개념 및 관계 분류를 통한 분야 온톨로지 구축
(Building Domain Ontology through Concept and Relation Classification)

황 금 하 † 신 지 애 †† 최 기 선 †††
(Jin-Xia Huang) (Ji-Ae Shin) (Key-Sun Choi)

요 약 본 논문에서는 분야 온톨로지 구축을 위하여 분야 상위 온톨로지를 구축한 다음, 분야 시소러스의 개념과 관계를 이용하여 분야 상위 온톨로지를 확장하는 방법을 제안한다. 이를 위하여 우선 일반 분야 시소러스와 분야 사전을 이용하여 분야 상위 개념 분류체계를 구축한다. 다음, 분야 시소러스의 개념을 분야 상위 온톨로지의 상위 개념으로 분류하고, 광의어(Broader Term: BT)-협의어(Narrower Term: NT) 및 광의어-관련어(Related Term: RT) 사이의 관계를 분야 상위 온톨로지에서 정의한 의미관계로 분류한다. 개념 분류는 두 단계로 진행되는데, 1단계에서는 빈도수 기반 방법, 2단계에서는 유사도 기반 방법을 적용하여 시소러스 개념을 분야 상위 온톨로지의 개념으로 분류한다. 관계 분류에서는 두 가지 방법을 적용하였는데, (i) 훈련데이타가 부족한 경우를 위하여 규칙기반 방법으로 BT-NT/RT관계를 *isa*와 기타 관계(*non-isa* 관계)로 분류하고, 다시 패턴기반 방법으로 *non-isa*관계를 온톨로지를 위한 의미관계로 분류한다. (ii) 훈련데이타를 충분히 가지고 있을 경우, 최대 엔트로피 모델(MEM)을 적용한 특징기반 분류 기법을 사용하되, k-Nearest Neighbors(k-NN)방법으로 훈련데이타를 정제하였다. 본 논문에서 제안한 방법으로 시스템을 구축하였고, 실험 결과 사람에 의한 판단 결과와 비교 가능한 성능을 보여 주었다.

키워드 : 분야 온톨로지, 분야 상위 온톨로지, 시소러스, 개념, 관계, 분류

Abstract For the purpose of building domain ontology, this paper proposes a methodology for building core ontology first, and then enriching the core ontology with the concepts and relations in the domain thesaurus. First, the top-level concept taxonomy of the core ontology is built using domain dictionary and general domain thesaurus. Then, the concepts of the domain thesaurus are classified into top-level concepts in the core ontology, and relations between broader terms (BT) - narrower terms (NT) and related terms (RT) are classified into semantic relations defined for the core ontology. To classify concepts, a two-step approach is adopted, in which a frequency-based approach is complemented with a similarity-based approach. To classify relations, two techniques are applied: (i) for the case of insufficient training data, a rule-based module is for identifying isa relation out of *non-isa* ones; a pattern-based approach is for classifying non-taxonomic semantic relations from *non-isa*. (ii) For the case of sufficient training data, a maximum-entropy model is adopted in the feature-based classification, where k-NN approach is for noisy filtering of training data. A series of experiments show that performances of the proposed systems are quite promising and comparable to judgments by human experts.

Key words : domain ontology, core ontology, thesaurus, concept, relation, classification

· 본 연구는 지식경제부 및 정보통신연구진흥원의 정보통신선도기반기술개발 사업의 지원으로 수행되었습니다.

† 학생회원 : 한국과학기술원 전자전산학부
hgh@world.kaist.ac.kr
†† 정 회 원 : 한국정보통신대학교 전산학과 교수
jiae@icu.ac.kr
††† 종신회원 : 한국과학기술원 전자전산학부 교수
kschoi@cs.kaist.ac.kr

논문접수 : 2008년 1월 3일
심사완료 : 2008년 8월 13일

정보과학회논문지: 소프트웨어 및 응용 제35권 제9호(2008.9)

1. 서 론

온톨로지는 사람과 컴퓨터, 또는 컴퓨터간의 개념 및 개념 표현을 공유하기 위한 개념화의 명시적 규약을 의미한다[1,2]. 해당 분야의 개념, 개념의 구체적 사례인 인스턴스(instance), 개념 혹은 인스턴스간 관계, 추론 규칙인 공리(axiom) 등 정보를 포함한다. 시소러스는 특정 주제 영역에서 사용하는 용어와 용어간 의미관계를 체계적으로 제시하는 어휘집으로서, 해당 분야의 개념, 용어, 개념 또는 용어간의 동의 및 유의 관계, 상하위 관계, 부분 전체 관계 등 정보를 제공한다. 이런 정보는 온톨로지를 구성하는 가장 기본적인 정보이기에, 시소러스를 이용하여 온톨로지를 구축할 수 있다[3,4].

시소러스와 온톨로지는 모두 개념과 관계 정보를 가지고 있지만, 시소러스의 관계는 주로 상하위 관계와 동의/유의 관계, 부분 전체 관계로 국한되어 있고, 기타 관계를 가지고 있더라도 세분화하지 않고 BT-NT/RT 관계로만 표현한다[3,5-7]. 예를 들면, 분야 시소러스인 Inspec[1][5]의 계층체계에는 *isa* 관계와 *non-isa* 관계가 혼재하여 있는데, 예를 들면 그림 1에서 "SCADA systems"와 "Data acquisition" 사이는 *non-isa* 관계임에도 불구하고 *isa* 관계와 똑 같이 표현된다. *isa* 관계로 구성된 엄격한 상하위 계층체계(taxonomic hierarchy)와 구별하기 위하여, 본 논문에서는 분야 시소러스의 이러한 계층체계를 BT-NT/RT계층체계라고 부르기로 한다.

시소러스가 지식 표현 자체에 그 목적을 두고 있다면, 온톨로지는 자동추론을 최종 목적으로 하고 있기에, 엄격한 상하위 관계(*isa* 관계)와 세분화된 관계 정보로 개념의 속성을 나타낸다. 예를 들면 동물 온톨로지에서는 동물의 식성을 표현하기 위하여 *eats*라는 관계가 필요하고, 피자 온톨로지에서는 피자의 토핑과 산지 속성을 나타내기 위하여 *hasTopping*과 *hasCountryOfOrigin*등 세분화된 관계를 사용한다[8].

분야 시소러스를 이용하여 온톨로지를 구축하기 위해서는, 그림 2에서처럼 BT-NT/RT관계를 세분화된 관계로 분류하여야 한다.

