한국지능시스템학회 논문지 2009, Vol. 19, No. 3, pp. 363-370

워드넷과 구글에 기반한 온톨로지 개체의 일반화

Generalization of Ontology Instances Based on WordNet and Google

강신재* · 강인수**
Sin-Jae Kang* and In-Su Kang**

* 대구대학교 정보통신대학 컴퓨터 · IT공학부
** 경성대학교 멀티미디어대학 컴퓨터정보학부

요 약

본 논문은 온톨로지의 지식을 확장하기 위하여 웹 페이지 등 텍스트에서 추출된 온톨로지 개체(ontology instances)를 일반화하는 방법을 제시한다. 이를 위해서는 단어 의미 중의성 해소 과정이 필수적인데, 구글, 워드넷과 같은 오픈 API와 어휘 리소스를 이용하여 비교사학습 방법으로 해결하는 방법을 제안한다. 실험 결과 기존 연구에 비해 15.8%의 성능 향상을 얻을 수 있었다.

키워드 : 단어 의미 중의성 해소, 워드넷, 구글, 온톨로지 확장

Abstract

In order to populate ontology, this paper presents a generalization method of ontology instances, extracted from texts and web pages, by using unsupervised learning techniques for word sense disambiguation, which uses open APIs and lexical resources such as Google and WordNet. According to the experimental results, our method achieved a 15.8% improvement over the previous research.

Key Words : Word Sense Disambiguation, WordNet, Google, Ontology Population

1. 서 론

본 논문은 웹 페이지에서 자연어처리를 통하여 자동 추출된 온톨로지 개체(ontology instances)를 일반화하여 온톨로지의 지식을 확장하는 기법을 다룬다. 하나의 온톨로지 개체를 그것의 상위 개념으로 일반화시키는 일은, 자연어 문장에 출현한 단어(word)를 그것의 의미(sense label)로 태깅(tagging)하는 단어 의미 중의성 해소(word sense disambiguation: WSD) 절차에 대응시킬 수 있다. 이 논문은 온톨로지 개체의 일반화를 위해, 오픈 API(application program interface)와 리소스를 활용하는 방법을 제시한다.

온톨로지 개체란 그림 1과 같이 두 인수(argument)와 그들의 관계(relation)를 하나의 지식 단위로 표현하는 온톨로지에서 관계에 연결된 각 인수를 의미한다. 그림 1의 지식 단위는 W3C(world wide web consortium)의 온톨로지 정의에서는 RDF(resource description framework)에 대응된다. 예를 들어, “서울대학교는 한국에 위치하고 있다”라는 하나의 단위 지식은 그림 1과 같은 RDF로 표현될 경우 “[서울대학교] - [~에 위치한다] - [한국]”이 될 것이다. 온톨로지 개체의 일반화란, 개체에 대응하는 용어(term)를 보다 일반화된 개념으로 매핑(mapping)하는 것을 의미한다.

접수일자 : 2008년 11월 14일
완료일자 : 2009년 5월 28일
본 논문은 지식경제부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

예를 들면, 전술한 [서울대학교]라는 개체는 [대학교] 혹은 [교육기관]이라는 개념으로 일반화될 수 있을 것이다. 이러한 일반화 관계 설정을 통해 상위 개체인 [교육기관]이 갖는 속성과 제약들을 그 하위 개체인 [서울대학교], [부산대학교] 등이 공유하도록 할 수 있는 것이다.

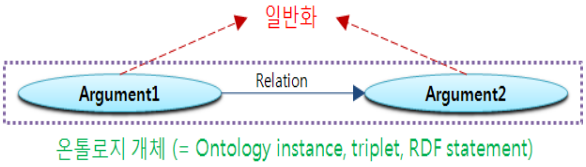

그림 1. 온톨로지 개체의 일반화
Fig 1. Generalization of ontology instances

온톨로지 개체 일반화는 보다 큰 문제인 온톨로지 학습(ontology learning)의 부분 문제에 속한다고 볼 수 있는데, 온톨로지 학습은 텍스트로부터 용어(term) / 동의어(synonym) / 개념(concept) / 개념체계(concept hierarchy) / 의미관계(relation) / 의미관계체계(relation hierarchy) / 공리(axiom) 등을 학습하는 것을 의미하며, 추가적으로 온톨로지 개념 노드에 개체(instance)들을 삽입하는 온톨로지 확장/증식(population) 단계가 포함되기도 한다. 본 연구에서 다루는 내용은 개념체계의 학습 및 온톨로지 확장에 해당된다.

용어를 일반화하기 위해서 접근할 수 있는 방법은 크게

세 가지 정도로 분류해 볼 수 있는데, 첫 번째 방법은 기구축된 온톨로지 개념으로 매핑하는 방법이고, 두 번째 방법은 의미자질(semantic feature) 또는 의미소(semantic primitive)를 이용하여 일반화하는 것이다. 세 번째 방법은 온톨로지 개체를 자연어처리 분야에서 광범위하게 사용되고 있는 워드넷(WordNet)의 synset으로 매핑하여 일반화하는 것이며, 이 방법에서는 단어 의미 중의성 해소 과정이 필요하게 된다. 본 연구에서는 단어 의미 중의성 해소 처리 시 학습데이터의 부족으로 인한 문제를 해결하기 위하여 웹 자원과 어휘 데이터베이스(WordNet)를 활용하여 자동으로 단어 의미의 중의성을 해소하고 그 결과를 워드넷의 synset으로 표현하고자 한다. 단어 의미별 구분 정보를 구축하기 위해 워드넷으로부터 synset 목록을 구축하게 되고, Google의 검색결과로부터는 대량의 문맥정보를 추출하여 이들 간 유사도 계산을 함으로써 단어 의미 중의성 해소 과정을 자동화하는 것이다.

