분야 시소러스를 이용한 코아 온톨로지 확장

황금하, 신지애*, 최기선
한국과학기술원 전산학과 / 시맨틱웹첨단연구센터
*한국정보통신대학교

hgh@world.kaist.ac.kr, *jiae@icu.ac.kr, kschoi@cs.kaist.ac.kr

Enriching Core Ontology with Domain Thesaurus

Jin-Xia Huang, Ji-Ae Shin*, Key-Sun Choi
Computer Science Division/Semantic Web Research Center,
Korea Advanced Institute of Science and Technology
*Information and Communications University

hgh@world.kaist.ac.kr, *jiae@icu.ac.kr, kschoi@cs.kaist.ac.kr

요 약

본 논문에서는 분야 시소러스의 개념과 관계를 이용하여 코아 온톨로지를 확장하는 방법을 제안한다. 분야 시소러스의 개념을 코아 온톨로지의 상위 개념으로 분류하고, 시소러스에서의 광의어(Broader Term: BT)-협의어(Narrower Term: NT) 및 광의어-관련어(Related Term: RT)들 사이의 관계는 코아 온톨로지에서 정의한 의미관계로 분류한다. 유사도와 빈도수 기반의 방법으로 개념 분류를 수행하였고, 관계 분류에서는 두 가지 방법을 적용하였는데, (i) 훈련데이터가 부족한 경우를 위하여 규칙기반 방법으로 BT-NT/RT관계를 isa와 기타 관계(non-isa관계)로 분류하고, 패턴기반 방법으로 non-isa관계를 온톨로지를 위한 의미관계로 분류한다. (ii) 훈련데이터를 충분히 가지고 있을 경우, 최대 엔트로피 모델(MEM)을 적용한 분류 방법을 사용하되, kNN방법으로 훈련데이터를 정제하였다. 본 논문에서 제안한 방법으로 시스템을 구축하였고, 실험 결과, 시스템 성능이 사람에 의한 판단 결과와 비교 가능한 수준이었다.

1. 소개

온톨로지는 해당 분야 개념, 인스턴스, 관계, 추론 규칙인 공리(axiom) 등 정보를 제공한다. 시소러스를 포함한 기존의 지식베이스도 개념, 인스턴스, 관계 등 의미정보를 부분적으로 포함하고 있다. 때문에 온톨로지 구축에 있어서 시소러스는 자주 이용되고 있다.

그러나, 시소러스의 계층(hierarchy) 관계에는 상위어와 하위어간의 isa관계와 부분-전체 관계가 포함되어 있을 뿐만 아니라 BT-NT간, 또는 BT-RT간의 다양한 관계도 포함되어 있으며, 이런 관계가 구체적으로 어떤 의미관계인지는 시소러스에서 명기하지 않고 있다. 이런 문제는 분야 시소러스에서 더욱 심각한데, 예를 들면, 분야 시소러스 Inspec의 계층 관계에는 isa와 기타 비-상하위(non-taxonomic) 관계가 혼재하여 있다(그림 1). 엄격한 상하위 계층 구조(taxonomic hierarchy)와 구별하기 위하여, 이런 분야 시소러스의 계층 구조를 우리는 BT-NT/RT계층 구조(BT-NT/RT hierarchy)라고 부르기로 한다.

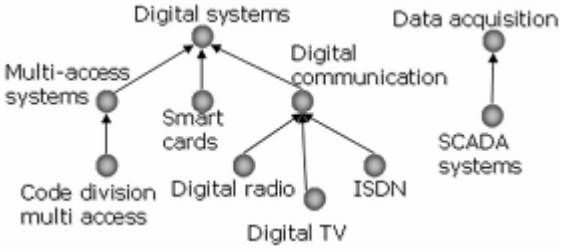

그림 1. 분야 시소러스 Inspec에서의 BT-NT/RT관계로 구성된 계층구조

분야 시소러스를 이용하여 온톨로지를 구축하기 위해서는, BT-NT/RT계층 구조는 상하위 계층 구조로 변환되어야 하고, BT-NT/RT관계는 의미관계로 분류되어야