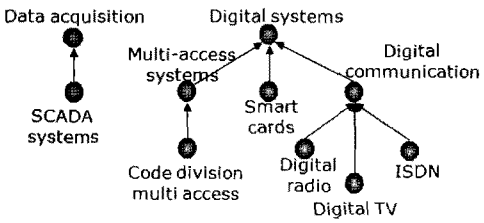

그림 1 분야 시소러스의 BT-NT/RT관계

1) http://scientific.thomson.com/media/scpdf/inspec_gettingstarted_en.pdf

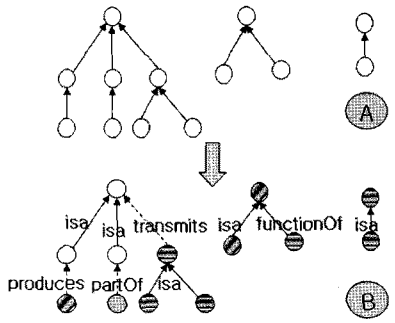

그림 2 시소러스에 대한 관계 분류: 온톨로지 구축을 위해서는 A부분이 보여주는 분야 시소러스에서의 BT-NT/RT관계를, B부분이 보여 주는 바와 같이, 온톨로지를 위한 의미관계로 분류하여야 한다.

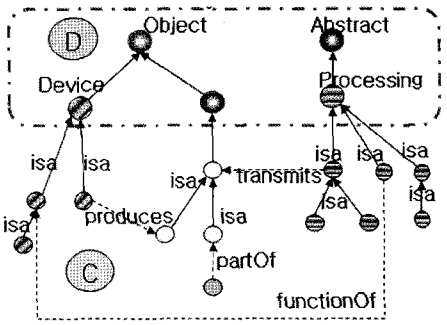

그림 3 개념 및 관계 분류를 통하여 구축된 온톨로지

관계 분류 후 BT와 RT가 상위어 없이 최상위 개념으로 남는 것을 방지하기 위하여, 본 논문에서는 개념 분류를 통하여 이들 용어를 분야 상위 온톨로지의 개념(의미 카테고리)으로 분류한다. 이러한 개념 및 관계 분류를 통하여 시소러스에서의 BT-NT/RT계층체계는 온톨로지를 위한 상하위 계층체계로 변환된다(그림 3). 개념 분류의 목표 카테고리로 시소러스의 자체 상위 개념이 아닌 온톨로지의 상위 개념을 이용하는데, 그 이유는 기존 분야 시소러스의 규모가 너무 작거나 그 분야가 목표 온톨로지의 분야와 조금씩 다르기에, 시소러스의 상위 개념이 목표 온톨로지의 분야를 대표하기에 부족한 경우가 많기 때문이다.

본 논문에서의 접근 방법은, 그림 2와 3이 보여 주는 바와 같이, 우선 분야 상위 온톨로지를 구축한 다음(D), 분야 시소러스에 대하여 개념 분류(B→C,D) 및 관계 분류(A→B)를 수행함으로써 분야 상위 온톨로지를 확장하는 것이다. 용어의 개념 정보는 관계 분류에도 도움 되기에, 우선 용어의 개념을 분류한 다음, 용어간 관계를 분류하기로 한다.

분야 상위 온톨로지는 해당 분야에서 가장 기본적이

고 중요한 상위 개념들을 포함하며[9-11], 해당 분야를 대표하는 확장 가능한 모델을 제공함으로써 다양한 소스로부터 획득한 데이타를 통합하기 위하여 구축된다[12]. 본 논문에서의 분야 상위 온톨로지는 IT분야 상위 개념 분류체계와 IT분야 의미관계를 포함하고 있는데, IT분야 상위 개념 분류체계는 대용량 IT분야 전문용어 사전[13]을 일반 분야 시소러스에 매핑하여 구축하였고, IT분야 의미관계는 온톨로지 개발자가 정의하였다.

분야 상위 온톨로지의 확장에서는 분야 시소러스 Inspec을 이용하였다. Inspec 시소러스는 전산, 제어공학, 전자 전기공학, 정보기술, 물리학 등 약 14개 분야를 포함하고 있으며 8,300개 이상의 용어와 15,901개의 BT-NT/RT 관계들을 가지고 있는데[5], 이 관계들은 그림 1이 보여 주듯이 BT-NT와 BT-RT관계가 구분없이 섞여 있다. 이를 이용하여 분야 상위 온톨로지를 확장하기 위하여, 분야 상위 온톨로지의 개념을 의미 카테고리로 간주하고, 빈도수 및 유사도 기반의 개념 분류 방법에 의하여 Inspec용어를 분야 상위 온톨로지 개념으로 분류한다. 또한 Inspec 시소러스의 BT-NT/RT관계를 분야 상위 온톨로지의 특정 의미관계로 분류하기 위하여 우선 훈련데이타가 부족한 경우를 위한 규칙 및 패턴기반 방법을 제안하였고, 충분한 훈련데이타를 확보한 후에는 MEM및 k-NN기법을 이용한 특징기반 관계 분류기법을 사용하되, 그 특징 정보로 규칙 및 패턴 기반 관계 분류에서 사용한 어휘 및 개념 정보를 사용하였다.

본 논문은 다음과 같이 구성되었다. 2장에서는 기존의 언어 지식베이스를 이용한 분야 온톨로지 구축 방법들에 대하여 조사한다. 3장에서는 IT 분야 상위 온톨로지의 구축 방법에 대하여 기술하고, 4장과 5장에서는 개념 및 관계 분류를 통한 분야 상위 온톨로지 확장 방법에 대하여 설명한다. 6장에서는 실험과 그 결과에 대하여 설명하고 7장에서는 결론을 맺는다.

2. 관련 연구

시소러스와 같은 기존의 지식 베이스는 해당 분야에서 자주 사용되는 전문용어 및 개념 정보를 가지고 있을 뿐만 아니라, BT와 NT/RT간의 관계 정보도 제공한다. 이들 중 일부 지식 베이스는 추론에 사용되는 제약조건(constraint) 정보를 제공하기도 한다[14]. 그러나 일반적으로 기존의 지식베이스에서는 온톨로지 구축에 필요한 정보의 일부분만을 얻을 수 있다. 따라서, 분야 시소러스를 온톨로지 구축에서 활용하는 연구에 대한 필요성이 제기 되어 왔고[3,15], 언어 지식 베이스를 이용한 온톨로지 구축 방법에 대한 연구도 꾸준히 진행되어 왔다.