2장에서는 관련된 기존 연구를, 3장에서는 온톨로지 개체를 일반화하기 위한 접근법을 정리하고, 4장에서는 워드넷과 구글을 이용하여 온톨로지 개체를 일반화하는 방법에 대해 설명한다. 5장에서는 실험 결과를 분석하고, 6장에서는 결론과 향후 연구계획을 제시한다.

2. 관련 연구

개념체계학습의 많은 연구들은 정규식 형태로 표현된 어휘-구문 패턴을 텍스트에서 탐색함으로써 상하위 개념용어들을 획득하는 방법[1]에서 출발하였으나, 이러한 패턴기반의 방법은 높은 정확률을 보이는 반면 낮은 재현율을 갖는 단점이 있어 전문가의 지속적인 패턴 추가 및 튜닝이 요구되는 어려움이 있다. 이 문제의 해결을 위해 기계학습을 통해 패턴을 자동학습하려는 시도를 한 연구[2]가 있으며 성능은 높지 않았으나 패턴기반방법의 문제인 재현율을 높이는 시도로 평가된다.

개념체계학습을 위해 주로 사용되어 온 비교사 군집 기법(unsupervised clustering)은 각 용어를 하나의 군집으로 고려하는 초기 비계층 용어 집합에서 출발하여, 유사 군집들을 더 큰 군집으로 그룹화하는 작업을 반복함으로써 개념학습과 개념체계 학습을 동시에 수행하는 방법이다. 그러나 군집기법 역시 텍스트로부터의 수집되는 군집간 유사도 계산을 위한 자질의 빈약함으로 인해 군집 결과가 만족스럽지 못한 단점이 있다. 이의 해결을 위해 군집화 과정에 사람을 개입시키거나[3], 외부 자원을 활용[4]하는 시도가 있었다. 그러나 무엇보다 계층적 군집화의 난제는 군집 과정에 얻어지는 상위 군집에 대응되는 개념 명칭의 부재에 있다. 이처럼 텍스트로부터의 개념체계학습은 자질부족문제와 상위개념 명명의 문제를 안고 있으며 현재까지의 기술은 획기적 돌파구를 찾지 못한 수준에 머무르고 있다.

온톨로지 학습 연구를 수행하고 있는 주요 기관 및 시스템은 Karlsruhe 대학의 Text2Onto (독일), Amir Kabir 대학의 HASTI (이란), Brussel 대학의 OntoBasis (벨기에), DFKI의 OntoLT (독일 인공지능연구센터), Economic 대학의 TextToOnto++ (체코), USC의 CBC (미국), Keio 대학의 DODDLE (일본), Roma 대학의 OntoLearn (이탈리아) 등이 있으며[5], 이 분야와 관련해서는 유럽대륙이 선도적 연구를 진행하고 있다. 현재 대부분의 기관들이 용어로부터 시작하여 의미관계추출단계까지의 온톨로지 학습을 수행하는 온톨로지 학습도구들을 개발하고 있다.

개념체계 학습의 최신 기술들로 York 대학(영국)에서는 구글 검색엔진을 이용하여 개념의 문맥정보를 수집하고 이를 바탕으로 빈도수 기반의 방법을 개발하였으며 60%대의 정확률을 보고하였다[6]. 이 기술은 개념체계 학습의 자질부족문제를 웹이라는 자원을 통해 부분적으로 해소하고 있으며 적용이 용이한 빈도수 기반의 간단한 방법을 사용하고 있다.

국내의 경우 온톨로지 학습과 관련하여 KAIST에서 수행한 개념체계 학습 연구[7]를 제외하고는 관련 연구를 찾기 힘들다.

3. 개체 일반화를 위한 접근법

온톨로지 개체를 일반화하기 위해서 접근할 수 있는 방법은 크게 3가지 정도로 분류해 볼 수 있는데, 입력문장이 “Equipment such as handsets, modems and extension bells sold in or for use in Australia often arrives packaged with one of these adaptors.”이고, 이 문장에서 추출된 하나의 온톨로지 개체가 “Argument 1: handsets, Relation: isa, Argument 2: equipment”와 같다고 가정할 때, 다음 각각의 접근방법에 따른 결과를 예시한다.

3.1 기구축 온톨로지

기구축 온톨로지를 이용하는 개체일반화는, 일반화를 위한 온톨로지를 새롭게 구축하는 것이 아니라 기존에 널리 사용되고 있는 온톨로지를 그대로 활용하고자 할 때 사용할 수 있는 접근법이다. 단, 매핑의 대상이 되는 온톨로지가 검증되고 널리 인정받는 것이어야 한다는 조건이 필요하다. 또한 온톨로지 구조상에 개념(concept)과 개체(instance)의 명확한 구분이 있어야 개체를 개념에 매핑할 수 있다. 이 방법은 대부분 수작업으로 이루어질 수밖에 없는데 사람이 하기 때문에 매핑 결과의 일관성이 보장되지 않는 문제가 있다. 제시된 예제에 대해 일반화된 결과는 다음 그림과 같은 형태가 된다.

