온톨로지 구축 및 단어 의미 중의성 해소에의 활용

강 신 재†

요 약

본 논문은 기존의 다양한 언어자원들을 이용하여 온톨로지를 구축하고, 이를 단어의미 중의성 해소에 활용하는 방법을 제시하고 있다. 온톨로지를 실용적으로 구축하기 위해서는 가도카와 시소러스의 개념 체계에 격 관계와 기타 의미관계와 같은 다른 의미관계를 추가하여 확장하는 방법을 선택하였다. 구축된 온톨로지를 단어 의미 중의성 해소에 활용하기 위해서는, 결합가 정보를 포함하고 있는 전자사전을 먼저 이용하여 단어의 의미를 결정하고, 결정하지 못한 단어의 의미는 온톨로지를 이용하여 결정하는 절차를 거친다. 이를 위해 온톨로지 내 개념들간의 상호정보가 말뭉치의 통계 정보에 근거하여 계산되는데, 이를 가중치로 간주하면 온톨로지는 가중치 그래프로 생각할 수 있으므로, 개념간 경로를 통하여 개념간 연관도를 알아 볼 수 있다. 실제 기계번역 시스템에서 본 방법은 온톨로지를 사용하지 않은 방법보다 9%의 성능 향상을 가져오는 결과를 얻을 수 있었다.

Ontology Construction and Its Application to Disambiguate Word Senses

Sin-Jae Kang†

ABSTRACT

This paper presents an ontology construction method using various computational language resources, and an ontology-based word sense disambiguation method. In order to acquire a reasonably practical ontology, the Kadokawa thesaurus is extended by inserting additional semantic relations into its hierarchy, which are classified as case relations and other semantic relations. To apply the ontology to disambiguate word senses, we apply the previously-secured dictionary information to select the correct senses of some ambiguous words with high precision, and then use the ontology to disambiguate the remaining ambiguous words. The mutual information between concepts in the ontology was calculated before using the ontology as knowledge for disambiguating word senses. If mutual information is regarded as a weight between ontology concepts, the ontology can be treated as a graph with weighted edges, and then we locate the weighted path from one concept to the other concept. In our practical machine translation system, our word sense disambiguation method achieved a 9% improvement over methods which do not use ontology for Korean translation.

키워드 : 온톨로지(Ontology), 단어의미 중의성 해소(Word Sense Disambiguation), 말뭉치 분석(Corpus Analysis), 기계번역(Machine Translation), 세종 전자사전(Sejong Electronic Lexicon of Korean)

1. 서 론

한국어정보처리에서 형태소 분석/구문 분석 이상의 처리를 하기 위해서는 온톨로지(ontology)라 불리는 의미지식베이스가 반드시 필요하기 때문에, 최근에는 이에 관련된 연구가 활발해지고 있다. 하지만 온톨로지내 개념의 크기(granularity)와 수, 지식표현문제, 구축 방법 등 여러 문제들로 인하여 아직까지 그 성과는 미미한 실정이다. 본 연구에서는 시소러스(thesaurus), 기계번역사전, 세종전자사전, 말뭉치 등 기존 언어자원들을 최대한 활용하여 온톨로지를 구축하여, 단어의미 중의성 해소에 활용해 보고자 한다.

※ 이 논문은 2002학년도 대구대학교 학술연구비 지원에 의한 논문임.
† 정 회 원 : 대구대학교 컴퓨터 · IT공학부 교수
논문접수 : 2004년 3월 12일, 심사완료 : 2004년 6월 11일

일반적으로 시소러스와 온톨로지의 구별을 하지 않고 사용하는 연구자들도 많으나, 본 연구에서는 다음과 같이 구별하여 사용하고자 한다.