한다. 그림 2에서, A부분이 나타내고 있는 분야 시소러스에서의 BT-NT/RT관계는, B부분이 보여 주는 바와 같이, 온톨로지 의미관계로 분류된다(관계 분류). 관계 분류 후 의미 관계에서의 NT가 상위어 없이 최상위 개념으로 남는 것을 방지하기 위하여, B부분의 BT와 NT/RT는 C와 D부분에서 보여 주는 바와 같이, 코아 온톨로지의 상위 개념(의미 카테고리)으로 분류된다(개념 분류). 이런 개념 및 관계 분류를 통하여 시소러스에서의 BT-NT/RT계층 구조는 온톨로지를 위한 상하위 계층 구조로 변환된다. 이를 위하여 개념 분류가 필요하고, 그 분류 목표 카테고리로는 시소러스의 자체 상위 개념이 아닌 온톨로지의 상위 개념을 이용한다. 여기에서 온톨로지 상위 개념을 이용하는 원인은, 기존 분야 시소러스는 일반적으로 목표 분야 온톨로지와 완전히 일치한 분야가 아니거나, 규모가 아주 작은 시소러스로서, 이의 상위 개념이 목표 온톨로지가 표현하고자 하는 분야를 대표하기에 역부족인 경우가 많기 때문이다. 이런 원인으로 인하여, 분야 온톨로지의 상위 개념과 의미 관계를 우선 정의하여야 하고, 다음 시소러스의 개념을 온톨로지의 상위 개념으로 분류하는 작업이 필요하게 된다.

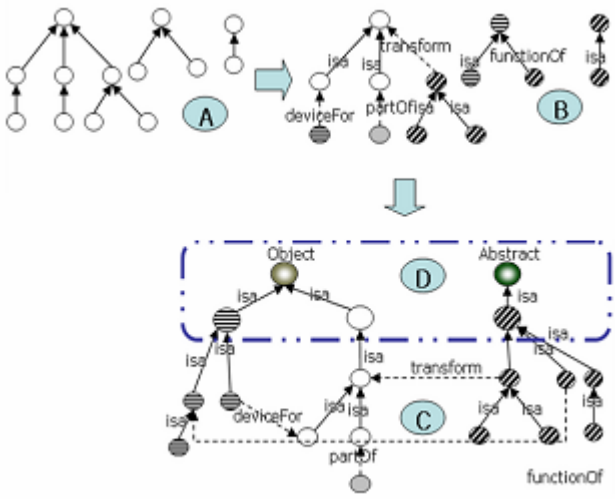

그림 2.시소러스에 대한 관계 분류(A→B) 및 개념 분류(B→C, D)를 통한 온톨로지 확장

본 논문에서 사용한 IT 코아 온톨로지는 CoreNet에서[2] IT분야와 관련된 200개의 상위 개념으로 구성된 개념 분류체계와, IT분야를 위해 제안된 185개의 의미관계로 구성되어 있다. 즉 IT 코아 온톨로지는, 상위 개념과 의미관계가 이미 정의된 작은 규모의 분야 핵심 온톨로지이다. 이 코아 온톨로지를 확장하기 위하여, 분야 시소러스 Inspec을 주로 사용하였다. 개념 분류를 위해 코아 온톨로지의 상위 개념을 의미 카테고리로 간주하고, Inspec의 개념과 용어들을 개념 분류를 통하여 의미 카테고리로 분류하였다. 이러한 개념 분류에서는 유사도 및 통계기반 방법을 이용하였다.

관계 분류에서는 Inspec 시소러스의 BT-NT/RT관계를 코아 온톨로지의 특정한 의미 관계로 분류하였다. 이를 위하여 훈련데이터 부재 시를 위한 비지도식 방법과, 충분한 훈련데이터를 확보한 후의 지도식 분류 방법을 제안하였다. 비지도식 관계 분류에서는 우선 BT-NT/RT관계를 개념 어휘 정보를 이용한 규칙기반 방법으로 isa와 non-isa관계로 분류한다. 다음, non-isa 관계를 패턴기반 방법으로 의미관계로 분류한다. 지도식 방법에서는 이미 분류된 의미관계 트리플(triple)을 훈련데이터로 사용하였고, 최대 엔트로피 모델(MEM)을 이용하여 BT-NT/RT관계를 isa관계를 포함한 의미관계로 분류하였다. 훈련데이터에서 노이즈를 제거하기 위하여 kNN기반 접근법을 사용하였다.

2. 관련 연구

지난 몇 년간 기존 지식 베이스를 이용한 온톨로지 구축 방법에 대한 연구가 꾸준히 진행되어 왔다. 시소러스와 같은 기존 지식 베이스는 해당 분야에서 자주 사용되는 전문용어 및 개념 정보를 가지고 있을 뿐만 아니라, BT와 NT/RT간의 계층 관계 정보도 제공한다. 이들 중 일부 지식 베이스는 추론에 사용되는 제약조건(constraint) 정보를 제공하기도 한다[5]. 그러나 기존의 지식베이스에서는 대개 온톨로지 구축에 필요한 정보를 부분적으로만 제공하고 있다. 때문에 분야 시소러스를 온톨로지 구축에서 활용하기 위한 연구에 대한 필요성이 제기 되어 왔다[12, 13].

