온톨로지의 구축과 학습: 상하위 관계

한국과학기술원 최기선 · 류법모

1. 서 론

"온톨로지"라는 용어에 대한 해석으로 "an ontology is an explicit formal specification of a shared conceptualization" 이라는 Gruber [1]의 정의를 가장 많이 인용하고 있다. 이 정의를 바탕으로 온톨로지의 세부적인 정의를 살펴보면, "shared"(공유)라 함은 개념이 해당 영역 구성원뿐만 아니라 컴퓨터 간에 합의된 지식에 바탕을 두고 있다는 것을 의미한다. "conceptualization"(개념화)라 함은 대상 세계에서 일어나는 현상에 연관된 개념들을 특정 목적을 위하여 표현하기 위한 추상적인 모델을 일컫는다. 또한 "formal"(형식적)이라는 것은 기계 가독형이어야 한다는 것을 의미하며, "explicit"(명시적)이라 함은 개념의 종류와 그들 간의 관계, 그리고 그 개념들의 사용에 있어서 주어지는 제약사항을 명백하게 정의한다는 것이다[2].

기존의 온톨로지들은 대부분 전문가의 수작업으로 구축되고 있지만[3], 시간 및 인적 제약 때문에 실용적인 온톨로지를 구축하기 어렵다. 앞으로 온톨로지에서 표현되는 여러 가지 관계 중에서 가장 핵심인 개념간 계층관계를 자동으로 추출하는 방법을 설명하고자 한다. 이 방법을 통하여 전문가의 수작업을 최소화할 수 있고, 여러 전문가들의 작업결과가 일관성을 가지게 된다. 따라서 기존의 온톨로지를 "구축"한다는 개념에서 온톨로지를 "학습"한다는 개념으로 전환하게 된다. 그림 1은 온톨로지 학습 단계를 케이크 모양으로 도식화한 것이다. 온톨로지 학습에서 가장 기본 단계인 "Terms" 단계에서는 온톨로지 구축을 위한 대상 용어를 추출하고 선정하며, "Synonyms" 단계에서는 선정한 용어들 사이의 동의어를 찾아서 그룹핑하고, "Concepts" 단계에서는 그룹핑된 용어들을 개념으로 표현하고, "Concept Hierarchies" 단계에서는 개념들 사이의 상하위 관계를 설정하고, "Relations" 단계에서는 상하위어 관계 이외의 다양한 관계를 표현하며, 마지막으로 "Rules" 단계에서는 개념 사이의 관계를 논리 형태로 표현한다. 전체 학습단계에서 "Concept Hierarchies"는 개념들을 조직화하는 가장 기본적이고 필수적인 단계이다[4]. 개념간 상하위 관계는 개념간 상속관계를 표현하기 때문에 지능형 시스템에서 상하위어 관계 탐색을 통한 추론기능을 제공한다.

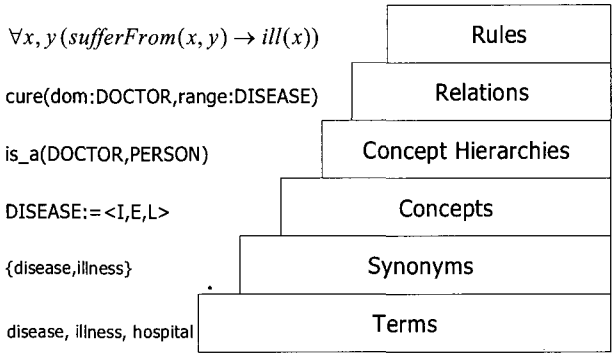

그림 1 온톨로지 학습 단계 케익

용어 계층 구조는 용어들 사이의 계층 관계를 설정하여 조직화시킨 것으로, 계층 구조에 포함된 모든 용어는 한 개 이상의 용어와 계층 관계를 가진다. 계층 관계는 *IS-A*, *PART-OF*, *INSTANCE-OF* 등의 관계를 포함한다. 여기에서는 온톨로지의 기본 프레임워크인 상하위어 관계(*IS-A*)를 포함한 여러 가지 용어간 계층관계를 자동으로 획득하기 위한 방법을 설명한다. 먼저 규칙 기반 학습 방법을 2절에서 설명하고, 통계 기반 방법을 3절에서 그리고 4절에서 용어의 전문성과 유사도를 이용한 방법을 설명한다.