시소러스란 "통제된 색인언어의 어휘집으로, 개념간의 특정 관계를 형식적으로 조직화하여 명시한 것"으로 초기 문헌정보학에서 어떤 문헌에 대한 색인 작업 시 적절한 색인 표목의 선택과 색인어의 통제를 위해 사용될 뿐만 아니라 검색 시에는 적절한 탐색어의 선택을 위해 사용되었으며, 점차 응용 영역이 확대되어 정보 검색, 전자상거래, 전문가 시스템, 자연언어처리 등의 여러 분야에서 다루어지고 있다[3]. 자연어처리에서의 시소러스는 상하위 개념어, 동형이의어, 다의어, 반의어, 부분-집합어, 관련어 등으로 구성되며, 대표적인 시소러스로는 WordNet[13], Goi-Taikei[22] 등이 있다.

온톨로지에 대한 사전의 전형적인 정의는 "존재의 본질을 연구하는 형이상학의 한 갈래"이나, 자연어 처리의 관점에서 보는 온톨로지는 "실세계(혹은 특정 도메인)에 존재하는 모든 개념들(concepts)과 그 개념들의 속성들(properties), 그리고 개념들이 상호간 의미적으로 어떻게 연결되어 있는가(semantic relations)에 대한 정보를 가지고 있는 지식베이스(knowledge base)"라 정의할 수 있다. 기계번역(machine translation)에서 온톨로지를 사용하는 주된 이유는 원시 언어 분석기(source language analyzer)와 목표 언어 생성기(target language generator)간 정보 교환 시 매개의 역할을 하며, 개념간 의미 제약(semantic constraint)을 저장하고 있는 온톨로지 개념망의 추론을 통하여 의미 중의성을 해소하기 위함이다[19, 23]. 온톨로지는 언어 독립적인 정보만 저장하고 있어서 지식 공유와 재사용을 중요시한다는 점과, 개념간 의미관계가 계층관계(taxonomic relation), 격관계, 동의관계 외의 "has-member, material-of, represent"와 같은 다른 다양한 의미관계도 포함하고 있다는 점에서 시소러스와 구별될 수 있다.

문장 중의 단어는 대부분 다른 단어와의 관계에 의해 하나의 유일한 의미로 결정될 수 있다. 이처럼 의미의 중의성이 존재하는 동형이의어나 다의어의 의미를 결정하는 과정을 단어 의미 중의성 해소(word sense disambiguation ; WSD)라 한다. WSD를 위해서는 어떤 자원을 어떻게 사용할 것인가를 결정해야 한다. 단어 의미 중의성 해소에서 사용될 수 있는 정보의 유형들을 정리해 보면, 품사 정보, 형태소 정보, 연어 정보, 의미 관계(계층 구조, 유의어 등), 구문 정보, 의미역 정보(semantic roles), 선택 제약 정보(selectional preferences), 도메인 정보, 빈도수 정보, 화용 정보 등이 있다. 본 연구에서는 온톨로지를 의미 분석 단계에서의 핵심 작업인 단어 의미 중의성 해소에 활용하여 그 성능을 향상시키고자 한다.

2. 온톨로지 및 WSD 관련연구

마이크로 코스모스(Mikrokosmos)[19]는 미국방성의 지원 아래 미국 뉴멕시코 주립대학에서 개발되고 있는 지식기반의 기계번역 시스템이다. 기존 다국어 기계번역 시스템들과는 달리 마이크로 코스모스는 대규모의 실용적인 기계번역을 지향하고 있다. 약 5,000개의 개념을 포함한 온톨로지를 이용하여 7,000여 단어의 스페인어 사전을 구축하여, 회사간 인수/합병에 관한 스페인어 기사에 대해 고품질의 의미 분석이 가능하다. 마이크로 코스모스 온톨로지는 기본적으로 방향성 그래프(directed graph)의 형태를 가지고 있다. 각각의 개념은 노드(node)에 해당하여 프레임(frame)으로 표현되며, 개념간의 관계는 슬롯(slot)과 필러(filler)로 구성된 링크(link)로 표현된다. 프레임은 명명된 슬롯의 집합으로 구성되는데 슬롯은 해당 개념(프레임)이 다른 개념과 어떤 관계로 연관되어 있으며, 또 어떤 특성을 가지고 있는지를 나타낸다. 또 마이크로 코스모스 온톨로지는 개념지식과 에피소드 지식을 명확하게 구분하여 개념지식만 포함하고 있다. 인스턴스와 에피소드를 따로 저장하고 있는 지식베이스가 바로 오노메스티콘(onomasticon)인데, 주로 고유명사와 같이 인스턴스화된 개념을 가지고 있다. 온톨로지에는 "CITY"라는 개념이 있고, 오노메스티콘에는 "SEOUL"이 저장되어 있으며, 각 언어 사전에는 seoul(영어사전), 서울(한국어사전), ソウル(일어사전)와 같이 저장되어 있는 것을 예로 들 수 있다.