지식 베이스로 온톨로지를 구축하는 연구에서, 어떤 이들은 기존 지식 베이스를 온톨로지 포맷으로 변환하는데 주력하고 있다[6]. 이런 연구에서는 새로운 온톨로지 지식의 생성이 없이, 기존의 시소러스 포맷을 RDF나 OWL과 같은 온톨로지 표현들로 변환한다. 이런 연구에서는 각각의 시소러스 표현 방식에 대하여 조사 연구 후, 패턴기반이나 규칙기반 방법으로 온톨로지로 변환한다. 또 일부 연구에서는 기존 지식 베이스로부터 유용한 정보를 추출하여 온톨로지 지식으로 변환해 준다[5, 7-8, 10]. 이런 연구에서는 기존의 논리 프로그램으로부터 제약 조건을 추출하여 온톨로지 지식으로 변환해 주고 있다.

기타 연구에서는 관계 정보를 확장함으로써 시소러스를 온톨로지로 리모델하고 있다[1, 8-9]. 이 중 일부는 [11, 7] 격 관계(case relation)와 의미관계를 시소러스의 상하위 계층 구조에 추가함으로 시소러스를 온톨로지로 확장한다. 격 관계는 기존 기계번역 시스템과 사전으로부터 얻어지며, 의미관계는 말뭉치의 상관관계 정보를 이용하여 획득한다. 다른 연구[1, 9]에서는 시소러스의 BT-NT/RT관계를 사람에 의하여 정의한 규칙이나 패턴을 이용하여 의미 관계로 분류하기도 한다.

관계 분류를 통하여 시소러스로 온톨로지를 구축하는 면에서, [9, 1]의 연구는 본 논문과 비슷하다고 볼 수 있다. 그러나 본 과제의 연구대상인 IT 분야와 Inspec 시소러스는 기존 연구보다 훨씬 넓은 분야에 대한 연구

로, 패턴을 수동적으로 정의하기엔 어려움이 따른다. 실제로 본 논문에서는 180여 개의 의미 관계가 정의되어 있다. 이런 문제를 해결하기 위하여 본 논문에서는 규칙과 패턴기반 관계 분류 외에 지도식 관계 분류 방법을 제안하였으며, 실험에서 좋은 성능을 볼 수 있었다. 본 과제가 기존 연구와 의 또 다른 차이는, 본 논문에서는 관계 분류 뿐만 아니라, 개념 분류도 수행함으로, 분야 시소러스의 BT-NT/RT계층을 분야 온톨로지를 위한 상하위 계층으로 변환시키고 있다는 점이다.

3. IT 코아 온톨로지와 분야 시소러스 INSPEC

본 논문에서는 **IT 코아 온톨로지**를 확장 대상 온톨로지로 설정하였다. IT 코아 온톨로지는 IT분야 상위 분류체계(top-level taxonomy)와 185개의 의미관계들로 구성되었다. IT분야 상위 분류체계는 200개의 IT분야 최상위 개념들과 이들 개념 사이의 상하위 계층관계를 포함하고 있으며, 이 분류체계는 일반 분야 시소러스인 CoreNet의 일부이기도 한데, CoreNet에는 2,900여 개의 개념(카테고리)와 50,000여 개의 한국어 상용 어휘를 포함하고 있다[2]. IT분야 최상위 개념은 CoreNet 개념 중 IT 분야에서 보편적으로 자주 사용되는 개념을 선택한 것인데, 이들 개념이 해당 분야에서 자주 사용되는 정도를 보편성(popularity)으로 설명한다면, CoreNet개념 "통신기기"는 IT분야에서도 보편성이 높은 개념이기에 IT분야 상위 개념으로 선택되었고, 반면에, "의약품"은 IT분야에서의 보편성이 낮기에 IT 분야 상위 분류체계에서 배제되었다. 그림 3은 일반 분야 시소러스인 CoreNet의 일부로서, CoreNet 개념 "인공물"의 하위 트리 구조를 나타내고 있다. 그림에서 회색 노드는 CoreNet 개념으로서 IT분야 분류체계에도 선택된 개념이고, 흰색 노드는 CoreNet개념이지만 IT분야 분류체계에서는 선택되지 않은 개념이다.