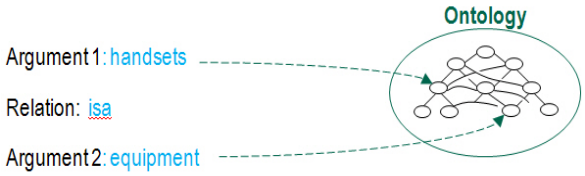

그림 2. 기존 온톨로지를 이용하여 일반화된 결과 예시
Fig 2. Example of generalized results by using existing ontology

3.2 의미자질(semantic feature)

두 번째 방법은 의미자질(semantic feature) 또는 의미원소(semantic primitive)를 이용하여 개체를 그것의 상위 개념으로 일반화하는 것이다. 의미자질이란 플러스나 마이너스 기호를 통해 해당 의미속성(human, male, adult 등)의 유무를 표현하는 표기법으로, 예를 들어 사람(man), 여성(woman), 소년(boy), 소녀(girl)를 의미자질로 표현하면 다음과 같다.

예) Man : [+HUMAN], [+MALE], [+ADULT]
Woman : [+HUMAN], [-MALE], [+ADULT]
Boy : [+HUMAN], [+MALE], [-ADULT]
Girl : [+HUMAN], [-MALE], [-ADULT]

온톨로지 개체 일반화에 있어 의미원소를 고려하는 기본 가정은 의미원소 집합으로부터 닫혀 있는 상위 레벨 온톨로지를 정의할 수 있다는 것이다. 예를 들어 10개의 의미원소가 있고 각 의미원소가 이진값을 가질 수 있다고 할 때 이 10개 의미원소들의 집합으로부터 정의되는 상위 레벨 온톨로지의 개념 노드의 수는 이론적으로 1024개로 고정된다. 의미원소를 이용한 상위 레벨 온톨로지 개념의 표현은 온톨로지 개체 일반화 작업을 온톨로지 개체에 각 의미원소의 값을 부여하는 작업으로 대체할 수 있게 하는 장점을 갖는다. 즉 '컴퓨터'라는 개체에 {artifact=TRUE, potential agent=TRUE}라는 의미원소들을 부여함으로써 '컴퓨터' 개체는 'artifact=TRUE'이면서 'potential agent=TRUE'인 개념에 자동으로 소속되게 된다. 이러한 의미원소 기반의 온톨로지 표현은 온톨로지 개체 일반화의 두 가지 큰 문제인 자질부족과 개념 명명의 문제를 동시에 해결하는 방안이 될 수 있다.

의미원소는 Wierzbicka[8]에 의해 제안된 이후 많은 비판 속에서도 현대 언어학에서 descriptive semantics를 위한 주요 방법 중 하나로 사용되고 있으며, 현재는 NSM(natural semantic meta-language)[1]이라는 이름으로 불린다. 초기 의미원소의 개수는 14개였으나 현재는 표 1에 보인 바와 같이 60여 개로 증가하였다.

표 1. NSM 의미원소
Table 1. NSM semantic primitive

Substantives:	I, YOU, SOMEONE, PEOPLE, SOMETHING/THING, BODY
Relational substantives:	KIND, PART
Determiners:	THIS, THE SAME, OTHER/ELSE
Quantifiers:	ONE, TWO, SOME, ALL, MUCH/MANY
Evaluators:	GOOD, BAD
Descriptors:	BIG, SMALL
Mental predicates:	THINK, KNOW, WANT, FEEL, SEE, HEAR
Speech:	SAY, WORDS, TRUE
Actions, events, movement, contact:	DO, HAPPEN, MOVE, TOUCH
Location, existence, possession, specification:	BE (SOMEWHERE),THERE IS, HAVE, BE (SOMEONE/SOMETHING)
Life and death:	LIVE, DIE
Time:	WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT
Space:	WHERE/PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE
"Logical" concepts:	NOT, MAYBE, CAN, BECAUSE, IF
Intensifier, augmentor:	VERY, MORE
Similarity:	LIKE

다른 사례인 MultiNet[2]에서 사용하는 이진 의미자질을 사용하여 제시된 예제를 일반화하는 경우 결과는 다음과 같다.

표 2. MultiNet에서 사용되는 의미자질
Table 2. Sematic features used in MultiNet

Name	Meaning	Examples	
		+	−
ANIMAL	animal	*fox*	*person*
ANIMATE	living being	*tree*	*stone*
ARTIF	artifact	*house*	*tree*
AXIAL	object having a distinguished axis	*pencil*	*sphere*
GEOGR	geographical object	*the Alps*	*table*
HUMAN	human being	*woman*	*ape*
INFO	(carrier of) information	*book*	*grass*
INSTIT	institution	*UNO*	*apple*
INSTRU	instrument	*hammer*	*mountain*
LEGPER	juridical or natural person	*firm*	*animal*
MENTAL	mental object or situation	*pleasure*	*length*
METHOD	method	*procedure*	*book*
MOVABLE	object being movable	*car*	*forest*
POTAG	potential agent	*motor*	*poster*
SPATIAL	object having spatial extension	*table*	*idea*
THCONC	theoretical concept	*mathematics*	*pleasure*

Argument 1: handsets
{ARTIF+, INSTRU+, MOVABLE+, SPATIAL+}
Relation: isa
Argument 2: equipment
{ARTIF+, INSTRU+, MOVABLE+, SPATIAL+}

이 방법은 모든 사람이 공감할 수 있는 의미자질 또는 의미소를 정의해야 한다는 점과 의미자질 부여가 자동으로 이루어지기 어렵기 때문에 전적으로 수작업으로 이루어져야 하며, 작업자에 따라 다른 결과가 나올 가능성이 높다는 문제가 있다. 또한 동일 작업자인 경우에도 온톨로지 개체에 의미자질을 할당할 때 일관성을 유지하기가 어렵다.

