시소러스를 이용한 온톨로지의 Is-a 관계 설정

황금하°, 이신목, 남윤영, 신지애*, 최기선
한국과학기술원 전산학과 / 시맨틱웹첨단연구센터
한국정보통신대학교*
{hgh°, smlee, yynam, kschoi}@world.kaist.ac.kr, jiae@icu.ac.kr*

Identification of Is-a Relation in Ontology based on Thesaurus

Jin-Xia Huang°, Sheen-Mok Lee, Yun-Yeong Nam, Ji-Ae Shin*, Key-Sun Choi
Dept. of Computer Science, Korea Advanced Institute of Science and Technology /
Semantic Web Research Center
Information and Communications University*

요 약

시소러스의 개념과 개념간 계층관계가 온톨로지 구축에 흔히 이용되고 있다. 다만 시소러스 계층관계는 is-a관계 뿐만 아니라 세분화되지 않은 관계도 포함되고 있기 때문에, 온톨로지의 기본 관계인 is-a관계를 분별하는 작업이 필요하다. 본 논문은 개념의 어휘표현 정보를 이용하여 온톨로지의 is-a관계를 설정하는 규칙을 제시하였고, 개념의 정의문 정보를 이용하여 is-a관계를 검수하는 방법을 제안하였다. IT분야 시소러스에 대한 is-a관계 설정 실험결과, 어휘표현 정보를 이용한 규칙 기반 is-a관계 설정은 85.83%의 정확도를 보였고, 정의문 정보를 이용한 is-a관계 판단의 일관성 평가 결과 일치도가 86.44%였다.

1. 서론

온톨로지는 개념화의 명시적 규약(explicit specification of conceptualization)으로서[1,2], 사람과 컴퓨터간의 개념 및 개념 표현을 공유하기 위하여 개념의 종류, 개념간 관계, 및 개념의 사용에 대한 제약사항을 형식적으로 명백하게 정의해 주기 위한 것이다.

온톨로지는 개념, 개념의 속성, 개념간 관계, 및 개념의 구체적 사례(instance) 등 정보를 가지게 된다. 시소러스와 의미망 등 대부분의 지식베이스에서도 개념과 의미관계를 제공하고 있다. 시소러스에 있는 의미관계는 동의 유의 관계 외에, 상하위 및 부분전체(part-whole) 등 계층관계가 있다. 기존 시소러스와 의미망 등에서의 개념과 계층관계 정보를 이용하여 온톨로지를 구축하는 방법이 많이 연구되어 왔다[3].

시소러스의 계층 관계에서 상위어는 보다 일반적이거나 포괄적인(generic) 의미를 가지고 있기에 broader term(BT)이라고 하고, 하위어는 상대적으로 한정되거나 특정된 의미를 가지고 있기에 narrower term(NT)이라고 한다. 시소러스에는 계층관계 외에도 하위어가 상위어의 기능, 주변시설, 응용 등 의미를 나타내는 관련어(related term)로서 상하위어가 서로 준계층관계(quasi-hierarchy)를 이루기도 한다[4]. 본 논문에서는 시소러스에서 계층관계와 준계층관계를 이루는 상위어와 하위어 사이의 관계를 모두 BT/NT관계라고 부른다(그림1).

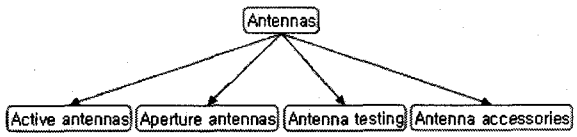

그림 1. 시소러스에서의 BT/NT관계

Is-a관계는 온톨로지에서 가장 기본적이고 중요한 관계이다. 본 연구에서는 시소러스의 BT/NT관계를 온톨로지의 is-a관계로(그림 2) 분별하여 설정하는 방법을 제시하고자 한다.

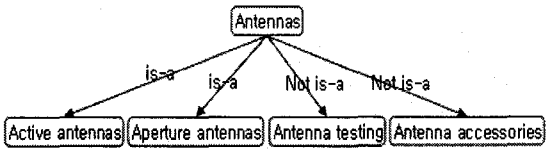

그림 2. BT/NT관계에 대한 is-a관계 설정

본 연구는 의미 기반의 IT839 서비스[5]를 위한 국가 IT 코어(Core) 온톨로지 구축 과제의 일환이다. 이 논문에서는 is-a관계 설정 대상으로 IT분야 시소러스인 INSPEC 시소러스[6]를 사용하였다.