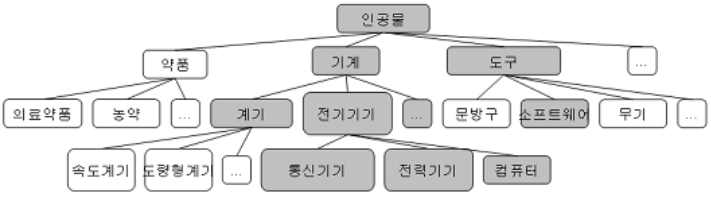

그림 3. IT 코아 온톨로지의 상위 분류체계(회색 노드로 구성된 계층구조)

IT 코아 온톨로지의 의미 관계는 정의역(domain)과 치역(range)을 제약으로 가지고 있다. 표 1에서, 관계 *theoryAbout*의 정의역은 *Theory*이고, 치역은 *Structure* 나 *Equipment* 두 가지 모두 가능하다. 정의역과 치역으로 제약된 이런 관계 트리플은 관계 분류 패턴으로 사용할 수 있다. IT 코아 온톨로지는 여전히 개발 중이므로, 현재는 의미관계들의 일부에만 정의역과 치역이 정의되어 있는바, 총 185개의 의미관계 중 108개의 의미관계에 대하여 258개의 관계 트리플이 정의되어 있다.

온톨로지 확장에 사용될 분야 시소러스로는 **Inspec 시소러스**[3]를 선정하였다. Inspec 시소러스는 컴퓨팅, 제어공학, 전자 전기공학, 정보기술 및 물리학 등 약 14개 분야를 어우르고 있다. Inspec은 8,300개 이상의 용어와 15,901개의 BT-NT/RT 관계들을 포함하며, 관계들은 그림 1에서 볼 수 있듯이 BT-NT 와 BT-RT 관계들이 혼재되어 있고, 구체적인 관계 유형은 명기되지 않고 있다.

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

표 1. IT 분야 온톨로지에서 정의된 의미관계

4. 개념 분류

본 절에서는 그림 2에서의 개념 분류(B→C, D) 방법에 대하여 설명하고자 한다. 개념 분류에서는 Inspec 용어를 IT 코아 온톨로지의 200개 카테고리로 분류한다. 그 첫 단계는 빈도수 기반의 접근 방법으로, 각 용어들을 보편성 점수(popularity score)에 근거하여 카테고리로 분류하게 된다. t 로 Inspec 용어를, h_t 로 t의 중심어를 표시하고, h_t 가 m개의 의미를 가지고 있는데 이런 의미들은 CoreNet 개념 $\{c_1,\cdots c_j,\cdots c_m\}$에 각각 대응된다고 가정한다. w_j로 c_j의 IT분야에서의 보편성 점수를 나타내고, IT분야에서 t의 소속 개념 c_t는 공식 (1)에 의하여 분류된다:

$$c_t = c_{h_t} = \arg\max_c \{w_j \mid h_t \in c_j, 1 \le j \le m\} \qquad (1)$$

위의 공식에서, 보편성 점수 w_j는 IT코아 온톨로지의 상위 분류체계를 구축하는 과정에서(그림3) 이미 획득한 점수로, 카테고리 c_j가 포함하고 있는 IT분야 용어의 개수와 정비례한다.

두 번째 단계는 유사도 방법으로 CoreNet 개념 c_t와 가장 가까운 IT 분야 분류체계에서의 상위개념을 찾는다. IT분야 분류체계의 개념 집합을 $C=\{C_1,\cdots C_i,\cdots C_n\}$로 표현할 경우 앞에서 설명한 바와 같이 n=200이고, C는 CoreNet 개념 집합의 분류집합이다. IT분야 분류체계에서의 t의 의미 카테고리는 식 (2)에 의해 분류된다.

$$C(t) = C(h_t) = \arg\max_C \sum_{i=1}^{n} Sim(c_t, C_i) \qquad (2)$$

CoreNet에서 노드 c의 깊이를 *depth*(c)라고 하고, 톱 노드의 깊이는 1, 그 하위는 2면, CoreNet에서 노드 c의 깊이라고 하면, c_t와 카테고리 C_i 사이의 유사도는 C_i와 c_j 사이의 거리의 최대역수이다. 본 논문의 실험에서, c_t가 카테고리 C_i의 하위 카테고리가 아니면 유사도는 0으로 한다. (식 3)

$$Sim(c_t, C_i) = \begin{cases} 0, \text{if } c_t \text{ is not hyponym category of } C_i \text{ in CoreNet;} \\ 1/(depth(c_t) - depth(C_i) + 1), \text{else.} \end{cases} \quad (3)$$