3.3 워드넷(WordNet)

워드넷[9]은 널리 알려진 대규모의 영어 어휘 의미 목록 데이터베이스이다. 워드넷은 영어 단어를 'synset'이라는 유의어 집단으로 분류하여 간략하고 일반적인 정의를 제공하고, 이러한 어휘목록 사이의 다양한 의미 관계를 기록한다. 현재 영어를 대상으로 구현된 대부분의 자연어처리 응용 시스템에서는 단어 의미 구분 등의 처리를 하기 위해 워드넷을 활용하고 있으며, 이를 편리하게 하기 위해 워드넷 검색, 유사도 계산 API 등 많은 라이브러리와 소프트웨어 도구들이 개발되어 제공되고 있다. 따라서 온톨로지 개체를 워드넷의 synset으로 매핑하여 일반화를 한다면 일반화된 온톨로지 개체의 활용가능성이 높아질 뿐만 아니라, 사람의 수작업이나 학습데이터 등이 필요 없이 일반화 과정을 자동화할 수 있게 된다. 다음 그림은 온톨로지 개체 일반화를 워드넷의 synset으로 매핑하여 처리하는 과정을 보여주는데, 단어 의미의 중의성을 해소(WSD, Word Sense Disambiguation)하는 과정이 필요하게 된다. 따라서 온톨로지 개체 일반화를 WSD의 특수한 문제로 볼 수 있다.

1) http://www.une.edu.au/bcss/linguistics/nsm/semantics-in-brief.php
2) http://pi7.fernuni-hagen.de/homepage

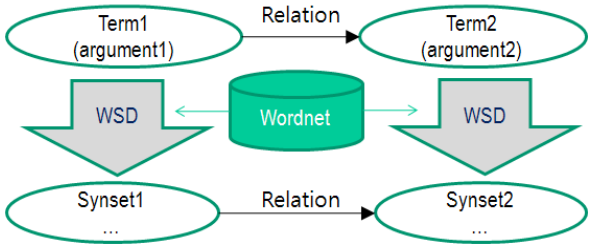

그림 3. 워드넷을 이용한 일반화 과정
Fig 3. Generalizing process using WordNet

동일 예제에 대한 일반화 결과는 다음과 같다.

Argument 1: handsets {SID-03488438-N}[3]
Relation: isa
Argument 2: equipment {SID-03294048-N}[4]

단어 의미 중의성 해소는 자연어처리 분야에서 오랫동안 연구해 온 주제이면서도 그 성능의 향상이 두드러지지 않는 연구 분야이다. 소규모로 정해진 단어를 대상으로 중의성을 해소할 때에는 의미 태깅된 말뭉치 등 학습 가능한 리소스가 어느 정도 확보되어 있으나, 모든 단어를 대상으로 WSD를 수행할 때에는 자료 부족 현상이 발생하여 성능의 향상을 꾀하기 어려운 실정이다. 본 연구에서는 자료 부족 현상을 해결하기 위하여 워드넷과 구글을 이용하여 자동으로 단어 의미의 중의성을 해소하고 그 결과를 워드넷의 synset으로 표현하고자 한다.

4. 연구내용

워드넷은 이를 활용하기 위한 다양한 API(검색, 유사도 계산 등)가 공개되어 있고, 구글은 PageRank 알고리즘을 개발/구현하여 현존 검색엔진 가운데 최상의 검색결과를 제공해주며, 사용자 프로그램에서 접근하여 결과를 사용할 수 있도록 API를 제공한다. 이 연구에서는 온톨로지 개체 일반화 과정을 자동화하기 위해 워드넷, 구글과 같은 공개 리소스/API를 적극 활용하여 단어의미 중의성을 해소하는 방법을 고안하였다.

단어 의미 중의성 해소(WSD)는 단어가 가지고 있는 여러 의미 가운데 그 단어가 사용된 문장에서의 의미를 결정하는 문제이다. 거의 대부분의 경우 단어의 의미는 주어진 문맥에서 정확히 결정될 수 있다. 모든 WSD 시스템은 접근 방법에 상관없이 WSD될 대상 단어(target word)의 문맥정보(contextual features)가 있어야 하고, 이 정보와 비교할 대상 단어의 의미 구분 정보(sense differentiation information)가 있어야 한다[10]. 이러한 정보를 자동으로 추출하기 위해 본 연구에서는 구글과 워드넷을 이용하여 그림 4와 같은 과정을 거친다.