2절에서는 INSPEC시소러스에 대하여 소개하고, 3절에서는 개념의 어휘표현을 이용한 is-a관계 설정 규칙을, 4절에서는 정의문 정보를 이용한 is-a관계 검수 방법을 제시하였다. 5절에서는 평가 방법 및 결과에

대하여, 마지막으로 6절에서는 토론과 향후 연구 방향에 대하여 소개한다.

2. INSPEC 시소러스

INSPEC은 IEE(The Institution of Electrical Engineers)가 제작하는 과학기술 문헌에 대한 기본 서지 정보 시스템으로, INSPEC이 수록 제공하는 주요주제 분야는 그림 3과 같다. 보다 정확하고 자세한 주제 정보를 제공하기 위하여 INSPEC은 색인 시스템으로 INSPEC분류체계를 제공하는 외에, INSPEC 시소러스도 가지고 있다[6].

- biomedical engineering
- biophysics
- **computing**
- **control engineering**
- **electrical and electronic engineering**
- geophysics
- **information technology**
- materials science
- oceanography
- nanobiotechnology
- nuclear engineering
- **physics**
- power & energy
- radar

그림 3. INSPEC이 제공 주요 14개 주제 분야.

Computing, control engineering, electrical and electronic engineering, information technology, 및 physic 등이 5개의 주된 분야이다.

INSPEC시소러스의 8,300여 개 개념어휘는 15,901개의 개념간 BT/NT관계를 형성하고 있는데, 이들 BT/NT관계가 바로 본 연구에서 진행한 is-a관계 설정 대상이다. INSPEC 시소러스의 개념들은 영어 어휘표현만 가지고 있는데, 본 연구에서는 TTA용어사전[7]을 이용하여 한국어 어휘표현을 추가하였고 정의문 정보도 부분적으로 추가하였다.

3. 개념의 어휘 표현 정보를 이용한 Is-a 관계 설정

주어진 BT/NT관계에서 두 개념의 본질적 속성(identity)이 같으면 이 두 개념이 is-a관계를 이룬다고 가정한다. 본 논문에서는 이 가정을 제1가정이라고 한다. 예를 들면, 그림 1에서, " Antenna" 와 " Active Antenna" 는 BT/NT관계를 이루는데, 이 두 개념이 같은 본질적 속성 " 안테나" 를 가지고 있기에, 이들은 is-a관계를 이룬다고 판단한다(그림2). " Antennas" 와 " Antenna testing" 은 BT/NT관계를 형성하지만, 본질적 속성이 같다고 볼 수 없으므로 이 관계는 is-a관계가 아니다. 본 논문에서는 BT/NT관계를 **btnt(NT, BT)**로 표기하고, is-a관계는 **isa(NT, BT)**로, 기타 관계는 **n-isa(NT, BT)**로 표기한다.

다음, 개념의 어휘표현에서 중심어가 해당 개념의 본질적 속성은 나타낸다고 가정한다. 이 가정을 제2가정이라 하고, 이와 제1가정에 의하여, 본 논문은 영어 및 한국어로 표현된 개념의 어휘 정보, 특히 중심어 정보를 이용한 is-a관계 설정 규칙을 제안한다. 이에는 동일 중심어 규칙, 중심어 관계의 이행 규칙, 중심어의 다양성 포용 규칙, 중심어의 약자 허용 규칙 등이 있다. 이런 규칙을 적용하기 위하여 중심어 인식이 필요한데, 그 방법에 대하여 다음 절에서 우선 소개한다.

3.1 중심어 인식

중심어 정보를 이용한 is-a관계 판단을 위하여 개념의 어휘표현에 대한 중심어 인식이 필요하다.