2. 규칙 기반 학습 방법

2.1 어휘 구문 패턴 기반 학습 방법

용어 계층 구조를 구축하기 위한 규칙 기반 방법 중 가장 널리 알려진 방법은 어휘-구문 패턴을 이용하는 방법이다. 이 방법에서는 어휘 정보와 구문 정보가 정규 표현의 형태로 표현되고, 말뭉치 또는 웹에서 패턴에 일치하는 부분을 추출하여 상하위어 관계를 설정한

다. 일반적으로 영어를 모국어로 사용하는 사람들이 "*a L_0 is a (kind of) L_1*" 패턴을 만족하면 L_1은 L_0의 "상위어"라고 하고, L_0는 L_1의 "하위어"라고 한다. 상위어와 하위어 관계는 전이적인 성질을 가진다. Hearst [5]와 Caraballo [6]는 식 (1)과 같은 어휘-구문 패턴을 이용하여 단어의 상하위 관계를 추출하였다.

$$NP_1 \{, NP_i\}^* \{,\} \text{ and other } NP_0 \quad \text{for all } NP_i,\ 1 \le i \le n,\ IS\text{-}A(NP_i, NP_0) \tag{1}$$

여기에서 $IS\text{-}A(w_1, w_2)$ 는 w_1이 w_2의 하위어임을 나타낸다. 예를 들어 문장 "… temples, treasuries, and other important civil buildings"에서 *IS-A* (temple, civil building), *IS-A*(treasury, civil building) 관계를 추출할 수 있다.

Berland [7]는 Hearst가 *IS-A* 관계를 추출한 방법과 유사한 방법으로 "전체-부분" 관계를 추출하였다. 이 연구에서는 그림 2와 같은 패턴을 사용하여 전체-부분 관계를 NANC (North American News Corpus, 약 100,000,000 단어)에서 6개의 명사(book, building, car, hospital, plant, school)의 "부분"을 추출하였다. 추출한 결과에서 추상적인 의미를 나타내는 명사를 간단한 규칙을 이용하여 제거하였고, likelihood 함수를 사용하여 추출한 "부분"들을 정렬하였다. 정렬한 "부분" 중에서 상위 50개를 선택하여 사람이 검증하는 방법으로 평가하였다. 50개의 상위 순서 "부분" 중 약 55% 정도가 실제 "부분"을 나타내는 명사였고, 20개의 상위 순서 "부분" 중 약 70% 정도가 실제 "부분"을 나타내는 명사였다.

A. *whole*/NN[-PL]'s/POS *part*/NN[-PL]
... building's basement ...
B. *part*/NN[-PL] of/PREP {the|a}/DET *mods*/[JJ|NN]* *whole*/NN
... basement of a building ...
C. *part*/NN in/PREP {*the* | *a*}/DET *mods*/[JJ|NN]* *whole*/NN
... basement in a building ...

표현 형식:
type_of_ word/TAG type_of_word/TAG ...
TAG의 종류:
NN=Noun, NN-PL=Plural Noun, DET=Determiner,
PREP=Preposition, POS=Possessive, JJ=Adjective

그림 2 전체-부분 관계를 추출하기위한 어휘-구문 패턴

2.2 정의문 패턴 기반 학습 방법

정의문 패턴 기반 방법은 용어의 정의문에서 많이 나타나는 패턴을 이용하여 계층 관계를 추출한다. Hearst [5]가 제안한 어휘-구문 패턴을 이용하는 방법과 유사하지만 정의문 패턴을 중심적으로 이용하는 특징이 있다. 정의문 패턴을 이용하면 일반적인 어휘-구문 패턴을 이용하는 경우보다 정확하게 계층 관계를 추출할 수 있는 장점이 있다. 정의문은 용어의 의미를 명확하고, 정확하고, 완벽하게 표현하는 문장이며, 과학 기술 문헌에서 특정한 개념, 동작, 객체 등을 설명하기 위하여 자주 나타난다. ISO 704 규정에서는 상위개념과 그 개념을 다른 개념과 구분 짓는 의미 특징(characteristics)를 이용하여 식 (2)와 같이 용어의 정의문을 구성한다 [8].