EDR[15]은 일영/영일 기계 번역을 위해서 일본 정부기관과 8개의 컴퓨터 관련 업체가 참여하여 개발한 것으로, 단어사전, 개념분류사전, 개념기술사전 등으로 구성된다. 단어사전에는 표제어 정보, 문법정보, 개념 코드와 같은 의미정보를 포함하고 있으며, 개념을 구조화한 개념사전에는 약 6,000개의 개념에 대해 개념 코드(concept identifier)와 개념 코드를 자연어로 표현한 표제 개념(head concept), 개념의 설명(concept explication)이 포함되어 있다. EDR 개념의 분류는 미리 분류범주를 정한 것이 아니라 말뭉치에서 각 개념의 용례색인을 추출하여 형태와 문맥이 비슷한 것끼리 우선 대분류한 다음 차례로 세분해 나가는 방식을 취하였다.

HowNet[28]은 중영 기계번역 시스템의 개발을 위해 만들어진 중국어 온톨로지이다. 우선, 개념 분류를 위해서 중국어가 뜻글자임을 이용하여 다음과 같은 의미소(sememe) 결정 과정을 거친다. 용언을 예로 들면, 가장 많이 쓰이는 6,000여 중국어 한자를 검토하여 뜻이 유사한 한자들을 모아나가는 과정을 통하여 우선 3,200개의 그룹을 만들고, 또 다시 합치는 과정을 통하여 최종적으로 800여 개의 의미소를 정의하였다. 이러한 과정을 거쳐 HowNet에서는 개체(entity) 142개, 사건(event) 813개, 속성(attribute) 117개, 속성이 가질 수 있는 속성값 433개, 구문 관계 52개, 의미관계 69개로 분류하여 사용하고 있다. HowNet은 다른 지식베이스에 비해 상당히 자세한 분류 정보를 가지고 구축되었으나, 온톨로지의 구축이 한 사람의 전문가에 의해서 이루어진데다가 그 품질이 객관적으로 증명되지는 않은 상태이다. 또 중국어에 전적으로 의존하여 만들어졌기 때문에 다른 언어에서 사용하기에는 어려움이 있을 수 있다.

SENSUS[18]는 USC/ISI에서 Pangloss 기계번역 프로젝트의 의미 분석을 위하여 구축한 대규모의 개념망이다. 이것은 PENMAN Upper Model, ONTOS, LDOCE, WordNet, 콜린스 대역 사전과 같은 온라인 사전, 의미망, 대역어 사전을 이용하여 구축되었다. 이 방법은 기존 자원을 활용하여 비교적 쉽게 온톨로지를 구축하는 방법을 제시하였으나, 결과물이 WordNet과 마찬가지로 영어에 의존적인 지식베이스이기 때문에 다른 언어에서의 사용을 위해서는 영어와 해당 언어의 대역 사전을 확보하여 SENSUS와 병합하는 과정을 더 거쳐야 한다.