지식 베이스로부터 온톨로지를 구축하는 연구는 크게 네 가지 동향으로 나누어 볼 수 있다. 첫째는 기존 지식 베이스를 온톨로지 포맷으로 변환하는 것인데[16], 새로운 온톨로지 지식의 생성이 없이, 기존의 시소러스 포맷을 RDF나 OWL과 같은 온톨로지 표현들로 변환한다. 이런 연구에서는 각각의 시소러스 표현 방식에 대하여 조사 연구 후, 이를 패턴기반이나 규칙기반 방법으로 온톨로지 포맷으로 변환한다. 둘째는 기존 지식 베이스로부터 유용한 정보를 추출하여 온톨로지 지식으로 변환하는 것 인데[14,17-19], 예를 들면 기존의 논리 프로그램으로부터 제약 조건을 추출하여 온톨로지 지식으로 변환해 주기도 한다. 셋째는 관계 정보를 확장함으로써 시소러스를 온톨로지로 리모델링하는 연구인데[6,7,18], 격 관계(case relation)와 의미관계를 시소러스의 상하위 계층체계에 추가함으로써 시소러스를 온톨로지로 확장하기도 한다[17,20]. 격 관계는 기존 기계번역 시스템과 사전으로부터 얻어 지며, 의미관계는 말뭉치의 상관관계 정보를 이용하여 획득한다. 넷째는, 시소러스의 BT-NT/RT관계를 사람에 의하여 정의한 규칙이나 패턴을 이용하여 의미관계로 분류하는 것인데[6,7], 이런 연구는 특히 분야 온톨로지의 구축에서 많이 사용된다.

분야 온톨로지 구축을 위하여 분야 상위 온톨로지를 먼저 구축한 다음, 이를 기반으로 기타 언어 자원을 통합하는 방법이 많이 사용되어 왔다[10-12, 21,22]. 분야 상위 온톨로지 구축에서는 전문가가 분야 상위 개념과 제약 조건을 정의하거나[23,24], 기존의 두 개 분야 상위 온톨로지를 하나로 통합함으로써 새로운 분야 상위 온톨로지를 구축하는 방법이 사용되었다[12].

분야 상위 온톨로지를 우선 구축한 다음 이를 이용하여 분야 온톨로지를 자동으로 구축하는 측면에서 본 연구는 이런 연구들과 유사하다[10-12, 21,22]. 다만 분야 상위 온톨로지 구축 방법으로, 본 연구에서는 일반 분야 시소러스와 분야 사전을 이용하여 IT분야 상위 개념 체계를 자동으로 구축하였다. 한편, 관계 분류를 통하여 분야 시소러스로 분야 온톨로지를 구축하는 면에서, 본 논문은 [6,7]의 연구와 유사하다. 그러나 본 과제의 연구 대상인 IT 분야와 Inspec 시소러스는 기존의 연구들보다 포괄적인 분야를 다루기 때문에, 패턴을 수동적으로 정의하기엔 어려움이 있다. 이 문제를 해결하기 위하여 본 논문에서는 규칙과 패턴기반 관계 분류 외에 지도식 관계 분류 방법을 제안하였다. 본 과제가 이런 기존 연구와의 또 다른 차이점은, 본 논문에서는 관계 분류뿐만 아니라, 개념 분류도 수행하여, 분야 시소러스의 BT-NT/RT계층을 분야 온톨로지를 위한 상하위 계층으로 변환시키고 있다.

분류 기법에 대해서는 많은 연구가 진행되어 왔는데,

특히 문서 분류에서는 신경망, 결정 트리, k-NN, 지지 벡터 기계, 베이지안 통계, MEM등 다양한 기계학습 기반 분류기법들이 사용되었다. 이러한 기법들은 문서 분류뿐만 아니라 관계 분류에서도 사용되는데, 예를 들면 어휘의 의미 역할 결정(semantic role labeling)[25]과 명사의 의미관계 분류[26] 등 문제에서도 이런 분류기법을 사용하는 것을 볼 수 있다. 다만 관계 분류에서는 분류 기법 자체에 대한 연구보다는 관계 분류에서의 특징 사용에 초점을 맞추고 있는 것을 볼 수 있다[26]. 본 연구에서는 개념 분류에서는 빈도수와 유사도 기반 방법을 사용하였고, 관계 분류에서는 훈련데이타 부족 시 규칙 및 패턴 기반 방법을 사용하였으며, 일정한 훈련데이타가 축적 된 후에는 MEM을 이용한 특징기반 방법을 사용하되, 훈련 데이타 정제를 위하여 k-NN방법과 결합 사용함으로써 정확도를 향상 시켰다.

3. IT 분야 상위 온톨로지

분야 상위 온톨로지는 일반 분야 온톨로지와 특정 분야 온톨로지를 이어 주는 역할을 한다. 여기에 속한 개념은 해당 분야에서의 중요한 개념으로, 일반인도 이해하기 쉬워야 하며, 해당 분야 하위 개념의 이해를 도울 수 있어야 한다. 분야 상위 개념 분류체계의 구축을 분야 상위 개념 선정과 개념간 분류체계 구축으로 나눌 수 있다.

본 연구에서 IT분야 상위 개념을 일반성, 보편성, 구체성의 세 가지 기준에 근거하여 선정한다.

일반성의 기준

분야 상위 개념은 비 전문가도 이해하기 쉬운 개념이어야 한다. 일반성 기준을 따르기 위하여, IT분야 상위 개념을 "전기기기"와 같이 일반 분야 시소러스에도 포함된 개념 중에서 선정하였다.

보편성의 기준

분야 상위 개념은 해당 분야에서 자주 사용되는 개념이어야 한다. 이는 해당 분야의 중요 개념을 선별하기 위한 기준이다. 예를 들면, 개념 "전기기기"는 IT분야에서 자주 사용되는 개념으로, IT분야 상위 개념으로 적합하지만, "농약"은 IT분야에서 출현 빈도가 낮은 개념으로 IT분야 상위 개념으로 적합하지 않다.

본 연구에서는 개념 분류를 통하여 대용량 IT분야 전문용어 사전[2][13]의 한국어 용어를 일반 분야 시소러스 CoreNet[3][27]의 개념으로 매핑한 후, 각 개념으로 매핑된 IT분야 용어수로 IT분야에서 해당 개념의 보편성 점수를 계산하였다. CoreNet 은 2,900 여 개의 개념(카테고리) 과 50,000 여 개의 한국어 상용어휘를 가지고 있는 일반분야 시소러스이다.

t 로 IT 분야 용어를, h_t 로 t 의 중심어를 표시하고, h_t 가 가지고 있는 m개의 의미는 CoreNet 개념 $\{c_1,\cdots c_j,\cdots c_m\}$에 각각 대응된다고 가정한다. 중심어 h_t가 개념 c_j 로 개념분류(의미태깅)되는 확률을 $\Pr(c_j \mid h_t)$로 표시하면, IT 분야에서 용어 t 의 상위 개념 c_t 는 식 (1)에 의하여 분류된다.

$$c_t = c_{h_t} = \arg\max_c \Pr(c_j \mid h_t), \forall 1 \le j \le m \quad (1)$$

위 공식에서, $\Pr(c_j \mid h_t)$는 일반 분야 의미태깅 말뭉치인 KAIST 의미 태깅 말뭉치로부터 얻었다[28]. 식 (1)로 대량의 IT분야 용어를 CoreNet 개념으로 매핑 한 후, 각 개념으로 매핑된 용어의 개수에 의하여 해당 개념의 보편성 점수를 결정한다.