우리는 Inspec 용어들에 대한 분류 대신 용어 중심어에 대한 분류를 실행하였는데 용어 t 의 중심어 h_t의 인식에서는 다음과 같은 패턴을 적용하여 중심어 인식을 수행하였다(*head*(term)은 중심어 인식 함수이다):

- <headword><prep.><otherword>,
  - 위에서 <prep.>={by, in, on, of, from, for, with, about}
  - Ex) *head*(learning by example) = learning
- <headword>_<domain>,
  - 위에서 <domain>은 해당 개념의 분야 정보를 나타낸다.
  - Ex) head(network_circuits) = circuits
- <otherword>-<headword>
  - Ex) head(unsolicited_e-mail) = mail
- <otherword&headword>,
  - 위에서 '&'는 해당 부호의 앞 뒤 단어 사이에 공백이 없이 연결된 경우를 나타낸다.
  - Ex) *head*(radiotelephony) = telephony
- <otherword headword>
  - 용어가 복합명사일 경우, 마지막 단어가 중심어이다.
  - Ex) head(state estimation) = estimation
- <headword>
  - 용어가 하나의 단어로만 구성되었을 경우, 이 단어 자체가 중심어이다.
  - Ex) *head*(antenna) = antenna

5. 관계 분류

본 절에서는 그림 2에서의 관계 분류(A→B) 방법에 관하여 설명하고자 한다. 관계 분류에서, 이미 분류된 관계 훈련데이터가 부족할 경우 비지도식 방법을 사용하였고, 훈련데이터가 어느 정도 축적된 후에는 지도식 방법이 도입되었다.

비지도식 관계 분류에서는 우선 규칙기반 방법으로, BT-NT관계를 isa관계와 non-isa관계로 분류하였다. 여기에서 우리는 [13]에서 제안된 동일 중심어 규칙, 중심어 관계의 이행규칙, 중심어의 다양성 포용 규칙 및 중심어의 약자 허용 규칙을 적용하였다. 그 다음, 위의 과정에서 non-isa관계로 분류된 관계들을 패턴기반 방법으로 의미관계로 분류한다.

본 논문에서는 규칙 기반 isa관계 분류 부분은 생략하고, 패턴기반 의미 관계 분류 부분에 대해서 설명하고자 한다. 또한 본 논문에서는 BT-NT/RT관계는 *btnt*(NT, BT)로, isa관계는 *isa*(NT, BT)로, non-isa관계는 *n-isa*(NT, BT)로 표기하기로 한다.

5.1 패턴기반 의미 관계 분류

규칙기반 isa관계 분류에서 BT-NT/RT관계들은 isa나 non-isa관계로 분류된다. 이 단계에서는 non-isa관계를 코아 온톨로지의 의미관계로 분류한다. 의미관계는 두 단계로 분류되는데, 우선 첫 단계는 개념 분류 단계로서 BT와 NT/RT를 IT분야 상위 개념으로 분류한다. 다음 두 번째 단계에서는 관계패턴을 이용하여 관계 분류를 진행한다.

앞에서 설명된 바와 같이, IT 코아 온톨로지의 의미관계는 정의역과 치역이 정의되었고, 이런 관계는 관계패턴으로 간주될 수 있다. 예를 들어, 주어진 BT-NT/RT 관계 *btnt*(bubble chambers, particle track visualisation)의 경우, NT/RT "bubble chambers" 는 *Equipment* 카테고리로 분류되고, BT "particle track visualization"은 *Processing* 카테고리로 분류된다. 그림4의 관계 패턴으로부터 정의역 *Equipment*와 치역 *Processing*을 가진 관계는 *equipmentFor*라는 것을 알 수 있다. 때문에 주어진 BT-NT/RT관계는 *equipmentFor*(bubble chambers, particle track visualization)로 분류된다.

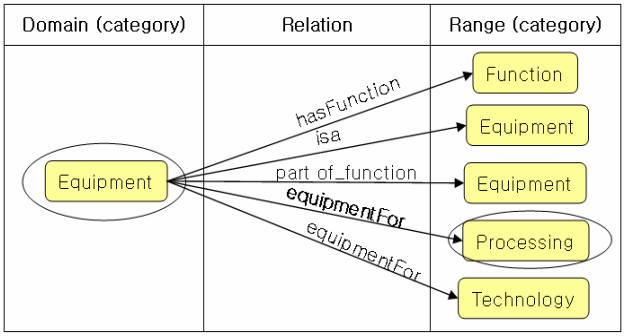

그림 4. 패턴기반 분류에서 의미관계는 관계 분류 패턴으로 사용 한다

동일 정의역과 동일 치역 사이에 두 가지 이상의 관계가 존재할 수도 있다. 그림4에서 주어진 정의역과 치역이 (*Equipment*, *Equipment*)인 경우, *isa*관계와 *part_of_function* 두 가지 관계가 가능한 것을 볼 수 있다. 이런 관계 애매성이 존재하는 경우, 패턴기반 관계 분류에서는 주어진 BT-NT/RT관계에 가능한 모든 관계를 부여하였다.