IT분야 개념의 영어 어휘표현에서 중심어는 기본적으로 복합용어에서의 마지막 부분에 위치하고, 단일 어휘로 구성된 경우 그 어휘 자체가 중심어로 된다. 마지막 어휘가 중심어로 안 되는 경우에는 다음의 패턴을 적용하여 중심어를 인식한다:

- <headword><preposition><otherword>
  <preposition>={by, in, on, of, from, for, with, about}
  예: " learning by example" 의 중심어는 " learning" .
- <headword>_<otherword>
  이런 경우 otherword는 domain정보를 나타낸다.
  예: " network_circuits" 의 중심어는 " circuits" .
- <otherword><headword>
  예: " unsolicited_e-mail" 의 중심어는 " mail"
- 기타: 일부 중심어는 빈자리나 " _" 표기 없이 기타 독립단어와 연결되어 나오는 경우 가 있다. 예를 들면 " radiotelephony" 의 중심어는 " telephony" 이다.

3.2 동일 중심어 규칙

앞에서 소개한 제2가정에 의하면, BT/NT관계에서, 두 개념이 같은 중심어를 가지면 이 두 개념은 본질적 속성이 같다는 결론에 이르게 된다. 제1가정까지 고려하여, BT와 NT가 같은 중심어를 가지면 is-a관계로 설정한다. 이것이 동일 중심어 규칙이다.

다음은 개념의 영어 어휘 표현에 대하여 동일 중심어 규칙을 적용한 예이다:

isa(active antenna array, antenna array)
isa(elastic waves, waves)

일부 개념의 영어 어휘 표현이 서로 다른 중심어를 가지고 있지만 한국어 어휘 표현이 같은 중심어를 가지는 경우가 있다. 예를 들면, 주어진 관계 btnt(personal digital assistant, terminal)에서, NT " personal digital assistant" 의 한국어 어휘 표현은 " 개인 휴대 정보 단말기" 로서, BT " terminal" 의 한국어 어휘 표현인 " 단말기" 와 같은 중심어를 가진다. 그러므로 이 관계는 is-a관계로 설정한다.

isa(personal digital assistant, terminal)

3.3 중심어 관계의 이행(transitivity) 규칙

BT/NT관계하에서, 두 개념의 중심어가 서로 다르지만, 본질적 속성이 같기에 서로 is-a관계를 이루면, 이 두 개념도 is-a관계를 가진다고 판단한다. 이는 is-a관계의 이행성을 이용한 규칙인데, 이를 중심어 관계의 이행 규칙이라고 한다. 이런 is-a관계를 이루는 중심어는

분야에 의존한다.

다음은 차세대 이동통신 분야의 주어진 BT/NT관계에서 서로 is-a관계를 이루는 중심어의 예이다:

isa(programs, listings)
isa(theory, methods)

위의 중심어 관계를 이용하여 주어진 BT/NT관계에 대하여 중심어 관계의 이행 규칙을 적용한 예는 다음과 같다:

isa(JAVA listings, complete computer programs)
isa(smoothing methods, filtering theory)

3.4 중심어의 다양성 포용 규칙

일부 포괄적인 의미를 가지는 개념 어휘는 그 하위 개념의 어휘표현이 다양할 수 있다. 예를 들면 주어진 BT/NT관계하에서, BT " equipment" 는 이의 NT " receivers" , " antennas" , " cameras" , " tubes" , " transmitters" 등과 모두 is-a관계를 이룬다. IT분야에서 이런 포괄적인 의미를 가지는 개념 어휘로 " equipment" , " accessories" , " applications" 등이 있다.

주어진 BT/NT관계에서, BT가 이런 하위 개념의 다양성을 포용하는 어휘를 중심어로 가질 경우, 중심어 관계의 이행 규칙에 의하여, NT들과 is-a관계를 이루게 된다. 이를 중심어의 다양성 포용 규칙이라고 한다. 관련된 예는 다음과 같다:

isa(radio receivers, radio equipments)
isa(antenna feeds, antenna accessories)
isa(radio tracking, radio applications)

3.5 중심어의 약자 허용 규칙

어떤 중심어는 그 하위 개념의 어휘적 표현에서 약자를 많이 사용한다. 이런 약자 허용 중심어로는 " languages" , " standards" , " networks" 등이 있다. 약자의 판단은 대문자 사용 여부로 판단 가능하다. 관련된 예는 다음과 같다:

isa(BASIC, high level languages)
isa(Bluetooth, telecommunication standards)
isa(ISDN, telecommunication networks)

4. 정의문 정보를 이용한 Is-a 관계 판단

개념의 어휘 표현 정보로 판단이 어렵거나 잘못된 경우는 정의문 정보를 이용하여 판단 수정할 수 있다. 정의문 정보를 이용하여 개념의 상위개념과 본질적 속성을 판단한다. 정의문 정보는 TTA사전 개념 정의[7], Wikipedia 정의문 정보, 네이버 백과사전 개념 정의와 기타 웹 검색엔진을 이용하여 얻은 정의문 정보를 사용 할 수 있다.