구체성 기준

개념의 일반화 정도나 보편화 정도가 지나치면 개별적 특징에 대한 기술이 불가능하므로 의미 정보 전달의 정확성을 저하시킬 수 있다. 이를 방지하기 위하여 구체성 기준을 제안하는바, 일정한 임계 값 이상의 보편성 점수를 가지는 개념 중, 일반 분야 시소러스에서 상대적으로 하위 노드에 위치한 개념을 선택한다. 구체성 기준은 일반성 기준과 보편성 기준과 서로 충돌 하기에, 일정한 선에서 타협점을 찾아야 한다.

위의 세가지 기준에 근거하여 IT분야 상위 개념을 선정하였는데, 우선 2,900여 개의 CoreNet 개념 중 800여 개의 개념을 자동으로 선택한 다음, 이 중에서 전문가가 200개의 개념을 최종으로 선정하였다.

선정된 개념 사이의 상하위 관계는 일반 분야 시소러스에서 해당 개념들의 상하위 관계를 그대로 승계한다. 즉, IT분야 상위 개념 분류체계는 일반 분야 시소러스인 CoreNet의 일부로, CoreNet개념 중 IT 분야에서 보편적으로 자주 사용되는 개념을 선택한 것이다. 그림 4는 CoreNet에서 개념 "인공물"의 하위 트리 구조를 부분적으로 보여주고 있다. 그림에서 회색 노드는 IT분야 상위 개념으로 선택된 CoreNet개념이고, 흰색 노드는 선택되지 않은 CoreNet개념이다.

IT 분야 상위 온톨로지는 IT분야를 위하여 제안된 의미관계 유형도 포함한다(그림 3). 이들 의미관계는 정

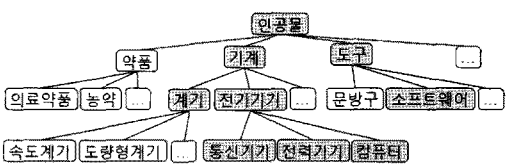

그림 4 IT 분야 상위 온톨로지의 개념 분류체계(회색)

2) http://korterm.or.kr
3) http://bola.or.kr/CoreNet_Project

표 1 IT분야 온톨로지에서 정의된 의미관계

Relation	Domain	Range
functionFor	Function	Analysis
functionIn	Function	Logic
functionOf	Function	Plan
theoryAbout	Theory	Structure
theoryAbout	Theory	Equipment
theoryOf	Theory	Information

의역과 치역을 제약으로 가지고 있는데, 정의역과 치역은 IT분야 상위 개념으로 일반화하였다(표 1). 이런 관계 트리플(triple)은 관계 분류에서 패턴으로 사용할 수 있다.

본 논문의 작성 당시, IT 분야 상위 온톨로지는 총 185개의 의미관계 유형을 가지고 있었고, 이 중 108개의 의미관계 유형에 대하여 258개의 관계 트리플이 정의되어 있었다. 이런 의미관계는 온톨로지 개발자가 정의하였다.

4. 개념 분류

본 절에서는 분야 온톨로지 구축을 위한 개념 분류(그림 2와 3: B→C, D) 방법에 대하여 기술한다. Inspec 용어를 200개의 IT 분야 상위 개념으로 분류하였는데 이는 두 단계로 나뉜다: 첫 단계에서는 용어 t를 CoreNet개념 c_t로 분류한 다음($t \rightarrow c_t$), 두 번째 단계에서는 CoreNet개념 c_t와 가장 가까운 IT 개념 C를 찾아($c_t \rightarrow C_t$), 용어 t를 IT분야 상위 개념 C_t로 최종적으로 매핑한다.

첫 번째 단계에서는, 개념의 보편성 점수에 근거하여 용어를 분류한다. t로 Inspec 용어를, h_t로 t의 중심어를 표시하고, h_t가 m개의 CoreNet 개념 $\{c_1,\cdots c_j,\cdots c_m\}$에 대응된다고 가정한다. w_j로 c_j의 IT분야 보편성 점수를 나타내면, IT분야에서 t의 개념 c_t는 식 (2)에 의하여 분류된다:

$$c_t = c_{h_t} = \arg\max_{c}\{w_j \mid h_t \in c_j, 1 \le j \le m\} \quad (2)$$

식 (2)에서, 보편성 점수 w_j는 IT분야 상위 온톨로지 구축 과정에서 이미 획득한 것으로, 제3장에서 설명한 대로 개념 c_j에 분류된 IT분야 용어의 개수와 정비례하기 때문에, 이 단계에서의 개념 분류는 빈도수 기반 방법을 사용한 셈이다.

두 번째 단계에서는, 유사도 기반 방법으로 CoreNet 개념 c_t와 가장 가까운 IT 분야 상위개념 C_t을 찾아, 용어 t를 최종으로 IT분야 상위 개념으로 매핑한다(식 (3)).

$$C_t = C_{h_t} = \arg\max_{C} \sum_{i=1}^{n} Sim(c_t, C_i) \quad (3)$$

CoreNet에서 노드 c의 깊이를 $depth(c)$라고 하고, 최상위 노드의 깊이를 1로 하면, c_t와 노드 C_i 사이의 유사도는 C_i와 c_t 사이의 거리의 최대역수이다. 본 논문의 실험에서, c_t가 C_i의 하위 노드가 아니면 이 두 노드 사이의 유사도는 0으로 한다(식 (4)).

$$Sim(c_t, C_i) = \begin{cases} 0, \text{if } c_t \text{ is not hyponym category of } C_i \text{ in CoreNet;} \\ 1/(depth(c_t) - depth(C_i) + 1), \text{else.} \end{cases} \quad (4)$$

본 논문에서는, Inspec 용어에 대한 분류 대신 용어의 중심어에 대한 분류를 수행하였는데, 중심어 인식은 다음의 패턴을 적용하여 수행하였다.

패턴 1: <headword><prep.><otherword>,
<prep.> ∈{by, in, on, of, from, for, with, about}
예: *head*(learning by example) = learning

패턴 2: <headword>_<domain>,
<domain>은 해당 개념의 분야 정보를 나타낸다.
예: *head*(network_circuits) = network

패턴 3: <otherword>-<headword>
예: *head*(unsolicited_e-mail) = mail

패턴 4: <otherword&headword> ("&"는 해당 부호의 앞 뒤 단어 사이에 공백이 없이 연결된 경우를 표시한다.)
예: *head*(radiotelephony) = telephony

패턴 5: <otherword headword>
복합명사일 경우, 마지막 단어가 용어의 중심어다.
예: *head*(state estimation) = estimation

패턴 6: <headword>
하나의 단어로만 구성되었을 경우, 단어 자체가 중심어로 된다.
예: *head*(antenna) = antenna

5. 관계 분류

본 절에서는 그림 2에서의 관계 분류(A→B) 방법에 대하여 설명하고자 한다. 관계 분류에서, 훈련데이타가 부족할 경우 규칙 및 패턴기반 방법을 사용하였고, 훈련데이타가 어느 정도 축적된 후에는 지도식 방법을 도입하였다.

