 해당하는 의미역을 기록한다. 2001년 세종계획 용언 분과에서 확장된 의미역은 총 35개이다.

[15] 1998-2000년도에 '행위주(Agent:AGT)' 하나였던 분류가 2001년에는 대상행위주(Affected Agent:AFA), 공여주(Donner:DON), 수령주(Recipient:RCP), 공조행위주(Joined Agent:JAG), 원인주(Causer:CAU), 행위주(Agent:AGT), 상태주(Positionner:POS), 비의도행위주(Effector:EFF), 소유주(Possessor:PSS), 피해주(Patient:PAT)로 세분되었다.

i. <orth>언도하다</orth>
<synSem>
<selRst>N0=인물(판사)|단체(법원|가정법원) N1=인물
N2=(사형|형|징역)</selRst>
<thtRol>N0=DON N1=GOL N2=THM</thtRol>
...
</synSem>

⇓

<owl:ObjectProperty rdf:ID="언도하다">
<rdfs:domain rdf:resource="#판사">
<rdfs:range rdf:resource="#형">
</owl:ObjectProperty>

괄호 밖에 있는 최소 상위어(인물, 사물, 추상)는 변별성이 없을 정도로 포괄적이어서 괄호 안의 하위어를 값(value)으로 넣는다.

4. 동의어(synonym), 반의어(antonym), 양태상위어(trohpr), 양태하위어(trohpo), 양태연쇄어(troser) 정보가 있어 이들을 활용하여 동의어와 양태연쇄어는 'equivalentProperty', 반의어는 'inverseOf', 양태상위어와 양태하위어는 'rdfs:subPropertyOf'로활용한다.

(a) 동의어 (synonym)

i. 동의어 <syn>의 값은 표제어와 동의어 간에 'equivalentProperty' 관계를 제시한다.

<orth>언도하다</orth>
<semRel>
<syn>선고하다</syn>
...
</semRel>

⇓

<owl:ObjectProperty rdf:ID="언도하다">
<owl:equivalentProperty rdf:resource="#선고하다">
</owl:ObjectProperty>

(b) 반의어 (antonym)

i. 반의어 <ant>의 값은 표제어와 반의어 간에 'inverseOf' 관계를 설정한다.

```
<orth>언도하다</orth>
<semRel>
    <ant>언도받다</ant>
    ...
</semRel>

        ⇓

<owl:ObjectProperty rdf:ID="언도하다">
    <owl:inverseOf rdf:resource="#언도받다">
</owl:ObjectProperty>
```

(c) 양태상위어와 양태하위어 (hypertroponym & hypotroponym)

i. 양태상위어 <trohpr>의 값은 표제어와 양태상위어 간에 'rdfs:subPropertyOf' 관계를 나타낸다.

```
<orth>언도하다</orth>
<semRel>
    <trohpr>판결하다</trohpr>
    ...
</semRel>

        ⇓

<owl:ObjectProperty rdf:ID="언도하다">
    <rdfs:subPropertyOf rdf:resource="#판결하다">
</owl:ObjectProperty>
```

ii. 양태하위어 <trohpo>의 값도 표제어와 양태하위어 간에 'rdfs:subPropertyOf' 관계를 제시한다. 그러나 양태상위어의 경우와는 포섭 방향이 반대이다.

```
<orth>조사시키다</orth>
<semRel>
    <trohpo>취조시키다|문초시키다</trohpo>
...
</semRel>

        ⇓

<owl:ObjectProperty rdf:ID="취조시키다">
    <rdfs:subPropertyOf rdf:resource="#조사시키다">
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="문초시키다">
    <rdfs:subPropertyOf rdf:resource="#조사시키다">
</owl:ObjectProperty>
```

(d) 양태연쇄어

i. 양태연쇄어 <troser>의 값은 표제어와 양태연쇄어 간에 'equivalentProperty' 관계를 설정한다.

```
<orth>교부하다</orth>
<semRel>
    <troser>나누어주다</troser>
    ...
</semRel>

        ⇓

<owl:ObjectProperty rdf:ID="교부하다">
    <owl:equivalentProperty rdf:resource="#나누어주다">
</owl:ObjectProperty>
```

지금까지 세종전자사전의 정형화된 데이터에서 OWL 온톨로지를 자동으로 구축하는 방법을 보였으며 이상의 논의대로 온톨로지를 구축함에 있어 세종전자사전을 활용하는 이점을 정리해 보면 다음과 같다.

1. 세종전자사전은 정형의 SGML 구체적으로는 XML 포맷으로 되어 있어 컴퓨터에 의한 자동 분석과 처리가 용이하여 이를 온톨로지로 변환하는 컴퓨터 프로그램 툴을 사용하여 빠르고 자동으로 대용량 온톨로지를 구축할 수 있다.

2. 세종전자사전이 업데이트됨에 따라 온톨로지도 즉시 또 자동으로 업데이트가 용이하다.

3. 온톨로지를 처음부터 모든 항목의 내용을 일일이 작성하는 것 보다 대략적으로라도 온톨로지 초안을 자동구축한 다음 수정, 보완하는 것이 시간적으로나 인력적으로 경제적이다.

4. 이러한 시도는 세종전자사전과 유사한 정형의 데이터를 활용하여 온톨로지 구축을 자동화할 수 있음을 시사한다.

4. 맺음말

본 연구는 정형의 전자사전에서 자동 프로그램을 이용하여 쉽고 빠르게 대용량 온톨로지를 작성할 수 있는 방법론을 제시했다. 본고에서는 세종전자사전을 정형화된 데이터의 모델로 삼았으며 여기에서 사용된 메커니즘은 다른 정형의 데이터에도 공히 적용될 수 있으리라 생각한다. 세종전자사전의 체언과 용언을 각각 OWL의 'Class'와 'Property'로 분류한 다음 전문분야(domain)별로 나누어 각 분야의 온톨로지를 구축할 수 있

음을 보았다. 이 과정에서 체언과 용언의 의미하위부류와 동의어, 반의어, 상위어/양태상위어, 하위어/양태하위어, 동위어/양태연쇄어, 관계어 등의 데이터를 적절히 활용할 수 있음도 보았다.

물론, 여기에서의 결과물이 모든 어휘를 총망라한 완전한 온톨로지라는 것은 아니며 최소한 구축하고자 하는 분야의 온톨로지 초안으로 충분히 활용가치가 있고 각 분야 전문가와 온톨로지 구축자가 직접 수정, 보완할 것을 전제로 한다.

향후 보다 세밀한 검증과 반례 검토를 통하여 본고에서 제시한 방법론을 대용량 OWL 온톨로지 자동구축 프로그램의 알고리듬으로 작성하고 이를 적용하여 온톨로지 자동구축 프로그램으로 구현할 수 있기를 기대한다.

<참고문헌>

Gruber, Thomas R. 1993. A Translation Approach to Portable Ontology Specifications. KSL 92-71, Knowledge Systems Laboratory, Stanford University. http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html.

홍재성 외, 2001. *21세기 세종계획 전자사전 개발분과 연구보고서*. 문화관광부, 국립국어연구원.

웹사이트

21세기 세종계획. http://www.sejong.or.kr

OWL Web Ontology Language Guide, W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-guide

OWL Web Ontology Language Overview, W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features

OWL Web Ontology Language Reference, W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-ref

OWL Web Ontology Language Semantics and Abstract Syntax, W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-semantics

W3C Semantic Web. http://www.w3.org/2001/sw/

W3C Web Ontology Working Group. http://www.w3.org/2001/sw/WebOnt/

접수 일자: 2005년 5월 10일
게재 결정: 2005년 6월 2일