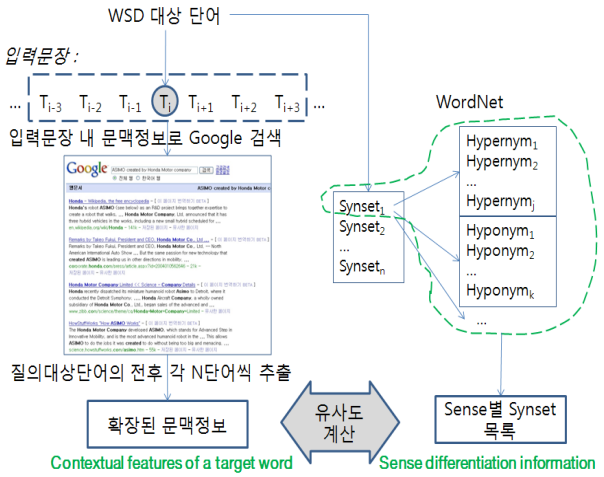

그림 4. WSD를 위한 정보의 추출
Fig 4. Information extraction for WSD

4.1 대상 단어의 문맥 정보(contextual features)

WSD 대상 단어 w의 직접 문맥정보로 w가 출현한 입력문장에서 w의 전후 N[5]개의 단어(CW, context window)를 추출하였다. 그러나 이 직접 문맥정보(direct contextual information)는 양이 적어 추가의 문맥정보를 획득할 필요가 있다. 이를 위해 직접 문맥정보를 질의어로 구글 사이트를 검색하고 검색된 상위 100건의 결과의 요약텍스트에서 간접 문맥정보를 추출하였다. 간접 문맥정보 추출 방법을 위해 w와 w의 직접 문맥 정보로 추출된 각 단어 s에 대해 요약텍스트 내에서 s의 전후 N개의 단어를 추출하는 방식을 취했다(그림 4의 왼쪽 부분). 간접 문맥정보 추출 전에 구글에서 검색된 요약텍스트로부터 HTML 태그와 불용어(stopword)를 제거하는 전처리를 거쳤다. 추출된 간접 문맥정보 내의 각 단어에 대해 원형(root form)을 복원한 후 원형이 워드넷에 등록되어 있는지를 검사하여 워드넷에 등록되지 않은 단어는 문맥 단어로 고려하지 않았다. 이러한 과정을 통해, 각 WSD 대상 단어 w에 대해, w가 출현한 문맥과 유사한 문맥에서 출현한 공기(collocation)단어와 그 빈도수를 문맥정보로 획득할 수 있다.

이 과정의 구현을 위해 구글 웹 검색결과를 JSON[6] 포맷으로 리턴하는 Google AJAX Search API[7]를 사용하였고, 불용어 제거(stopping)와 스테밍(stemming)을 위해서는 Apache Lucene[8]에 포함되어 있는 모듈을 이용하였다.

4.2 의미 구분 정보(sense differentiation information)

본 연구에서는 단어의 의미 표현을 위해서 워드넷의 synset ID를 이용하는데, 각 단어별 의미의 구분을 위해 의미에 해당하는 synset 뿐만 아니라 상위어, 하위어, 전체어, 부분어에 해당하는 synset들까지 모두 모아 synset 목록을 만들어 사용한다(그림 4의 오른쪽 부분). 이의 구현을 위해 WordNet 3.0 데이터베이스[9]와 워드넷 검색 API인 JWI

3) synset은 "handset French_telephone"이고, 주석(gloss)은 "telephone set with the mouthpiece and earpiece mounted on a single handle"임
4) synset은 "equipment"이고, 주석(gloss)은 "an instrumentality needed for an undertaking or to perform a service"임
5) [표 5] 실험결과에 의해 WSD 대상 단어의 전후 4단어씩 추출하였다.
6) http://www.json.org/
7) http://code.google.com/apis/ajaxsearch/
8) http://lucene.apache.org/java/docs/index.html
9) http://wordnet.princeton.edu/obtain

2.1.3 (MIT Java WordNet Interface)[10]을 사용하였다.

4.3 단어 의미 결정 알고리즘

단어의 의미를 결정하기 위해서는 입력문장으로부터 유도된 문맥정보(단어목록)와 단어의미별 synset 목록(단어목록)을 비교하여 유사도 값이 가장 높은 단어의 의미(synset)를 선택하는 부분이 필요하다.

이 연구에서 제안하는 WSD 대상 단어의 의미 결정 알고리즘은 다음과 같다. 먼저, WSD 대상 단어를 w에 대해 w의 총 n개 의미들 중 i-번째 의미를 $s_i(w)$라 하고, w의 문맥정보를 Context(w)라 하고, 워드넷에서 얻어지는 $s_i(w)$의 의미구분정보를 Meaning($s_i(w)$)라 하자. 또한 w의 n개 의미들 중 최다빈도 의미를 $s_{mostfreq}$라 하자. 다음으로 전체 n개 의미들 각각에 대해 Context(w)와 Meaning($s_i(w)$) 사이의 유사도 Sim(w, s_i)를 계산한다. 다음으로 w의 n개 의미들 각각에 대해 얻어진 n개 유사도들 중 최대값에 해당하는 의미를 s_{maxsim}라 하자.

의미의 빈도정보를 이용하지 않는 실험은 w의 최종 의미 s_{final}로 s_{maxsim}을 바로 결정하는 것이고, 의미의 빈도정보를 반영하는 실험은 s_{maxsim}과 $s_{mostfreq}$가 같다면 w의 최종 의미는 s_{maxsim}으로 결정하고, 만약 $s_{mostfreq}$과 s_{maxsim}가 다르다면 Sim(w, $s_{mostfreq}$)와 Sim(w, s_{maxsim})의 비율 R(=Sim(w, $s_{mostfreq}$)/Sim(w, s_{maxsim}))을 계산하여 R이 임계치를 넘는 경우 w의 최종 의미를 $s_{mostfreq}$으로 결정하고 임계치를 넘지 못하는 경우 w의 최종 의미를 s_{maxsim}로 결정하는 것이다. 문맥정보를 문장 단위로 추출하지 않고, WSD 대상 단어의 주변 문맥정보로 한정하고, 의미의 빈도정보를 반영하는 WSD 알고리즘을 다음 표로 정리하였다.