그러나 이러한 관계 애매성은 관계 수가 많을수록 더 심각해 졌는데, 이를 해결하기 위하여 지도식 의미관계 분류 방법을 도입하였다

5.2 지도식 의미관계 분류

실험 데이터가 축적됨에 따라 지도식 분류를 위한

MEM기반 분류 방법이 도입되었다. 각 관계 트리플을 하나의 이벤트로(훈련데이터에서의 한 예) 간주하였고, BT 및 NT/RT의 어휘 정보를 특징으로 사용하였다. 용어의 어휘정보만 특징으로 사용한 원인은 말뭉치에서 BT와 NT/RT 모두를 포함하는 용례를 찾기 어려웠기 때문이다.

특징 추출에서 이벤트의 기본 특징 정보는 다음과 같다:

1) BT와 NT/RT의 중심어
2) 이벤트가 동일 중심어 규칙을 만족 하는가: 예이면 값은 1; 아니면 0.
3) 이벤트가 이행성 규칙을 만족 하는가: 예이면 값은 1; 아니면 0.
4) 이벤트가 다양성 포용 규칙을 만족 하는가: 예이면 값은 1; 아니면 0.
5) 이벤트가 중심어 약자 허용 규칙을 만족시킨다면: 예이면 값은 1; 아니면 0.

이외에 비교 실험을 위하여 두 가지 특징이 추가로 사용되었다:

6) 카테고리특징: BT와 NT/RT의 개념 분류 카테고리
7) Isa 특징: 규칙기반 방법으로 isa로 분류되면 값은 1, 아니면 0.

위에서, 카테고리 특징(특징 6)은 패턴기반 분류에서 사용한 특징을 반영한 것이고, BT와 NT/RT의 중심어 특징(특징 1)은 패턴기반 분류에서의 관계 애매성 문제 해결을 위하여 추가로 사용한 특징으로 볼 수 있다. 또한 특징 2)~5)는 규칙 기반 방법으로 isa와 non-isa관계 분류 시 사용했던 특징으로, 주어진 관계가 isa관계인지 여부에 도움이 될 것으로 기대하였다. 반면에 Isa 특징(특징 7)은 규칙 기반 방법의 판단 결과를 직접 특징으로 사용하는 것으로 비교 목적으로 사용하였다.

분류 대상 이벤트에 대한 훈련 데이터 구축에서는 kNN방법으로 훈련 데이터를 정제하였는데, 전체 훈련 데이터에서 분류 대상 이벤트와 가장 유사한 k개의 이벤트를 훈련데이터로 선정하였다. 유사 이벤트 추출을 위하여 코사인 유사도 방법을 사용하였는데, 유사도 계산에서는 위에서 제안한 특징 정보를 이용하였다.

6. 실험 및 평가

개념 분류 평가에서는 적용률(coverage)과 정확도(accuracy)를 사용하였다. 적용률은 얼마나 많은 용어가 코아 온톨로지의 의미 카테고리로 분류되는지를 평가하기 위해 사용되며, 정확도는 얼마나 많은 용어가 정확하게 그것이 속해야 할 카테고리로 분류되는지를 평가하기 위한 것으로 전문가에 의해 행하여진다.

용어가 속한 개념은 해당 용어의 중심어가 속한 개념과 같다는 가정하에, 22만개의 용어를 가진 IT 분야 사전에서 빈도수가 가장 높은 180개의 중심어를 실험 데이터로 사용하였다. 이 실험에서 우리는 78%의 적용률과 81%의 정확도를 얻을 수 있었다.

패턴기반 의미관계 분류에서는 개념 분류 결과가 필요적이다. 그러나 지도식 의미관계 분류에서는 그렇지 않기에 개념 분류가 관계 분류에 주는 영향을 “카테고리 특징”으로 평가하였다(표2 참조). 관계 분류에서 사용한 개념 분류 결과는 전문가의 검증을 거치지 않은 자동 개념 분류 결과를 직접 사용하였다. 다만 위에서 언급한 180개의 중심어에 대해서는 이미 사람의 수정을 거친 개념 분류 결과를 적용하였다.