규칙 및 패턴기반 관계 분류에서는 우선 규칙기반 방법으로, BT-NT관계를 *isa* 관계와 *non-isa* 관계로 분류하였다[15]. 그 다음, 위의 과정에서 *non-isa* 관계로 분류된 관계들을 패턴기반 방법으로 온톨로지를 위한 의미관계로 분류한다. 서술의 간략함을 위하여 본 논문에서는 BT-NT/RT관계는 *btnt*(NT, BT)로, *isa* 관계는 *isa*(NT, BT)로, *non-isa* 관계는 *n-isa*(NT, BT)로 표기하기로 한다.

5.1 규칙기반 *Isa* 관계 분류

• 동일 중심어 규칙

BT/NT관계에서의 두 개념이 같은 중심어를 가지면

isa 관계로 설정한다. 다음은 개념의 영어 어휘 표현에 대하여 동일 중심어 규칙을 적용한 예이다:

isa(active antenna array, antenna array)

isa(elastic waves, waves)

• 중심어 관계의 이행 규칙

BT/NT관계에서, 두 개념의 중심어가 *isa* 관계를 이루면, 이 두 개념도 *isa* 관계를 가진다고 판단한다. 이는 *isa* 관계의 이행성을 이용한 규칙이다.

다음은 차세대 이동통신 분야의 주어진 BT/NT관계에서 서로 *isa* 관계를 이루는 중심어의 예이다:

isa(listings, programs)

isa(methods, theory)

위의 *isa* 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)

• 중심어의 다양성 포용 규칙

포괄적인 의미를 가지는 일부 개념의 하위 개념은 다양한 어휘표현을 가질 수 있다. 예를 들면 주어진 BT/NT관계하에서, NT "receivers", "antennas", "cameras", "tubes", "transmitters"는 BT "equipments"와 모두 *isa*관계를 이룬다. IT분야에서 이런 포괄적인 의미를 가지는 개념 어휘로 "equipments", "accessories", "applications" 등이 있다. 주어진 BT/NT관계에서, BT가 이런 다양성 포용 중심어를 가지는 경우, 이들의 관계는 *isa*관계로 될 가능성이 많다. 다음은 이 규칙을 적용한 예이다:

isa(radio receivers, radio equipments)

isa(antenna feeds, antenna accessories)

isa(radio tracking, radio applications)

이 규칙과 아래의 약자 허용 규칙은 분야 데이타에 대한 관찰에 근거하여 경험적으로 얻어졌다.

• 중심어의 약자 허용 규칙

일부 개념은 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)

5.2 패턴기반 의미관계 분류

본 단계에서는 규칙기반 *isa* 관계 분류에서 *non-isa*로 분류된 관계 트리플을 온톨로지를 위한 의미관계로 분류하는데, 우선 BT와 NT/RT에 대하여 개념 분류한 다음, 관계 패턴을 적용함으로 관계 분류를 수행한다.

제3장에서 설명한 바와 같이, IT 분야 상위 온톨로지의 의미관계에는 정의역과 치역이 정의되었고, 이런 관계 트리플은 관계 패턴으로 간주될 수 있다. 예를 들어, 주어진 BT-NT/RT 관계 btnt(bubble chambers, particle track visualisation)의 경우, NT/RT "bubble chambers"는 개념 Equipment로 분류되고, BT "particle track visualization"은 개념 Processing으로 분류된다. 그림 5의 관계 패턴으로부터 정의역 Equipment와 치역 Processing은 관계 *equipmentFor*를 가지는 것을 알 수 있다. 때문에 주어진 BT-NT/RT관계는 *equipmentFor*(bubble chambers, particle track visualization)로 분류된다.

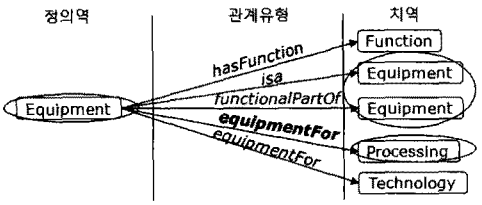

그림 5 정의역과 치역이 정의된 관계 패턴

동일 정의역과 동일 치역은 두 가지 이상의 관계유형을 가질 수 있다. 그림 5에서 주어진 정의역과 치역이 (Equipment, Equipment)인 경우, *isa*관계와 *functionalPartOf* 두 가지 관계가 가능한 것을 볼 수 있다. 이런 관계 애매성이 존재하는 경우, 주어진 BT-NT/RT관계에 가능한 모든 관계를 부여한다. 관계 유형이 많아질수록 관계 애매성 문제는 더 심각해지는데, 이를 해결하기 위하여 통계기반의 지도식 의미관계 분류 기법을 도입한다.

5.3 지도식 의미관계 분류

실험 데이타가 축적됨에 따라 지도식 분류를 위한 특징기반 분류 기법을 도입하였는데, 여기에서 각 관계 트리플은 하나의 이벤트(훈련데이타에서의 한 예)로 간주된다. 분야 시소러스가 제공하는 BT-NT/RT관계에서의 용어 쌍이 실제 같은 문장에 나타나는 용례를 찾기 어려웠고, 문맥 정보를 포함한 훈련데이타의 구축이 어려웠기에, 본 연구에서는 규칙 및 패턴 기반 방법에서 사용하였던 어휘 정보를 특징으로 활용하였다. 본 연구에서 사용한 기본 특징 정보는 다음과 같다:

- 중심어 특징: BT와 NT/RT의 중심어
- 이벤트가 동일 중심어 규칙을 만족 하는가?
- 이행성 규칙을 만족 하는가?
- 이벤트가 다양성 포용 규칙을 만족 하는가?
- 이벤트가 중심어 약자 허용 규칙을 만족시키는가?

위에서, BT와 NT/RT의 중심어 특징은 패턴기반 분

류에서의 관계 애매성 문제 해결을 위하여 추가로 사용한 특징이다. 나머지 특징들은 규칙기반 방법으로 *isa*와 *non-isa* 관계 분류 시 사용했던 특징으로, 주어진 관계가 *isa* 관계 여부에 대한 판단에 도움이 될 것으로 기대하였다.

이외에 비교 실험을 위하여 두 가지 특징을 추가로 사용하였다.

· 개념 특징: BT와 NT/RT의 개념 분류 카테고리

· *Isa* 특징: 규칙기반 방법으로 *isa*로 분류되는가?