CYC[9]는 MCC에서 10여년의 기간을 통하여 구축한 대규모의 일반 상식(common sense) 지식베이스를 가리키는데, 약 10만개의 개념 정의와 100만개의 사실(fact)/규칙(rule)과 추론 엔진(inference engine)을 가지고 있다. CYC는 지식베이스, 환경(편집/검색 도구 인터페이스 등), 지식표현언어(CycL)의 3개의 구성 요소로 이루어져 있다. 가장 일반적인 개념 3,000개를 정의하여 이를 Upper CYC 온톨로지라 부르고, 이 상위 온톨로지 밑에 방대한 양의 사실들이 연결되어 있다. CYC는 평범한 문맥에서의 일반적인 의미만을 포함하고 있기 때문에 특정 도메인 지식베이스를 구축하기 위해서 재사용하기에는 어려움이 있으며, 또 CYC를 여러 자연어처리 시스템에서 어떻게 활용하는가에 대한 연구가 부족한 실정이다.

단어 의미의 중의성을 해소하기 위해서는 지금까지 많은 연구들이 진행되어 왔는데, 이들은 사용하는 데이터의 형태에 따라서 지식베이스(사전, 시소러스 등)를 이용하는 방법과 말뭉치를 이용하는 방법으로 분류할 수 있고, 방법론에 따라서는 크게 규칙을 이용한 방법, 확률 통계를 이용한 방법 등으로 분류할 수 있다[20].

지식베이스를 이용한 방법은 기계가독사전(machine-readable dictionary, MRD)과 시소러스 같은 지식베이스를 사용하는데, 주로 사전의 뜻풀이말이나 예문, 주제 코드 등을 이용한다. 이러한 리소스가 확보된 경우 손쉽게 적용해 볼 수 있다는 이점이 있으나, 사용하는 대부분의 사전이나 시소러스가 전산처리의 목적이 아닌 일반 사용자를 위한 목적으로 수작업에 의해 만들어졌기 때문에 정보의 일관성에 문제가 있을 수 있으며, 또 사전의 정의나 기술에 제한적인 어휘를 사용하기 때문에 실제 문장에 적용하기에는 한계를 가지고 있다. Roget와 같은 시소러스에서는 의미 계층 구조 및 유사어의 공기 정보 등을 주로 이용하며, WordNet과 같은 지식베이스를 사용하는 경우는 표제어 정의, 유사어, 의미 계층 구조, 전체/부분어 등과 같은 기타 개념 관계들을 이용한다. Agirre[12]와 Li[27]는 WordNet에서의 개념간 유사 관계를 이용했으며, Yarowsky[11]는 자료 획득 병목 현상을 최소화하기 위해 Roget 시소러스의 카테고리를 기반으로 한 통계 모델을 제안하였다.

말뭉치를 이용하는 방법은 원시 말뭉치에서 통계 정보를 추출하는 방법과 의미가 태깅된 말뭉치에서 보다 정확한 통계 정보를 추출하여 사용하는 방법이 있다. 의미 태깅된 말뭉치를 사용하는 방법은 구축하기에 많은 시간과 비용이 든다는 단점과, 상세한 의미 구분이 가능하여 비교적 정확한 통계를 얻을 수 있다는 장점이 있다. 반면에 원시 말뭉치를 이용하는 방법은 대용량의 말뭉치를 확보할 수 있으나, 상세한 의미 구분이 불가능하다. 말뭉치를 이용하는 방법은 일반적으로 자료 부족 문제가 야기되는 문제점이 있으며, 언어의 동적인 특성 즉, 실제 사용되는 문장들의 성격을 잘 반영할 수 있다는 장점이 있다. 이 방법은 주로 확률 통계를 이용하는 방법론과 같이 사용된다.