정의문 정보를 이용한 is-a관계 판단 방법을 아래에서 예를 이용하여 설명하고자 한다.

4.1 개념의 정의문 정보 이용

예: isa(antenna array, antenna)

" antenna array" 의 한국어 정의문은 " 2개 이상의 안테나 소자를 적절한 간격으로 배치한 안테나이다." 로서, 이 정의문으로부터 알 수 있는바 " antenna array" 의 본질적 속성은 " antenna" 이다. 때문에 이 두 개념은 is-a관계를 이룬다.

예: n-isa(subroutines, software)

" subroutines" 의 Wikipedia 정의는 " In computer science, a subroutine is a portion of code within a larger program." 으로서, 이 정의문에 의하여 이 두 개념간 관계는 부분전체관계로, is-a관계가 아니다.

4.2 NT의 중심어 정의문 정보 이용

예: isa(수은 정류기, 회로소자)

위 관계에서, NT 의 중심어 " 정류기" 의 정의는 " 교류전력에서 직류전력을 얻기 위해 정류작용에 중점을 두고 만들어진 전기적인 회로소자." 로서, 그 본질적 속성이 " 회로소자" 임을 알 수 있다. 때문에, isa(정류기, 회로소자)가 성립하고, 동일 중심어 규칙에 의하여 isa(수은 정류기, 정류기)가 성립되기에, is-a관계의 이행성에 의하여 isa(수은 정류기, 회로소자)가 성립하게 된다.

4.3 BT의 의미에 애매성이 있는 경우

BT/NT관계에서 BT의 의미에 애매성이 있는 경우, NT가 그 중의 한 가지 의미와 is-a관계를 이루면, 두 개념간에는 is-a관계가 성립한다고 판단한다. 의미의 판단은 정의문 의미 해석 또는 대역어 사전으로 확실한 근거가 있는 경우에만 이 지침을 적용한다.

예: isa(accident prevention, safety)

그림 4와 같이, 개념 " safety" 에는 " 안전" 과 " 안전책" 이라는 두 가지 의미가 있다. 때문에 NT " aerospace safety" 와 " marine safety" 는 " 안전" 을 뜻하는 개념으로 " safety" 와 is-a관계가 성립되고, " accident prevention" 은 " 안전책" 의 의미로 " safety" 와 역시 is-a관계가 성립된다.

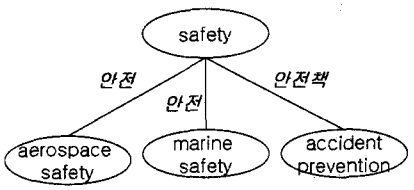

그림 4. BT의 의미에 애매성이 있는 경우의 관계 설정

5. 평가

이 절에서는 3절에서 소개한 개념 어휘 표현 정보를 이용한 is-a관계 설정의 정확도에 대한 평가와, 4절에서

소개한 정의문을 이용한 is-a관계 설정의 일관성에 대한 평가를 진행하고, 그 방법 및 결과를 소개한다.

5.1 어휘표현 정보를 이용한 규칙기반 Is-a관계 설정 평가

어휘표현 정보를 이용한 규칙 기반 설정 결과를 R1이라 하고, 이를 정의문 정보를 이용한 전문가에 의한 검수 결과(R2)와 비교하여 정확도를 평가한다 (공식1).

$$Accu = \frac{|R1 \cap R2|}{|R2|} \quad (1)$$

INSPEC시소러스의 개념 및 개념 관계를, IT839에 대응하는 4개의 분야로 분류하여 평가하였다(표1). 전체 관계수(칼럼2) 중에서 is-a관계(칼럼3)의 비례를 is-a관계 설정 실험의 베이스라인으로 사용하였다(칼럼4). 공식1에 의하여 어휘표현 정보를 이용한 규칙 기반 is-a관계 설정 정확도를 평가하였다(칼럼5). 전체 4,999개 관계에 대한 설정 결과, 정확도는 85.83%로, 베이스 라인 79.59%보다 높은 정확도를 보였다.