개념 특징은 패턴기반 분류에서 사용한 특징을 반영한 것이고, 반면 *isa* 특징은 규칙기반 방법의 판단 결과를 직접 사용하는 것이다.

본 연구에서는 MEM기반 분류기를 사용하였고, 각 분류 대상 이벤트를 위한 훈련 데이타를 정제하기 위하여 k-NN방법을 사용하였는데, 전체 훈련 데이타에서 분류 대상 이벤트와 가장 유사한 k개의 이벤트를 훈련 데이타로 선정하는 방법이다. 유사 이벤트 추출을 위하여 코사인 유사도 방법을 사용하였는데, 유사도 계산에서도 위에서 제안한 특징 정보를 이용하였다.

6. 실험 및 평가

6.1 개념 분류 평가

적용률(coverage)과 정확도(accuracy)를 평가의 척도로 사용하였다. 적용률은 얼마나 많은 용어가 분야 상위 개념으로 분류되는지를 평가하기 위하여 사용되며(식 (5)), 정확도는 얼마나 많은 용어가 정확하게 분류되는지를 평가하기 위한 것이다(식 (6)).

$$\text{coverage}=\frac{\text{실제로 분류된 용어 수}}{\text{분류 대상 용어 수}} \quad (5)$$

$$\text{accuracy}=\frac{\text{정확히 분류된 용어 수}}{\text{실제로 분류된 용어 수}} \quad (6)$$

용어가 속한 개념은 중심어가 속한 개념과 같다는 가정하에, 22만개의 용어를 가진 IT 분야 전문용어 사전[13]에서 빈도수가 가장 높은 180개의 중심어를 평가 데이타로 사용하였다. 이 실험에서 78%의 적용률과 81%의 정확도를 얻을 수 있었다.

6.2 관계 분류 평가

관계 분류에서 개념 분류 결과를 사용하는데, 위에서 언급한 180개의 중심어에 대해서는 사람의 수정을 거친 개념 분류 결과를 적용하였고, 기타 중심어는 자동 개념 분류 결과를 적용하였다.

6.2.1 규칙기반 *Isa* 관계 분류에 대한 평가

본 실험에서는 우선 자동 분류의 정확도를 평가한 다음, 이를 수동 분류의 일관성과 비교 평가하였다.

자동 분류의 정확도에 대한 평가 실험에서, Inspec 시소러스의 12,821개 BT-NT/RT 관계를 5.1절에서 설명한 규칙 기반 방법으로 *isa* 관계와 *non-isa* 관계로 분류하였다. 본 실험에서는 모든 분류 대상에 대하여 관계 분류를 수행하였기에, 관계 분류의 적용률은 100%이다. 자동 분류의 정확도는 사람이 평가하는데 식 (6)에서 "용어 수"대신 "관계 수"를 대입하였다.

비교 평가를 위한 1차 베이스라인으로, 모든 관계 후보에 대하여 *isa*관계 부여시의 정확도를 취하였는데, 이는 실험데이타에서 *isa*관계의 비례와 동일하다. 5.1절에서 제안한 규칙을 평가하기 위하여 규칙을 순차적으로 적용하면서 정확도를 평가하였다(표 2).

표 2 규칙 기반 *isa* 분류 방법에 대한 평가

결과	접근법	특징	정확도
1	모두 *isa*	모두 *isa* 관계로 설정	74.21%
2	규칙1	동일 중심어 규칙	77.30%
3	+규칙2	+중심어 관계 이행 규칙	77.93%
4	+규칙3	+중심어의 다양성 포용 규칙	81.71%
5	+규칙4	+중심어의 약자 허용 규칙	82.02%

표 2에서 볼 수 있는 바와 같이, 규칙 기반 *isa* 관계 분류 방법으로 베이스라인보다 약 7.8%의 정확도 향상이 가능하였다. 본 논문에서 제안한 규칙 중, 중심어의 관계 이행 규칙이 많은 역할을 하지 못하였는데(결과 3과 2를 비교), 그 원인은 본 실험에서 중심어 사이의 *isa* 관계는 사람이 부분 데이타에 대한 관찰을 통하여 추가한 것으로, 해당 중심어 수는 92개 밖에 안되기에, 전체 데이타에서의 적용률이 낮았기 때문이다.

자동 분류 정확도를 수동 분류 일관성과 비교하기 위하여, 두 사람이 같은 분류 대상에 대하여 동시에 분류하되, 결과가 같은 경우를 정답으로 간주하여 평가하였다. 위의 12,821개 BT-NT/RT관계 중 임의로 선택된 675개의 관계에 대한 분류 평가 결과, 수동 분류의 일관성은 82.49%였다. 이는 표 2에서의 자동 분류 정확도 82.02%보다 미약하게 높은 수준이었다.

다만 두 실험의 데이타가 서로 다르기 때문에 직접 비교를 위하여, 위 675개 관계에 대한 자동 분류의 정확도를 따로 평가한 결과, 그 정확도는 83.41%로, 사람에 의한 분류 일관성보다 오히려 조금 높은 수준을 보여주었다. 분류 실험에 참여한 전문가가 해당 실험을 수행할 때까지 약 2-3개월간 관계 분류 작업을 수행하였으며, 각기 4,000~5,000개의 관계를 분류한 경험이 있었는데, 분류 참여자의 경험이 아직 많지 않아 분류 일관성에 영향을 주었던 것 같다. 또 다른 원인은, 자동 분류는 같은 규칙을 적용 시 늘 일관된 결과를 제공하는 반면, 수동 분류는 많은 훈련을 거친 전문가더라도 여전히 때와 분류대상에 따라 그 일관성이 영향 받기 때문인 것으로 간주된다.

6.2.2 패턴기반 의미관계 분류에 대한 평가

Inspec 시소러스의 BT-NT/RT관계 12,821개에 대하여 *isa* 관계 분류를 수행하고 사람이 1-3차례 검수한 결과, 최종적으로 3,307개의 *non-isa* 관계를 얻을 수 있었다. 이 3,307개의 *non-isa* 관계에 대하여 패턴기반 방법으로 의미관계 분류를 수행하였는데, 31.09%의 적용률과 약 90%의 정확도를 얻을 수 있었다. 제3장에서 설명하였듯이, 본 실험에서 사용한 패턴은 108개 관계 유형에 대한 258개의 관계 트리플인데, 이는 패턴기반 관계 분류의 목표 카테고리 수가 108개임을 뜻한다.

여기에서 정확도에 비하여 적용률이 많이 낮은데, 그 원인은, 많은 관계 유형 수에 비하여 정의역과 치역이 정의된 패턴 수가 너무 적기 때문이다. 실제 온톨로지 구축에서 관계 종류가 많고, 정의역과 치역이 다양하기에 패턴을 일일이 정의하기에는 어려움이 있다. 이 또한 통계기반 지도식 분류 기법을 도입해야 하는 한 원인이기도 하다.