Yarowsky[10]는 각 담화에서 단어는 하나의 의미를 가지며(one sense per discourse), 또 각 연어에서 단어는 하나의 의미를 가진다(one sense per collocation)는 제약을 이용하였다. 먼저 대상 단어의 예제 문장을 수집하여 대상 단어의 각 의미를 잘 나타낼 수 있는 초기 연어(seed collocation)를 선정하고 초기 연어가 포함된 문장을 중심으로 학습하여 새로운 연어를 추출해 나가는 방법으로 연어를 계속 확장해 간다. 이 방법은 원시 말뭉치를 사용하는 까닭에 세분화된 의미 구분이 불가능하다는 문제점이 있다. Leacock[8]은 WordNet과 원시 말뭉치를 이용한 의미 중의성 해소 방법을 제안하였다. WordNet의 유의어 집합에서 중의성이 없는 유의어를 중의성 해소에 사용하는데, 먼저 대상 단어의 각 의미별로 중의성이 없는 유의어를 찾아내고 이 유의어들을 포함하는 문장들을 대량의 말뭉치에서 뽑아내어 사용한다. 이 방법은 WordNet을 한국어에 바로 적용할 수 없다는 어려움이 있다.

규칙을 이용한 방법은 초창기 자연어 처리 연구에서 주로 이루어졌던 방법으로, 언어의 깊은 이해를 통해 얻은 일반적인 규칙을 수작업으로 구축하여 사용하는 방식이다. 이는 규칙의 구축이 매우 힘들뿐 아니라, 적용되는 도메인에 따라 규칙을 새로이 만들어야 하고, 모든 문장을 처리할 수 있는 견고한 규칙을 만들기 어렵고, 또한 구축된 규칙의 일관성을 보장하기도 어려운 한계점을 가지고 있다.

한국어에서의 WSD에 대한 연구들도 말뭉치와 사전 등을 이용하고 있다. 조정미[6]는 말뭉치와 사전을 이용한 WSD 방법을 제안하였다. 간단한 사전 분석을 통해 의미 분별하고자 하는 단어의 의미 지시자와 단어의 분류 정보를 추출하였으며, 말뭉치로부터는 정규화 과정을 거친 다음, 목적어 관계의 선택 제한 지식을 이용하여 단어간 유사성을 학습하였다. 또한, 말뭉치의 자료 부족 현상을 해소하기 위하여 선택 제한 지식을 명사 분포와 동사 분포로 표현하고 이를 이중적으로 의미 분별에 적용하였다. 그러나, 선택 제한 지식으로 목적격만 사용하였으며, 또 사전으로부터의 지식 획득이 수작업에 의존하기 때문에 정보의 구축이 어렵다는 문제점이 있다.

서희철[4]은 기존의 의미 계층 구조를 이용한 방법에서 고려하지 않았던, 유사어의 개별 특징을 고려해서 다의어의 의미를 결정하는 방법을 제시하였다. 품사 정보가 부착된 1,000만 어절의 말뭉치에서 유사어의 용례를 추출하여 학습 데이터로 사용하였다. 한국어 명사 '배', '밤', '고개'에 대해서 실험을 하였는데, 의미 벡터를 이용한 방법보다 평균 9.5% 정도의 성능향상이 있었다. 그러나 시스템의 확장을 위해서는 보다 신뢰성 있는 의미 계층 구조의 정의와 더 많은 양의 학습 데이터가 필요하다는 문제점이 있다.

이승우[5]는 한국어에 대한 의미 부착 말뭉치가 거의 없는

실정을 감안하여 비교사(unsupervised) 학습 방식을 채택하였는데, 초기의 공기 정보를 사전으로부터 학습하고 원시 말뭉치로부터 국소 문맥을 학습하였다. 또한 단어의 의미를 결정하는 과정에서 국소 문맥 데이터베이스 단서들의 의미를 학습하고 공기 정보를 추가적으로 학습하였다. 또 두 단어 사이의 의미 유사도 계산을 위해 의미 계층 구조를 WordNet으로부터 차용하였다. 한국어 명사 '가사'와 '공사'를 대상으로 실험하였는데, 기존의 교사(supervised) 학습 방법보다 나은 89.8%의 정확률을 보였다. WordNet에서부터 한국어 명사의 의미 계층 구조를 얻는 과정이 쉽지는 않지만, 사용할만한 한국어 의미 계층 구조가 없고, 의미 부착 말뭉치가 부족하다는 점을 고려해 볼 때, 흥미 있는 시도라 할 수 있다.