4.1절과 4.2절에서 구축한 목록들 간 유사도를 계산(Sim(w, s_i))하기 위해서는 아래와 같은 여러 유사도 척도를 사용하여 실험하였다.

먼저 의미의 빈도수를 사용하지 않는 단순 단어 매칭에 의한 중복도 계산 실험을 하였고, 또한 각 목록에 있는 모든 단어쌍 간 Pirro & Seco[11], Jiang & Conrath[12], Lin[13], Resnik[14] 등이 제안한 유사도 계산을 한 후, 평균값의 비교를 통해 의미를 결정하였다. 각 유사도 척도에 관한 내용은 다음과 같다.

워드넷에서 유사도를 계산하는 척도들은 대부분 정보 이론의 정보량(IC, Information Content)에 기반한 방법인데, c가 워드넷에 있는 개념이고, p(c)가 주어진 말뭉치에서 개념 c가 나타날 확률이라고 할 때, 정보량은 다음과 같이 정의할 수 있다.

$$IC(c) = -\log p(c) \quad (1)$$

Resnik[14]은 자주 등장하지 않는 단어가 자주 등장하는 단어보다 더 많은 정보를 가지고 있다는 전제하에, 각 개념들의 IC값을 안다면 주어진 두 개념 사이의 유사도를 계산할 수 있다고 주장하였다. 두 개념이 공통으로 가지는 정보의 양(MSCA, Most Specific Common Abstraction)에 따라 유사도가 결정되는 것인데, 다음과 같이 유사도 식을 정의하였다. S(c_1, c_2)는 개념 c_1과 c_2를 포함하는 개념들의 집합이다.

10) http://www.mit.edu/~markaf/prj/jwi/

표 3. 제안하는 WSD 알고리즘
Table 3. Proposed WSD algorithm

```
Algorithm WSD (Sentence sen, Word w) {
    입력문장 sen에서 WSD 대상단어 w의 전후 4단어씩
        추출하여 직접문맥정보 추출
    직접문맥정보로 구글 검색하여 상위 100건의 검색결과
        로부터 요약텍스트 추출
    요약텍스트에서 HTML 태그, 불용어 제거
    Context(w) = 직접문맥정보로 추출된 각 단어에 대해
                 요약텍스트 전후 4단어씩 추출하여
                 간접문맥정보 추출

    for (i = 0; i < n; i++) {
        Meaning(si(w)) = 워드넷으로부터 si(w)의 syn-
            set, 상위어, 하위어 추출
        Sim(w, si) = Context(w)와 Meaning(si(w)) 사이
            의 유사도 계산
    }

    smaxsim = n개의 Sim(w, si) 중 최대값에 해당하는
            의미
    if (smaxsim = smostfreq)
        sfinal = smaxsim
    else if (R > 임계값) {
        sfinal = smostfreq
    }
    else
        sfinal = smaxsim
}
```

w	WSD 대상 단어
sen	w가 포함되어 있는 입력문장
n	w가 가진 총 의미의 수
si(w)	w의 총 n개 의미들 중 i-번째 의미
Context(w)	w의 문맥정보 (from Google)
Meaning(si(w))	si(w)의 의미구분정보 (from WordNet)
Sim(w, si)	Context(w)와 Meaning(si(w)) 사이의 유사도
smostfreq	w의 n개 의미들 중 최다빈도 의미
smaxsim	w의 n개 의미들 각각에 대해 얻어진 n개 유사도들 중 최대값에 해당하는 의미
R	Sim(w, smostfreq) / Sim(w, smaxsim)
sfinal	WSD 과정을 거쳐 최종 선택된 단어 w의 의미

$$sim_{res}(c_1, c_2) = \max_{c \in S(c_1, c_2)} IC(c) \quad (2)$$

Jiang & Conrath[12]와 Lin[13]은 Resnik 수식에서 최대 유사도 1을 얻을 수 없다는 문제점을 보완하여 각각 아래와 같은 유사도 수식을 제안하였다.

$$sim_{JC}(c_1, c_2) = 1 - \frac{IC(c_1) + IC(c_2) - 2sim_{res}(c_1, c_2)}{2} \quad (3)$$

$$sim_{Lin}(c_1, c_2) = \frac{2sim_{res}(c_1, c_2)}{IC(c_1) + IC(c_2)} \quad (4)$$

Pirro & Seco[11]는 말뭉치에 의존하는 IC 기반 유사도

계산 방법의 문제를 피하기 위해 Tversky[15]가 제안한 특징 기반의 유사도 이론을 응용하여 다음과 같은 유사도 수식을 제안하였다.

$$sim_{tvr}(c_1, c_2) = 3IC(msca(c_1, c_2)) - IC(c_1) - IC(c_2) \quad (5)$$

$$sim_{PS}(c_1, c_2) = \begin{cases} sim_{tvr} & \text{if } c_1 \neq c_2 \\ 1 & \text{if } c_1 = c_2 \end{cases} \quad (6)$$

본 연구에서는 새로운 유사도 계산 방법을 제안하는 것이 아니라, 위에서 제시된 것과 같이 기존에 많이 활용되고 있는 유사도 수식들은 그대로 이용하지만, 이의 입력 정보라 할 수 있는 확장 문맥정보와 의미구분정보를 자동으로 구축하여 실험함으로써, 기존 비교사학습 방법으로는 성취하지 못했던 최다빈도 의미 선택 방법의 성능에 가깝게 갈 수 있는 효율적인 비교사학습 방법(WSD 알고리즘)을 제안하였다. 즉, WSD에 필요한 양질의 정보(문맥정보, 의미구분정보)를 어떻게 하면 효율적으로 획득할 수 있는가에 초점을 맞추어 연구를 진행하였다.