6.2.3 지도식 의미관계 분류에 대한 평가

특징기반 분류 실험에서는 기존 MEM 툴킷[4][29]을 사용하였다. 규칙 및 패턴기반 방법에 의하여 분류되고 전문가에 의하여 검수된 14,730개의 의미관계 트리플(*isa* 관계 포함) 중 10%인 1,473개 관계 트리플을 실험 데이타로 사용하였고, 나머지 90%는 훈련데이타로 사용하였다. 훈련데이타에서 사용된 관계 종류는 모두 185가지로서, 이는 분류 목표 카테고리가 185개임을 뜻한다.

비교를 위한 1차 베이스라인으로, 모든 관계 후보에 *isa* 관계 부여시의 정확도를 취하였는데, 이는 실험데이타 중의 *isa* 관계 비례와 동일하다. 표 3에서 보여 부는 바와 같이, 이 1차 베이스라인의 정확도는 53.73%이다. 다음, 5.3절에서 설명한 기본 특징을 이용한 분류결과를 2차 베이스라인으로 취하였는데, 이의 정확도는 59.61%이다(결과 2).

표 3이 보여 주는 바와 같이, BT-NT/RT의 개념 분류 카테고리 정보는 관계 분류에 도움이 안 된 반면(결과 3과 2, 5와 4 비교), *isa* 특징과 k-NN기반의 훈련데이타 정제 기법은 관계 분류 정확도를 크게 향상시키는 것을 볼 수 있다(결과 4와 2, 6과 4 비교).

표 3 지도식 의미관계 분류 실험 결과

결과	접근법	특징	정확도
1	All *isa*	모두 *isa*로 설정	53.73%
2	MEM	기본 특징	59.61%
3	MEM	기본 특징+개념 특징	58.86%
4	MEM	기본 특징+*isa* 특징	62.46%
5	MEM	기본 특징+개념 특징+*isa* 특징	61.71%
6	MEM+k-NN	기본 특징+*isa* 특징	66.12%

4)http://homepages.inf.ed.ac.uk/s0450736/software/maxent

지도식 분류 기법으로 패턴기반 방법의 적용률이 낮은 문제를 해결할 수 있지만, 정확도는 많이 떨어지는 것을 볼 수 있다. 표 3의 결과 3(기본 특징 + 개념 특징)에 대하여 분석한 결과, 이 중 *isa*관계 분류의 정확도는 89.58%인 반면, 기타 의미관계의 정확도는 24.19%밖에 되지 않았다. 이를 패턴기반 방법의 90%에 달하는 정확도와 비교할 때, 패턴기반 방법에서의 목표 카테고리 수가 108개로서 지도식 방법에서의 목표 카테고리 수 185개 보다 현저히 적은 점을 고려하더라도, 여전히 낮다고 할 수 있다.

이런 낮은 정확도의 원인을 찾기 위하여 수동 분류 일관성 평가를 수행하였다. 일관성 평가는 식 (6)을 따르되, 다만 두 전문가가 같은 답을 주는 경우를 정답으로 간주한다. 실험 데이타로는 *isa* 관계를 제외한 90개의 의미관계 분류 대상을 임의로 선택하였는데, 이 실험 데이타에 대한 전문가들의 관계 분류 일관성은 15.87%인 반면 자동 분류 정확도는 14.44%로서 서로 비슷하였다. 이는 전문가들도 BT-NT/RT관계를 185개나 되는 목표 의미관계로 분류하는데 많은 어려움이 있는 것을 볼 수 있었다. 다시 말하면, 너무나 많은 관계 유형은 분류 정확도를 저하시키는 원인이다. 때문에, 실제 온톨로지 구축에서 지도식 의미관계 분류기법을 활용하기 위해서는 관계 유형 수를 대폭 줄여야 한다. 예를 들면, 본 연구에서 사용한 *eMemberPartOf, iMemberPartOf, functionalPartOf* 등 7개의 *partOf* 하위 관계는 그 상위인 *partOf* 관계로 통합 가능하고, *isTypeOf*는 *isa* 관계로 통합할 수 있다.

이외에 실험데이타에 대한 관찰에서, 185개 관계 유형 중 *isa, usedFor, partOf, isTechnologyOf* 등 출현 빈도가 100이상인 상위 6%의 관계 유형이 전체 관계 트리플 수의 87%이상을 차지하는 반면, 나머지 94%의 관계 유형은 13%의 관계 트리플에서만 사용되는 것을 볼 수 있었다. 출현 빈도가 낮은 관계 유형은 훈련 데이타가 부족하기에 지도식 기반 방법의 적용에서 정확률 저하의 원인이 되기에, 정확도 향상을 위해서는, BT-NT/RT관계를 185개 관계로 분류하는 다중 분류 문제를 (multi-class classification), 이진 분류 문제로 변환하고 (예: *isa* 및 *non-isa, usedFor* 및 *non-usedFor* 등), 출현 빈도가 높은 관계 유형에만 한정하여 관계 분류를 수행함으로써, 적용률의 적은 희생으로 상대적으로 높은 정확도를 기하여야 한다.

7. 결 론

본 논문에서는 분야 상위 온톨로지를 구축한 다음, 분야 시소러스의 개념과 관계에 대한 분류를 통하여 분야 상위 온톨로지를 확장함으로, 최종적으로 분야 온톨로지

를 구축하는 방법을 제안하였다. 분야 상위 온톨로지는 일반 분야 시소러스와 특정 분야 용어 사전을 이용하여 구축하였다. 분야 시소러스의 용어와 개념을 분야 상위 온톨로지의 개념으로 분류하고, 시소러스에서의 BT-NT/RT관계는 분야 상위 온톨로지에서 정의한 의미관계로 분류하였다. 개념 분류에서는 유사도와 통계 기반 방법을 적용하였고, 관계 분류에서는 훈련데이타가 부족한 경우를 위하여 먼저 규칙기반 방법으로 BT-NT/RT 관계를 *isa*와 *non-isa* 관계로 분류한 다음 패턴기반 방법으로 *non-isa* 관계를 온톨로지를 위한 의미관계로 분류하였다. 또한 훈련데이타를 충분히 축적한 경우를 위하여 어휘정보를 활용한 특징기반 의미관계 분류기법을 제안하였는데, k-NN기법으로 훈련데이타를 정제함으로 분류 정확도를 향상시켰다. 본 논문에서 제안한 방법으로 시스템을 구축하였는데, 그 성능이 사람에 의한 판단 결과와 견줄만한 수준이었다.