6.1 패턴기반 의미 관계 분류에 대한 평가

Inspec 시소러스의 BT-NT/RT 12,821개 관계들에 대하여 isa관계 분류를 진행한 결과 3,307개의 non-isa 관계를 얻을 수 있었다. 이 3,307개의 non-isa관계에 대하여 패턴기반 방법으로 의미관계 분류 진행한 결과, 31.09%의 적용률과 약 90%의 정확도를 얻을 수 있었다. 정확도 평가는 공식 (4)를, 적용률 평가 공식은 공식 (5)를 따랐다.

$$Accu = \frac{|R1 \cap R2|}{|R2|} \quad (4)$$

공식 (4)에서, R1은 자동적으로 분류되는 관계, R2는 전문가가 결정하는 관계이다.

공식 (5)에서, 적용률은 non-isa관계 개수를 분모로, 이 중 특정된 의미관계로 분류된 관계 개수를 분자로 한다.

$$\text{Coverage} = \frac{|\text{Identified NotISA relations}|}{|\text{NotISA relations}|} \quad (5)$$

6.2 지도식 의미관계 분류에 대한 평가

MEM기반 분류 실험에서는 기존 MEM 툴킷 [4]를 사용하였다. 비지도식 방법에 의하여 분류되고 전문가에 의하여 검수된 14,730개의 의미관계 트리플(isa관계 포함) 중 10%인 1,473개 관계 트리플을 실험 데이터로 사용하였고, 나머지 90%는 훈련데이터로 사용하였다. 훈련데이터에서 사용된 관계 종류는 모두 185가지로서, 이는 분류 대상 카테고리가 185개임을 뜻한다.

실험 결과, 표 2가 보여 주는 바와 같이, BT-NT/RT의 개념 분류 카테고리 정보는 관계 분류에 도움이 안 된 반면, kNN기반의 훈련데이터 정제 방법은 관계 분류 정확도를 현저하게 향상시키는 것을 볼 수 있다.

지도식 분류 방법은 패턴기반 방법에서의 낮은 적용률을 극복 할 수 있었지만, 정확도는 많이 떨어지는 것을 볼 수 있었다. 표2의 두 번째 행(기본 특징 + 카테고리 특징)의 실험 결과를 분석한 결과, 이 중 isa관계 분류의 정확도는 89.58%인 반면, 기타 의미 관계의

정확도는 24.19%밖에 되지 않았다. 이 정확도는 패턴 기반 방법의 90%에 달하는 정확도와 비교하면, 패턴 기반 방법에서는 관계 분류 목표 카테고리가 지도식 방법에서의 카테고리 수보다 거의 절반 적은 점을 고려하더라도, 여전히 너무나 낮다고 봐야 한다.

접근법	특징	정확도
MEM	기본 특징	59.61%
MEM	**기본 특징 + 카테고리 특징**	**58.86%**
MEM	기본 특징 + Isa 특징	62.46%
MEM	기본 특징 + 카테고리 특징 + Isa 특징	61.71%
MEM+kNN	**기본 특징 + Isa 특징**	**66.12%**

표2. 지도식 의미관계 분류 실험 결과

이런 낮은 정확도의 원인을 찾기 위하여 전문가에 의한 의미 관계 분류 결과에 대하여 일관성 평가를 진행하였다. 일관성 평가는 공식 4)를 따르되, 다만 R1와 R2를 전문가1과 전문가2가 분류한 관계로 적용하였다. 실험 데이터로는 isa관계를 제외한 90개의 의미 관계를 임의로 선택하였는데, 이 실험 데이터에 대한 전문가들의 관계 분류 일관성은 15.87%로 나온 반면 자동시스템의 정확도는 14.44%로서, 이는 전문가의 일관성보다 약간 낮은 수치인 정도이다. 이 결과로부터, 전문가들도 BT-NT/RT관계를 185개나 되는 의미관계로 분류하는데 있어서 많은 어려움을 느끼는 것을 볼 수 있었다. 관계 유형이 급격히 증가 하는 것이 관계 분류의 효율성과 정확도를 현저히 저하시키는 원인이 되고 있다.