$$X = Y + \text{차별적 의미특징} \tag{2}$$

여기에서 X는 정의될 용어를 말하며, Y는 X에 대한 상위개념이다. "차별적 의미특징"이란 동위어(cohyponyme) 들로부터 그 용어를 구별해 주는 특징적인 의미 속성을 말한다. 어휘의미론적으로 말하면, X의 내포적 의미자질 집합(set of intensional semantic features) 중에서 Y의 내포적 의미자질 집합을 제거하고 남은 의미자질을 말한다. 동치관계를 나타내는 "="는 연결동사(connective verb) 라고 불리는 동사들로 표현된다. 영어의 경우 be, mean, consist of 등의 동사가 이에 해당한다. 아래의 영어 단어 "knife"에 대한 정의문에서 밑줄로 표시한 부분이 상위 개념이고, 이탤릭체로 표시한 부분은 의미특징이다[9]. 즉 "knife"의 상위 개념은 "instrument"이고, "which" 이하 절이 의미특징을 표현하는 부분이다.

예) A **knife** is an instrument *which is used for cutting.*

정의문 패턴을 이용하여 상하위어 관계를 추출하기 위해서는 정확한 정의문을 확보하는 일이 중요하다. 많은 경우, 전문 용어 사전의 정의문을 이용할 수 있지만 신조어인 경우는 기존에 출판된 전문용어 사전에서 정의문을 찾을 수 없다. 따라서 웹 검색을 이용하여 정의문을 먼저 추출하고, 정의문 패턴을 적용하여 용어의 상위어를 찾을 수 있다. 예를 들어 기계 학습 분야에서 많이 사용하는 전문용어인 "지지 벡터 기계"의 정의문은 영어 대역어 "support vector machine"를 이용하여 웹에서 검색할 수 있다. 웹 검색 엔진에서 "a support vector machine is a(n)"라는 구 (phrase) 검색 기능을 이용하여 검색한 한 가지 결과는 다음 예와 같다.

예) A **support vector machine** is a supervised learning algorithm *developed over the past decade by Vapnik and others.*

검색된 정의문에 정의문 패턴을 적용하여 "support vector machine"의 상위개념 "supervised learning algorithm"를 추출하여 계층 관계 *IS-A*("support vector machine", "supervised learning algorithm")를 만들 수 있다.

2.3 수직 관계 기반 학습 방법

도메인의 개념이 용어로 표현되는 경우, 기존의 용어에 수식어를 붙여서 새로운 용어를 만드는 경우가 많다[8]. 따라서 용어 구성 단어 사이의 수직 관계를 이용하여 용어간 상하위어 관계를 추출하는 방법이 많이 사용되고 있다. Velardi [10]와 Cimiano [11]는 주어진 두 용어 t_1과 t_2에서 t_2가 t_1에 매칭되고, t_1이 추가적으로 다른 용어나 형용사에 의하여 수식되는 경우 *IS-A*(t_1, t_2) 관계가 성립하는 특성을 이용하였다. 예를 들어 t_1 = "read only memory"이고, t_2 = "memory"인 경우 *IS-A*("read only memory", "memory") 관계가 성립한다. Cimiano의 실험에서 이 방법은 정확률 50 %, 재현율 3.77 %를 보였다. 수직관계에 의한 방법이 올바른 *IS-A* 관계만을 생성하지는 않는다. 예를 들어 두 용어 "exclusive OR gate"와 "OR gate" 사이에는 수직관계 조건을 만족하지만 두 용어는 대등한 관계이기 때문에 *IS-A* 관계가 성립하지 않는다.