표 1. 작업 대상 분야 및 대상 개념/관계 수

분 야 명	관계수	Is-a 관계수	Baseline 정확도(%)	정확도 (%)
차세대 이동통신 분야	1,730	1,373	79.36	88.78
디지털TV/방송 분야	2,288	1,812	79.20	86.89
지능형 로봇 분야	912	735	80.59	77.63
DC 및 S/W 솔루션 분야	69	59	85.51	85.51
통계	4,999	3,979	79.59	85.83

5.2 일관성 평가

정의문 정보를 이용한 is-a관계 검수는 그 판단이 애매한 경우가 있기에 관계 판단에 대한 일관성 평가가 필요하다. 일관성 평가는 그림 5와 같이, 두 명 이상의 전문가에 의하여 설정된 관계(R3, R4)에 대하여 이들의 일치도를 평가하는 것이다(공식 2).

$$Cons = \frac{|R3 \cap R4|}{(|R3|+|R4|)/2} = \frac{|R3 \cap R4|}{|R3|} \quad (2)$$

개념	상위개념	설정관계	일관성검사	전문가
maximum_entropy_methods	entropy	0		전문가1
maximum_entropy_methods	entropy	1		전문가2
minimum_entropy_methods	entropy	0		전문가1
minimum_entropy_methods	entropy	1		전문가2
flash_memories	EPROM	0		전문가1
flash_memories	EPROM	0		전문가2

그림 5. 일관성 검사 대상

표1의 4개 분야 INSPEC 시소러스 개념관계에 대하여 평가한 결과, 두 명 이상의 전문가가 작업한 관계 수는 총 2,994개이고, 이 들 관계에 대하여 공식 2에 따라 평가한 결과 86.44%의 일치도를 보였다.

5. 결론 및 향후연구

시소러스에서의 개념과 개념간의 관계(BT/NT관계)가 온톨로지 구축에 흔히 이용되고 있다. 본 논문은 BT/NT관계를 온톨로지에서의 기본 관계인 is-a관계로 설정하는 방법들을 제안하였다. 구체적으로, 개념의 어휘표현 정보를 이용한 is-a관계 설정 규칙과, 정의문 정보를 이용한 관계 검수 방법을 제시하였다. IT분야 온톨로지 구축을 위하여 IT분야 시소러스인 INSPEC시소러스를 이용하여 실험하고, 이에 대한 관계 설정 정확도 및 검수 작업의 일관성에 대한 평가도 진행하였다.

향후 어휘표현 정보, 정의문 정보, 용례 정보 등 다양한 정보를 이용하여 is-a관계를 포함한 온톨로지 개념 관계와 제약조건 등을 설정하는 방법에 대하여 연구를 계속하고자 한다. 또한 본 논문은 BT/NT관계가 주어진 상황에서의 is-a관계 설정에 초점을 두고 있지만, 향후에는 BT/NT관계가 주어지지 않은 경우의 개념간 관계 설정 방법에 대해서도 연구하고자 한다.

감사의 글

본 논문은 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참고 문헌

[1]Gruber, T.R., " A Translation Approach to Portable Ontology Specifications" , Knowledge Acquisition, 5(2), pp.199-220, 1993

[2] 최기선, 류법모, " 온톨로지 구축과 학습: 상하위 관계" , 시맨틱웹과 온톨로지 기술동향 특집, 정보과학회지, 제24권, 제4호, pp.24-30, 2006년 4월

[3] 최호섭, 임지희, 배영준, 최수일, 옥철영, " 온톨로지 구축 방법과 사례" , 시맨틱웹과 온톨로지 기술동향 특집, 정보과학회지, 제24권, 제4호, pp.31-44, 2006년 4월

[4]Helen M.Townley and Ralph D.Gee, " Thesaurus-Making: Grow Your Own Word-Stock" , London: Deutsch, 1980, pp.23

[5] 정보통신부, " IT839 전략" , 2006.4. http://help.mic.go.kr/eBook/report_2006/it839/it839/default1.html

[6] The Thomson Corporation, " INSPEC User Guide" , http://www.thomsonscientific.co.kr/images/INSPEC%202.0%20QRC%20Kor.pdf

[7] 정보통신용어사전, 한국정보통신기술협회 발간, http://word.tta.or.kr/faq/faq.jsp