다만 *isa* 관계 이외의 기타 의미관계에 대한 분류 정확도는 여전히 매우 낮은데, 이는 IT 분야 상위 온톨로지에서 채택한 의미관계 수가 너무 많기 때문인 것으로 관찰되었다. 이런 문제를 해결하기 위하여 분야 상위 온톨로지의 상위 의미관계를 정의하고, 최종적으로 의미관계 분류체계를 구축하는 연구가 진행되고 있다. 의미관계 분류체계가 구축되면, 관계 분류에서 목표 카테고리를 상위 관계로 국한시킴으로써, 관계 유형의 수를 줄이고, 관계 분류의 정확도를 향상시킬 수 있을 것이다. 또 다른 해법으로 다중 관계 분류 문제를 이진 분류 문제로 변환하여, 출현 빈도수가 높은 관계 유형에 대해서만 관계 분류를 수행하는 것도 적은 적용률의 희생으로 높은 정확도를 얻는 방법이 될 것이다.

참 고 문 헌

[1] 최기선, 류법모, "온톨로지 구축과 학습: 상하위 관계", 정보과학회지, 24(4), 2006.4.

[2] 최호섭, 임지희, 배영준, 최수일, 옥철영, "온톨로지 구축 방법과 사례", 정보과학회지, 24(4), 2006.4.

[3] 고영만, "시소러스 기반 온톨로지에 관한 연구", 성균관대학교, 정보관리 제5집, 2006.

[4] Gruber, T.R., "A Translation Approach to Portable Ontology Specifications," Knowledge Acquisition, 5 (2), 1993.

[5] "Inspec v2.0 Getting Started Guide," http://scientific.thomson.com/media/scpdf/inspec_gettingstarted_en.pdf

[6] Soergel, D., B. Lauser, A. Liang, F. Fisseha, J. Keizer, S. Katz, "Reengineering Thesauri for New Applications: the AGROVOC Example," Journal of Digital Information, 4(4), Mar 2004.

[7] Kawtrakul, A., A. Imsombut, A. Thunkijjanukit, D. Soergel, A. Liang, M. Sini, G. Johannsen, J. Keizer, "Automatic Term Relationship Cleaning and Refinement for AGROVOC," Workshop on the 6th Agricultural Ontology Service, Jul 2005.

[8] Drummond, N., M. Horridge, R. Stevens, C. Wroe, S. Sampaio, "Pizza Ontology v1.5," http://www.co-ode.org/ontologies/pizza/, 2007.

[9] Navigli, R., P. Velardi, "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites," Computational Linguistics, 30 (2), 2004.

[10] "Summary Report on Taxonomic Databases Working Group(TDWG) Core Ontology Meeting," Edinburgh, UK, May 2006.

[11] Oberle, D., S. Lamparter, A. Eberhart, S. Staab, S. Grimm, P. Hitzler, S. Agarwal, R. Studer, "Semantic Management of Web Services using the Core Ontology of Services," W3C Workshop on Frameworks for Semantics in Web Services (Position Paper), 2005.

[12] Doerr, M., J. Hunter, C. Lagoze, "Towards a Core Ontology for Information Integration," In Journal of Digital information, 4(1), Apr 2003.

[13] KORTERM, http://korterm.or.kr/, IT분야 전문용어 사전.

[14] D. Sleeman, S. Potter, D. Robertson, and M. Schorlemmer, "Ontology Extraction for Distributed Environments," In Proceedings of Workshop on Knowledge Transformations for the Semantic Web (ECAI-02), Jul 2002.

[15] 황금하, 이신목, 남윤영, 신지애, 최기선, "시소러스를 이용한 온톨로지 구축에서의 Isa 관계 설정", 한국정보과학회 제 33회 정기 총회 및 추계학술대회 논문집, 서울, 2006.10.

[16] Assem, M.V., V. Malaisé, A. Miles, G. Schreiber, "A Method to Convert Thesauri to SKOS," In Proceedings in the 3rd European Semantic Web Conference, Jun 2006.

[17] Alani, H., "Ontology Construction from Online Ontologies," The 5th International Semantic Web Conference (Position paper), Nov. 2006.

[18] Golbeck, J., G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler, B. Parsia, "The National Cancer Institute's Thesaurus and Ontology," Journal of Web Semantics, 1 (1), Dec 2003.

[19] Wielinga, B., Schreiber, G., Wielemaker, J., & Sandberg, J.A.C., "From thesaurus to ontology," International Conference on Knowledge Capture, Oct 2001.

[20] Kang, S.J., J.H. Lee, "Semi-Automatic Practical Ontology Construction by Using a Thesaurus," Computational Dictionaries, and Large Corpora, Workshop on Human Language Technology and Knowledge Management (ACL2001), Jul 2001.

[21] Mika, P., D. Oberle, A. Gangemi, M. Sabou, "Foundations for service ontologies: Aligning owl-s to dolce," The 13th International World Wide Web Conference. 2004.

[22] A. Gangemi, F. Fisseha, J. Keizer, J. Lehmann, A. Liang, I. Pettman, M. Sini, M. Taconet, "A Core Ontology of Fishery and its Use in the Fishery Ontology Service Project," EKAW04 Workshop on Core Ontologies in Ontology Engineering, Oct 2004.
[23] Gangemi, A., P. Mika, M. Sabou, D. Oberle. "An Ontology of Services and Service Descriptions," Technical report, Laboratory for Applied Ontology (ISTC-CNR), 2003.
[24] Breuker, J., R. Hoekstra. "Epistemology and ontology in core ontologies: FOLaw and LRI-Core,. two core ontologies for law," EKAW04 Workshop on Core Ontologies in Ontology Engineering, Oct 2004.
[25] C. Baker, M. Ellsworth, K. Erk, "SemEval'07 Task 19: Frame Semantic Structure Extraction," The 4th International Workshop on Semantic Evaluations (SemEval-2007), Jun 2007.
[26] R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, D. Yuret, "SemEval-2007 Task 04: Classification of Semantic Relations between Nominal," In the Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Jun 2007.
[27] Choi, K.S., H.S. Bae, "Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy," The Global WordNet Conference, Jan 2004.
[28] 황금하, 이주호, 최기선, "소스-채널 모델을 이용한 한국어 전단어 의미태깅 시스템", 2004년도 한국인지과학회 춘계학술대회. 2004. 06.
[29] Zhang, L., "Maximum Entropy Toolkit for Python and C++," 2004.

황 금 하

1991년 중국 길림대학 물리학과 졸업(학사). 2000년 한국과학기술원 전산학과 졸업(공학석사). 2000년~현재 한국과학기술원 전산학과 박사과정. 1994년~1996년 중국 연변과학기술대학 전산실(직원) 2001년~2003년 Microsoft Research Asia(Assistant Researcher). 관심분야는 자연언어처리, 지식추출, 기계번역, 기계학습 등

신 지 애

정보과학회논문지 : 소프트웨어 및 응용
제 35 권 제 2 호 참조

최 기 선

정보과학회논문지 : 소프트웨어 및 응용
제 35 권 제 2 호 참조