3. 통계 기반 학습 방법

통계 기반 방법은 분포 가정(distributional hypothesis)을 기반으로 용어간 계층관계를 설정한다. 분포 가정에서는 말뭉치에서 유사한 문맥을 공유하는 용어들은 유사한 의미를 가진다고 가정한다. Pereira [12]는 주어진 명사들을 그 명사들을 직접목적어로 가지는 동사들의 분포를 이용해서 군집화하였고, 한 개의 명사가 여러 개의 군집에 포함되는 것을 허용했다. 또한, 생성된 군집들을 결정 어닐링(deterministic annealing) 방법을 사용하여 하향식 방법으로 계층 구조를 만들었다. 어닐링 파라메터가 증가함에 따라서 기존의 군집들이 불안정한 상태가 되고, 불안정한 상태가 임계치를 넘어가는 군집들은 분할하였다. Caraballo [13]는 명사구들이 접속사로 결합된 패턴(예: "executive vice-president and treasurer")과 동격어 명사구 패턴 (예: "James H.Rosenfield, a former CBS Inc. executive")을 추출한 뒤, 해당 명사구에 포함된 명사들을 이용하여 각 명사들의 문맥을 벡터로 표현하였다. 각 명사들의 문맥 벡터 사이의 코사인 유사도를 이용하여 상향식 계층적 클러스터링을 수행하였다.

대부분의 통계 기반 학습 방법은 문맥 정보 사이의 유사도 계산을 전제하고 있다. 기존의 통계적인 자연언어처리 방법에서 많이 사용하고 있는 다양한 유사도 계산 방법이 계층관계 학습에도 널리 적용되고 있다[14]. 이 중에서 코사인 유사도 계산 방법과 상대 엔트로피 유사도 계산 방법이 대표적으로 사용된다. 코사인 유사도 계산 방법은 두 개의 벡터 사이의 상관계수를 정규화한 것이다. 두 개의 벡터가 각각 유사도를 비교하고자 하는 두 개의 용어의 문맥정보를 대표한다고 할 때, 이 계산 방법에서는 두 문맥정보의 관련성 척도를 나타낸다. 코사인 유사도 계산 방법은 식 (3)과 같이 표현된다.

$$Sim(t_1,t_2)=\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (3)$$

여기에서 (x_1, x_2, ⋯, x_n)와 (y_1, y_2, ⋯, y_n)은 각각 두 용어 t_1과 t_2의 자질에 대하여 가중치를 나타내는 벡터이다.

또 다른 유사도 계산 방법인 상대 엔트로피 계산 방법에서는 두 용어의 문맥정보를 확률질량함수으로 표현한 뒤, 두 개의 확률질량함수 사이의 상대 엔트로피를 계산하는 방법으로 두 용어 사이의 의미거리를 추정한다. 두 개의 확률질량함수 p(x), q(x)에 대하여 상대 엔트로피는 식 (4)와 같이 정의된다.

$$D(p\|q)=\sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \quad (4)$$

여기에서 $0\log(0/q)=0$, $p\log(p/0)=\infty$로 정의한다. 상대 엔트로피는 Kullback-Leibler divergence로 알려져 있으며, 두 개의 확률분포의 다른 정도를 측정한다. 이 성질은 항상 음이 아닌 실수를 가지며, $p=q$인 경우에 항상 $D(p||q)=0$ 이 된다 [15].

표 1 유사도 관계 계산을 위한 문맥 정보의 예

	호텔	숙소	주소	주말	테니스
호텔	-	14	7	4	6
숙소	14	-	11	2	5
주소	7	11	-	10	3
주말	4	2	10	-	5
테니스	6	5	3	5	-

표 1은 크기가 5인 작은 크기의 문맥 정보라고 가정하자. 표의 숫자는 두 단어가 같은 문맥에서 나타난 횟수를 의미한다. 이 예를 이용하여 유사도 계산 방법을 설명한다."호텔"과 "숙소"의 문맥정보 벡터는 각각 x=(0, 14, 7, 4, 6), y=(14, 0, 11, 2, 5)이고 코

사인 유사도 계산 방법에서 두 벡터 사이의 유사도는 다음과 같다.

$$\cos(x,y)=\frac{0\times14+14\times0+7\times11+4\times2+6\times5}{17.2\times18.6}=0.36$$

상대엔트로피를 구하기 위해서는 확률 밀도 함수를 먼저 계산하여야 한다. "호텔"에 대한 확률밀도함수 p는 (0.0, 0.45, 0.22, 0.13, 0.19)이고, "숙소"에 대한 확률밀도함수 q는 (0.44, 0.0, 0.34, 0.06, 0.16)이다. 이 값을 이용하여 "호텔"에 대한 "숙소"의 상대 엔트로피를 구하면 다음과 같다.

$$D(p\|q)=0.22\cdot\log\frac{0.22}{0.34}+0.13\cdot\log\frac{0.13}{0.06}+0.19\cdot\log\frac{0.19}{0.16}\approx0.016$$

이 외에도 두 벡터 X, Y 사이의 유사도 계산 방법은 coefficient, dice coefficient, Jaccard coefficient, overlap coefficient 등이 있다.