7. 결론

본 논문은 분야 시소러스의 개념과 관계에 대한 분류를 통하여 코아 온톨로지를 확장하는 방법을 제안하였다. 분야 시소러스의 개념을 코아 온톨로지의 상위 개념(top-level concept), 즉 의미 카테고리로 분류하고, 시소러스에서의 BT-NT 및 BT-RT 사이의 관계는 코아 온톨로지에서 정의한 의미관계로 분류한다. 개념 분류에서는 유사도와 통계 기반의 방법을 제안하였다. 관계 분류에서는 비지도식 방법과 지도식 방법 을 적용하였다. 비지도식 방법에서는, 훈련데이터가 부족한 경우를 위하여 규칙기반 방법으로 BT-NT/RT관계를 isa와 non-isa관계로 분류하고, 패턴기반 방법으로 non-isa 관계를 온톨로지에서의 의미관계로 분류하였다. 지도식 방법에서는 훈련데이터를 충분히 가지고 있는 경우에 한하여, 최대 엔트로피 모델(MEM)을 적용한 분류 방법을 사용하였다. 특정된 관계에 대하여 정제된 훈련데이터를 추출하기 위하여 kNN방법을 사용한 결과 정확도 향상에 많은 기여를 하는 것을 볼 수 있었다. 본 논문에서 제안한 방법으로 시스템을 구축하고 실험한 결과, 시스템 성능이 사람에 의한 판단 결과와 비교 가능한 수준을 보여 주었다.

그러나 isa관계 이외의 기타 의미 관계의 분류 정확도는 여전히 매우 낮은데, 이는 IT 코아 온톨로지에서 사용하기로 한 의미관계가 너무나 많기 때문인 것으로 드러났다. 또한 관계 분류를 위하여 사용된 특징이 주로 어휘 정보에만 국한된 것도 하나의 원인으로 간주된다.

관계 분류를 위하여 말뭉치에서의 문맥 정보를 어떻게 발굴하고 활용하는지가 다음 과제로 남아 있다. 현재 진행 중인 또 하나의 중요한 과제는 코아 온톨로지의 의미 관계 계층 구조(relation hierarchy)를 구축하는 작업이다. 의미 관계 계층 구조가 구축되면, 관계 분류에서 목표 카테고리를 상위 관계로 국한시킴으로, 관계 분류의 정확도를 향상시킬 수 있기를 기대하고 있다.

감사의 글

본 논문은 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참고 문헌

[1] Dagobert Soergel, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer and Stephen Katz. Reengineering Thesauri for New Applications: the AGROVOC Example. Journal of Digital Information, 4(4), March 2004.

[2] Key-Sun Choi, Hee-Sook Bae, Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy , In Proceedings of the Global WordNet Conference, pp. 320~325, 2004.1, Brno, Czech.

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

[4] Le Zhang. 2004. Maximum Entropy Toolkit for Python and C++. Available from http://homepages.inf.ed.ac.uk/s0450736/software/maxent/manual.pdf

[5] 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 (affiliated to ECAI-02), July 2002

[6] Mark van Assem, Véronique Malaisé, Alistair Miles, and Guus Schreiber: A Method to Convert Thesauri to SKOS. In Proceedings in the 3rd European Semantic Web Conference, June 2006, pp. 95-109

[7] Harith Alani, Position paper: Ontology Construction from Online Ontologies. In Proceedings of the 5th International Semantic Web Conference,

November 2006
[8] Golbeck, Jennifer, Gilberto Fragoso, Frank Hartel, Jim Hendler, Jim Oberthaler, Bijan Parsia "The National Cancer Institute's Thesaurus and Ontology," Journal of Web Semantics, 1(1), December 2003.
[9] Asanee Kawtrakul, Aurawan Imsombut, Aree Thunkijjanukit, Dagobert Soergel, Anita Liang, Margherita Sini, Gudrun Johannsen, and Johannes Keizer, Automatic Term Relationship Cleaning and Refinement for AGROVOC, Workshop on The Sixth Agricultural Ontology Service, July 25-28, 2005. Vila Real, Portugal.
[10] Wielinga, B., Schreiber, G., Wielemaker, J., & Sandberg, J.A.C. From thesaurus to ontology. In Proceedings of International Conference on Knowledge Capture, Victoria, Canada, October 2001
[11] Sin-Jae Kang and Jong-Hyeok Lee, Semi-Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries, and Large Corpora, In Proceedings of ACL 2001 Workshop on Human Language Technology and Knowledge Management, Toulouse, France, July 6-7, 2001
[12] 고영만,「시소러스 기반 온톨로지에 관한 연구」, 성균관대학교 정보관리연구소, 정보관리 제 5집(2006), pp.5 ~ 22
[13] 황금하, 이신목, 남윤영, 신지애, 최기선. 시소러스를 이용한 온톨로지 구축에서의 Isa 관계 설정. 한국정보과학회 제 33회 정기 총회 및 추계학술대회 논문집, 서울, 2006.10