5. 실 험

제안한 WSD 방법의 평가 및 기존 연구[6]와의 비교를 위하여 SemCor 3.0 말뭉치[11] 가운데 처음 10개의 파일을 선택하였다. SemCor는 Brown 말뭉치의 일부를 워드넷의 synset을 이용하여 수작업으로 태깅한 말뭉치로, WSD의 성능을 객관적으로 평가하기에 좋은 테스트 집합이다. 선택된 파일에는 총 1,024개의 문장과 5,463개의 명사가 포함되어 있다.

최근 WSD 시스템의 성능을 평가하는 Senseval-1, 2, 3 평가 대회 및 워크샵[12]을 통한 WSD 성능 수준을 정리해 보면, 단어의 의미를 최다빈도의 의미(most-frequent sense)로 결정하는 방법을 베이스라인(Baseline)이라 할 때, 의미 태깅된 학습말뭉치 등을 사용한 교사학습(supervised learning)의 경우는 그 최고 성능이 베이스라인의 성능을 근소하게 상회하는 정도이고, 대부분의 비교사학습(unsupervised learning) 방법은 베이스라인의 성능에 적지 않은 차이로 뒤쳐져 있다.

본 연구에서 제안한 단어 의미 결정 알고리즘에서의 Sim(w, $s_{mostfreq}$)와 Sim(w, s_{maxsim})의 비율 R(=Sim(w, $s_{mostfreq}$)/Sim(w, s_{maxsim}))은 표 4에 제시된 바와 같이 임계치가 0.4일 때 가장 좋은 성능을 보였다.

근접 문맥정보를 추출하기 위하여 WSD 대상단어의 전후 단어를 각각 N개씩 추출하여 실험하였으며, 근소한 차이이긴 하지만 전후 4단어씩 추출했을 때가 가장 좋은 결과를 보였다.

표 4와 표 5에 의해 비율 R이 0.4이고, 문맥정보의 추출단위(CW)가 4일 때, 실험 결과가 가장 좋게 나옴을 알 수 있다. 본 연구의 결과와 기존 연구결과와의 비교는 표 6에 제시되었다.

11) http://www.cs.unt.edu/~rada/downloads.html#semcor
12) http://www.senseval.org/

표 4. 비율 R에 따른 실험결과(Resnik 유사도식, CW=4 적용)
Table 4. Experimental results according to the ratio R

R Threshould (Resnik, CW=4)	Correct Nouns	Precision(%)
0.2	4459	81.6
0.3	4460	81.6
0.4	4466	81.7
0.5	4431	81.1
0.6	4336	79.4
0.7	4111	75.3
0.8	3683	67.4

표 5. 근접 문맥정보를 위한 추출 단어수에 따른 결과 (Resnik 유사도식, R=0.4 적용)
Table 5. Experimental results according to the number of extracted words for adjacent context information

Context Window Size (RESNIK 유사도, R=0.4)	Correct Nouns	Precision(%)
1	4431	81.1
2	4456	81.6
3	4463	81.7
4	4466	81.7
5	4464	81.7
6	4464	81.7

의미의 빈도정보를 전혀 이용하지 않는 경우에는 Context(w)와 Meaning(si(w)) 사이의 유사도를 단순 단어 매칭방법으로 계산한 것이 가장 좋은 성능을 보였고, 의미의 빈도정보를 반영하는 경우에는 Jiang, Lin, Resnik 등의 기존 워드넷 유사도 계산 방법을 적용한 실험이 동일하게 좋은 성능을 보였다.

본 연구는 비교사학습의 범주에 들어가지만 베이스라인에 가까운 성능을 보이고 있으며, 워드넷과 구글을 사용했던 기존 연구[6]와 비교해 볼 때도 15.8% 정도의 성능 향상을 보이고 있음을 확인할 수 있다. [6]은 WSD 대상 단어별 문맥정보를 구분하여 추출하지 않고 입력문장 전체를 사용하고 있어서, 문맥정보의 구분을 섬세하게 하지 못하고, 유사도 계산을 단순 단어 매칭에 의한 점수계산을 하기 때문에 본 연구보다 좋지 못한 결과를 보인 것으로 추정된다.

표 6. 실험결과
Table 6. Experimental results

Experiments		Total Nouns	의미의 빈도정보 불이용 (R=1)		의미의 빈도정보 이용 (R=0.4)	
			Correct Nouns	Precision (%)	Correct Nouns	Precision (%)
최다빈도 의미 선택		5463	-	-	4459	81.6
Klapaftis et al., [6]		5463	3218	58.9	3601	65.9
근접 문맥 정보 (CW=4)	단어 매칭	5463	3879	71.0	-	-
	Pirro & Seco	5463	2382	43.6	4455	81.5
	Jiang	5463	2382	43.6	4466	81.7
	Lin	5463	2382	43.6	4466	81.7
	Resnik	5463	2382	43.6	4466	81.7

6. 결론 및 향후계획

본 논문은 웹 페이지에서 자연어처리를 통하여 자동 추출된 온톨로지 개체(ontology instances)를 일반화하여 온톨로지의 지식을 확장하기 위하여, 자연어 문장을 의미 태깅(semantic annotation)하고자 할 때 필수적인 절차인 단어 의미 중의성 해소 과정을 오픈 API와 리소스를 이용하여 해결하는 방법을 제시하고 있다.