4. 용어의 전문성과 의미 유사도를 이용한 방법

이 절에서는 용어의 전문성과 용어간 의미 유사도를 이용하여 용어의 계층 구조를 구축하는 방법을 설명한다. 먼저 용어의 전문성을 이용하여 주어진 용어의 상위어 후보를 선택한 후, 용어간 의미 유사도를 이용하여 선택된 후보 중에서 최적의 상위어 후보를 결정한다.

4.1 용어의 전문성

용어의 전문성 (specificity)은 용어가 포함하는 전문적인 정보의 양을 정량적으로 표현한 것이다 [16]. 어떤 용어가 도메인 전문적인 정보를 많이 포함하고 있을 때 전문성이 높고, 반대로 일상적인 용어일수록 전문성이 낮다고 가정한다. 이 방법에서는 용어의 구성정보와 문맥정보를 이용하여 주어진 도메인 D 에서 사용되는 용어 t 의 전문성을 식 (5)와 같이 실수(R)로 표현한다.

$$Spec(t|D)\in R^{+} \quad (5)$$

전문분야 개념은 자신을 다른 개념들과 구분시킬 수 있는 고유한 특징 집합을 가진다. 비슷한 특징 집합을 가지는 개념들은 유사한 의미를 표현한다. 어떤 개념을 표현하는 특징 집합에 새로운 특징을 추가하여 더 전문적인 개념을 만들 수 있다. 일반적으로 기존의 개념 X 와 X 에 새로운 특징을 추가하여 생긴 개념 Y 사이에는 상하위 관계가 성립된다. 즉 X 는 Y 의 상위 개념이고, X 의 특징 집합은 Y 의 특징 집합의 부분집합이다 [8]. 전문분야 개념이 전문용어로 표현될 때 다음과 같은 두 가지 특징을 관찰할 수 있다. 첫째, 기존의 전문용어에 새로운 특징을 추가하는 수식어를 부가하여 더 전문적인 개념을 표현하는 용어가 만들어진다. 예를 들어 표 2에서 "insulin-dependent diabetes mellitus"는 "diabetes mellitus" 에 "insulin-dependent" 라는 수식어가 부가되어 만들어진 더 전문적인 용어이다. 이 방법으로 생성된 전문용어는 추가된 수식어의 전문성만큼 전체 용어의 전문성이 증가한다. 이 경우에는 용어의 구성단어들이 용어의 특징을 표현하는 정보로 사용된다. 둘째, 기존 전문용어의 구성단어와 전혀 다른 단어를 이용하여 더 전문적인 개념을 표현하는 경우가 있다. 예를 들어 표 2에서 "Wolfram syndrome"은 상위어 "insulin-dependent diabetes mellitus"의 구성단어와 전혀 다른 단어들로 구성되어 있다. 이 경우에는 용어의 문맥정보가 용어의 특징을 표현하는 정보로 사용된다.

표 2 MeSH[1] 트리의 일부분. 노드 번호는 용어 사이의 계층구조를 나타낸다.

노드 번호	용어
C18.452.297	diabetes mellitus (당뇨병)
C18.452.297.267	insulin-dependent diabetes mellitus (인슐린 의존형 당뇨병)
C18.452.297.267.960	Wolfram syndrome (볼프람 증후군)

정보이론에서는 정보량을 "불확실성" 또는 "놀라움"의 개념으로 설명한다. 출현 확률이 낮은 메시지가 채널의 출력에서 나타나기 전에는 "불확실성"이 높다고 이야기한다. "불확실성"이 높은 메시지가 실제로 나타난 경우 "놀라움"의 정도는 커지고, 그 메시지를 표현하기 위한 비트수는 다른 출력에 비해 길어진다. 따라서 그 메시지의 정보량은 높아진다 [17]. 도메인 D 와 관련된 말뭉치에서 나타나는 용어들이 어떤 채널의 출력에서 관찰되는 일련의 메시지라고 가정하면, 용어 t 가 관찰되는 사건 x 의 정보량 $I(x)$ 를 말뭉치의 각종 통계정보를 이용하여 계산할 수 있다. 그리고 $I(x)$ 를 식 (6)과 같이 용어 t 의 전문성 $Spec(t|D)$ 으로 사용한다.

$$Spec(t|D)\approx I(x) \quad (6)$$

이 경우, 정보량 $I(x)$ 는 식 (7), (8), (9)와 같은 성질을 가진다.

$$I(x)=0,\ \ p(x)=1 \text{ 일 때} \quad (7)$$

1) 미국 의학도서관(NLM, National Library of Medicine)에서 관리하는 의학용어 리스트이다(http://www.nlm.nih.gov/mesh/).

말뭉치에서 나타날 확률이 1인 용어 t 가 실제 말뭉치에서 출현할 경우 얻을 수 있는 정보량은 없다.

$$I(x) \geq 0, \quad 0 \leq p(x) \leq 1 \text{ 일 때} \tag{8}$$

용어 t 가 말뭉치에서 나타날 경우, 정보의 손실을 초래하는 경우는 없다. 즉 말뭉치에서 나타나는 모든 용어는 정보량을 계산할 수 있으며, 0 이상의 값을 가진다.

$$I(x_i) > I(x_j), \quad p(x_i) \leq p(x_j) \text{ 일 때} \tag{9}$$

용어 t_i 가 t_j 보다 말뭉치에서 나타날 확률이 낮을 때, 실제 말뭉치에서 t_i 가 나타날 경우, 얻을 수 있는 정보량이 t_j 가 나타날 경우 얻을 수 있는 정보량보다 많다. 즉 말뭉치에서 출현 확률이 낮은 용어일수록 정보량이 많아지고 전문성이 높아진다.

4.2 용어간 의미 유사도

특정 분야의 개념은 지식 전달 방식에 따라서 서로 다른 형태로 표현된다. 자연 언어를 이용하여 지식을 전달하는 경우, 개념은 해당 분야의 전문용어로 표현될 수 있다. 개념은 그 개념을 설명하는 특징들의 집합으로 표현되고, 그 특징들은 다른 특징들과 결합하면서 새로운 개념을 생성한다. 용어 관리의 중요한 부분 중의 하나는 표층에서 나타나는 용어의 언어 현상을 분석하여 대응하는 개념의 특징을 파악하는 것이다. 특징 집합에 새로운 특징이 추가될수록 더 전문적인 개념을 나타내고, 그 반대의 경우는 광범위한 개념을 나타낸다. 용어간 의미유사도는 용어의 특징 집합 사이의 포함 관계의 정도를 정량적으로 표현한 것이다. 두 특징 집합이 완전히 일치하거나, 포함 관계에 있거나, 부분적으로 겹치는 관계에 있거나, 또는 전혀 겹치지 않는 경우를 표현한다. 동일한 용어도 사용되는 분야에 따라서 서로 다른 특징 집합을 가진다. 따라서 용어간 유사도도 분야 의존적인 성질을 가진다. 용어 간 의미 유사도를 표현하는 대표적인 표층 언어 현상은 용어의 구성 단어 특징과 용어의 문맥 정보 특징이 있다. 용어의 구성단어 특징은 언어의 조합성 (compositionality)을 이용하여 설명할 수 있다. 조합성은 복잡한 표현의 의미는 내부 구조와 구성 성분의 의미에 결정된다는 이론이다. 즉 구성 성분의 의미를 알고 있고, 구성 성분들이 결합되는 방법을 알고 있으면 전체 표현의 의미를 알 수 있다. 따라서 구성 단어들의 특징을 조합하여 용어의 의미를 파악할 수 있다. 이와는 반대로 말뭉치에서 공기하는 단어들의 유사도를 이용하여 용어간 유사도를 파악할 수 있다. 두 용어가 비슷한 문맥에서 사용되는 경우 의미적으로 유사하다. 이 이외에도 사전적 정의문이 유사한 경우 두 용어가 유사하다고 판단할 수 있다. 의미적으로 유사한 용어를 정의할 때, 비슷한 단어를 이용하여 정의하기 때문에 정의문에 나타나는 단어들을 비교하여 유사한 정도를 판단할 수 있다.