단어 의미 중의성 해소 처리 시 학습데이터의 부족으로 인한 문제를 해결하기 위하여 오픈 소스(WordNet, Apache Lucene)와 오픈 API(Google)를 최대한 활용하여 자동으로 단어 의미의 중의성을 해소하고 그 결과를 워드넷의 synset으로 표현하였다. 이를 통해 온톨로지 자동구축 도구의 일부분에 해당하는 개체 일반화 모듈을 개발하여 온톨로지의 확장 및 구축의 용이성을 높이고 자연언어처리를 위한 지식베이스인 워드넷과 구글의 검색결과를 활용한 온톨로지 개체의 일반화 자동 도구를 개발함으로써 온톨로지 구축의 생산성을 높이고 온톨로지 학습의 핵심 기술을 확보하였다.

온톨로지 개체 생성 기술은 시맨틱 기술의 산업적 적용과 성공의 핵심적 요소기술로 파악되고 있으며, 특히 각 조직 및 포털에 대규모로 축적되어 온 비구조적 정보를 시맨틱 문서로 전환할 수 있도록 함으로써 산업적, 경제적 유발효과가 크기 때문에 본 연구주제를 해결하기 위한 노력이 지금보다 더욱 많아질 것으로 전망된다.

향후 연구계획으로는 제시한 온톨로지 일반화 방법론을 한국어에 적용하는 것과 워드넷 미등록어의 처리를 위한 연구를 계속할 계획이다.

참 고 문 헌

[1] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," In *Proceedings of the 14th International Conference on Computational Linguistics*, 1992.

[2] R. Snow, D. Jurafsky, and A. Y. Ng, "Learning syntactic patterns for automatic hypernym discovery," In *Proceedings of Advances in Neural Information Processing Systems*, 2004.

[3] D. Faure, and C. Nedellec, "Knowledge acquisition of predicate argument structures from technical texts using machine learning: the system ASIUM," In *Proceedings of the European Knowledge Acquisition Workshop (EKAW)*, 1999.

[4] P. Cimiano, and S. Staab, "Learning concept hierarchies from text with a guided agglomerative clustering algorithm," In *Proceedings of ICML-2005 Workshop on Learning and Extending Ontologies by using Machine Learning Methods*, 2005.

[5] P. Buitelaar, and P. Cimiano, "Ontology learning from text," *Tutorial Notes at 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

[6] I. P. Klapaftis, and S. Manandhar, "Google & WordNet based word sense disambiguation," In *Proceedings of ICML-2005 Workshop on Learning and Extending Ontologies by using Machine Learning Methods,* 2005.

[7] P. M. Ryu, and K. S. Choi, "An Information-theoretic approach to taxonomy extraction for ontology learning," In *Ontology Learning from Text: Methods, Evaluation and Applications, Frontiers in Artificial Intelligence and Applications,* IOS Press, Amsterdam, Vol. 123, July 2005.

[8] A. Wierzbicka, *Semantic Primitives.* Frankfurt a. M.: Athenäum-Verl, 1972.

[9] C. Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, Communication),* MIT Press, May 1998.

[10] E. Agirre, and P. Edmonds, *Word Sense Disambiguation: Algorithms and Applications,* Springer, 2006.

[11] G. Pirrò, N. Seco, "Design, Implementation and Evaluation of a New Similarity Metric Combining Feature and Intrinsic Information Content". *ODBASE 2008, LNCS*, Springer Verlag, 2008.

[12] J. Jiang, and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," In *Proc. ROCLING X*, 1997.

[13] D. Lin, "An Information-Theoretic Definition of Similarity," In *Proc. of Conf. on Machine Learning*, pp. 296-304, 1998.

[14] P. Resnik, "Information Content to Evaluate Semantic Similarity in a Taxonomy," In *Proc. of IJCAI* 1995, pp. 448-453, 1995.

[15] A. Tversky, Features of similarity, *Psychological Review*, Vol. 84, No. 2, pp. 327-352, 1977.

저 자 소 개

강신재(Sin-Jae Kang)

1995년 : 경북대학교 컴퓨터공학과 공학사
1997년 : 포항공과대학교 컴퓨터공학과 공학석사
2002년 : 포항공과대학교 컴퓨터공학과 공학박사
1997년～1998년 : SK Telecom 정보기술연구원 연구원
2007년 : 오스트리아 University of Innsbruck, DERI 연구소 방문교수
2002년～현재 : 대구대학교 컴퓨터·IT공학부 부교수

관심분야 : 시맨틱 웹, 소셜 웹, 온톨로지, 자연어처리
Phone : 053-850-6584
E-mail : sjkang@daegu.ac.kr

강인수(In-Su Kang)

1995년 : 경북대학교 컴퓨터공학과 공학사
1999년 : 포항공과대학교 컴퓨터공학과 공학석사
2006년 : 포항공과대학교 컴퓨터공학과 공학박사
1995년～1997년 : 포스데이타
2006년～2008년 : 한국과학기술정보연구원
2008년～현재 : 경성대학교 컴퓨터정보학부

관심분야 : 자연어처리, 정보검색, 시맨틱 웹
Phone : 051-663-5147
E-mail : dbaisk@ks.ac.kr