4.3 용어의 전문성과 유사도 기반 방법

전문적인 용어일수록 용어 분류체계에서 하위 계층에 위치하는 경우가 많기 때문에 용어의 전문성은 주어진 도메인 D의 전문용어 사이에 계층관계를 표현하는 필요조건으로 사용할 수 있다. 그림 3의 도메인 용어 계층구조 T_D 에서 용어 t_1 이 다른 용어 t_2 의 상위어인 경우 t_1 의 전문성은 t_2 의 전문성보다 작다. 이 조건을 이용하면 용어 t_1 과 t_2 가 의미적으로 충분히 유사하고, t_1 의 전문성이 t_2 의 전문성보다 작은 경우 t_1 이 t_2 의 상위어가 될 가능성이 매우 높다. 또 다른 예를 보면 t_1 의 전문성이 t_3 의 전문성보다 낮지만 두 용어가 의미적으로 유사하지 않기 때문에 두 용어 사이에 상하위관계가 성립할 가능성이 낮게 된다.

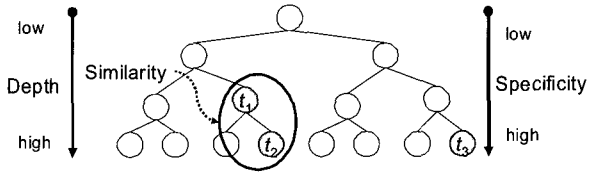

그림 3 도메인 용어 계층구조 T_D 에서 용어간 상하위어 관계와 용어의 전문성 조건, 용어간 유사도조건의 관계. 이 그림에서 두 용어 사이의 거리가 가까울수록 두 용어가 의미적으로 유사하다고 판단한다.

계층 구조 구축 과정은 현재의 계층 구조에 연속적으로 새로운 용어를 추가하는 과정을 반복한다. 계층 구조는 그림 4와 같이 초기에 비어 있는 상태에서 시작하여 반복적으로 새로운 용어를 추가하여 풍부한 구조를 가진다. 추가되는 용어는 용어의 전문성 값을 이용하여 정렬한다. 전문성이 높은 용어는 계층구조에서 하위 레벨에 위치하는 경향이 있고, 전문성이 낮은 용어는 상위 레벨에 위치하는 경향이 있다. 따라서 일반적인 용어부터 차례로 계층구조에 추가하면 계층구조는 상위 레벨부터 차례로 하위 레벨 방향으로 성장한다.

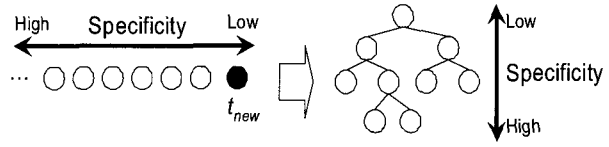

그림 4 클래스에 포함된 용어가 그 클래스의 계층 구조에 전문성이 낮은 용어부터 차례로 순차적으로 등록된다.

5. 요 약

온톨로지의 기본 개념, 응용 분야 및 학습 단계에 대하여 간단하게 설명하였고, 온톨로지 학습단계에서 전문 분야의 개념간 계층 관계 학습 방법에 대하여 자세하게 알아보았다. 전문분야 개념을 표현하는 전문 용어 사이의 계층 관계를 학습하는 방법은 크게 규칙 기반 방법, 통계 기반 방법 그리고 용어의 전문성과 유사도를 이용하는 방법으로 나눌 수 있다. 규칙 기반 방법은 비교적 정확한 결과를 얻을 수 있는 장점이 있지만 재현율이 낮은 단점이 있다. 기존의 통계 기반 방법에서는 재현율이 높은 장점이 있지만 정확률이 낮은 단점이 있다. 또한 이 방법에서는 순수하게 통계 정보만 이용하기 때문에 오류에 대한 분석이 어려운 단점이 있다. 용어의 전문성과 용어간 유사도를 이용한 방법에서는 용어의 전문성을 이용하여 기존의 계층 구조에서 상위어 후보를 선택하고, 용어간 유사도를 이용하여 선택한 후보를 정렬하여 최적의 후보를 찾는다. 이 방법은 상위어 선정 과정을 두 단계로 분리하여 수행하기 때문에 오류 분석이 용이한 장점이 있다. 향후 온톨로지 학습 과정에서 계층 관계뿐 아니라 인과 관계 및 다양한 관계의 학습과 관련된 연구가 진행되어야 한다.

참고문헌

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

[2] 이재호, "시맨틱 웹의 온톨로지 언어", 정보과학회지, 제21권, 제3호, pp. 18-27, 2003.

[4] Lassila, O., McGuinness, D., "The Role of Frame-Based Representation on the Semantic Web," Technical Report KSL-01-02, Knowledge Systems Laboratory, Stanford University, 2001.

[5] Hearst, M. A., "Automatic Acquisition of Hyponyms from Large Text Corpora," Proceedings of the Fourteenth International Conference on Computational Linguistics, 1992.

[6] Caraballo, S. A, "Automatic construction of a hypernym-labeled noun hierarchy from text," Proceedings of ACL, 1999.

[7] Berland, M., Charniak, E., "Finding Parts in Very Large Corpora," Proceedings of ACL, 1999.

[8] ISO, "Terminology work-Principle and methods," ISO 704 Second Edition, 2000.

[9] Pearson, J. "Terms in Context," Series of Studies in Corpus Linguistics Vol. 1, John Benjamins Publishing Company, 1998.

[10] Velardi, P., Fabriani, P., and Missikoff, M., "Using Text Processing Techniques to Automatically enrich a Domain Ontology," Proceedings of the ACM International Conference on Formal Ontology in Information Systems, 2001.

[11] Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S., "Learning Taxonomic Relations from Heterogeneous Evidence," Proceedings of ECAI2004 Workshop on Ontology Learning and Population, 2004.

[12] Pereira, F., Tishby, N., and Lee, L., "Distributational clustering of English words," Proceedings of ACL, pp. 183-190, 1993.

[13] Caraballo, S. A. and Charniak, E. "Determining the Specificity of Nouns from Text," Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 63-70, 1999.

[14] Lee, L., "Measures of Distributional Similarity," Proceedings of ACL, pp. 25-32, 1999.

[15] Manning, C. D., Schutze, H., "Foundations of Statistical Natural Language Processing," The MIT Press, 1999.

[16] Lenat, D.B. et. al., "Cyc: toward programs with common sense," Communications of the ACM, 33(8), pp.30-49, 1990.

[16] Ryu, P., Choi, K., "Measuring the Specificity of Terms for Automatic Hierarchy Construction," Proceedings of ECAI2004 Workshop on Ontology Learning and Population, 2004.

[17] Cover, T.M. & Tomas, J.A., "Elements of Information Theory," New York: John Wiley and Sons Inc., 1991.

최 기 선

1978 서울대학교 수학과(학사)
1980 한국과학기술원 전산학과(석사)
1986 한국과학기술원 전산학과(박사)
1987~1988 일본 NEC C&C 정보연구소 연구원
1988~현재 한국과학기술원 전산학과 교수
1997~1998 미국 스탠포드대학 CSLI 객원교수
2002~2003 일본 NHK 방송기술연구소 초빙연구원
2006~현재 한국인지과학회 회장
2003~현재 국가지정 언어자원특수소재은행장 http://bola.kaist.ac.kr
2002~현재 ISO/TC37/SC4 언어자원관리표준 Secretary
2002~현재 TermNet 회장
2000~현재 ACM TALIP, IJCPOL 편집위원, IAMT council member
1998~현재 전문용어언어공학연구센터 http://korterm.or.kr
관심분야 : 온톨로지, 텍스트마이닝, 인공두뇌, 지식획득, 창의계산론, 언어공학, 시맨틱웹
E-mail : kschoi@cs.kaist.ac.kr http://ci.kaist.ac.kr/

류 법 모

1995 경북대학교 컴퓨터공학과(학사)
1997 포항공과대학교 컴퓨터공학과(석사)
2000~현재 한국과학기술원 전산학과 박사과정
1997~1999 한국전자통신연구원(ETRI) 자연어처리연구실 연구원
1999~2004 (주)케이포엠 기술연구소 연구원
관심분야 : 자연언어처리, 온톨로지학습
E-mail : pmryu@world.kaist.ac.kr