허정[7]은 사전의 뜻풀이말에서 추출한 통계적 의미정보에 기반한 동형이의어 중의성 해결 시스템을 제안하였다. 5,246 문장의 사전 뜻풀이말을 학습 코퍼스로 하여 의미정보를 추출하였으며, 의미정보 내의 명사와 용언을 동시에 고려하여, 비교적 작은 말뭉치라 할 수 있는 사전만을 이용하여 제한된 의미 계층 구조를 유추하고 이용함으로써 대용량의 의미 계층 구조가 없는 경우에 적합한 모델이다. 체언과 용언 의미정보를 모두 고려한 모델로, 체언과 용언이 동형이의어 중의성 해결에 영향을 주는 정도를 결정하기 위하여 9개의 동형이의어 명사를 대상으로 실험하였다. 이 방법 역시 시스템의 확장 시 수작업으로 의미 정보를 구축해야 하고, 구문 정보를 중의성 해결에 사용하지 않았다는 점이 아쉬운 부분이다.

이상과 같이 수행되어온 기존 WSD 연구들을 볼 때에 Roget의 시소러스나 WordNet 정도를 WSD에 이용한 연구는 있지만 실제 온톨로지를 WSD에 활용한 연구는 거의 없는 실정이다. 이는 온톨로지의 구축 자체가 어렵기 때문으로 보인다. 기존 연구들을 토대로 단어 의미 중의성 해소에서 사용할 수 있는 정보의 유형들을 정리해 보면, 품사 정보, 형태소 정보, 연어 정보, 의미 관계(계층 구조, 유의어 등), 구문 정보, 의미역 정보(semantic roles), 선택 제약 정보(selectional preferences), 도메인 정보, 빈도수 정보, 화용 정보 등이 있다. 지금까지의 연구들의 결과는 평가 대상 및 평가 기준 등이 모두 다르기 때문에, 수치 결과를 직접 비교할 수는 없지만, 일반적으로 수작업으로 태깅된 말뭉치를 이용하여 공기 정보, 구문 정보 등을 추출하여 사용한 방법이 가장 좋은 결과를 보이고 있다.

3. 온톨로지 구축

단어 의미 중의성 해소를 위한 실용적인 온톨로지를 구축하기 위해, 다음과 같은 두 가지 전략을 세웠다.

첫째, 가도카와 시소러스[24]에서 사용하는 개념과 그 계층구조를 그대로 도입한다[1]. 가도카와 시소러스는 총 1,110개의 개념과 4단계의 계층구조를 가지고 있으며, L1, L10, L100 레벨에 속해 있는 개념들은 각각 10개의 하위 개념들로 나뉜다. 비록 가도카와 시소러스가 일본어를 대상으로 만들어지긴 하였으나, 개념 부류가 1,110개 정도이기 때문에 일본어에만 존재하는 개념에 의해 개념 부류가 독특하게 나뉘어졌다고는 볼 수 없다. 만약, 개념 부류의 수가 더 많았다면 그러할 가능성은 높아질 것이다. 즉, 1,110개의 부류는 다른 언어에 대해서도 그대로 활용될 수 있으며, 이는 추후 연구결과를 통하여 입증될 것이다. 또한 실험에 사용될 COBALT-J/K와 COBALT-K/J 기계번역 시스템[2]의 전자사전에는 이미 각 표제어의 의미별로 가도카와 시소러스의 의미 코드가 포함되어 있기 때문에, 향후 온톨로지 활용 시 별도의 사전 작업없이 온톨로지의 적용 및 평가가 가능하다. 이러한 접근은 실용적인 온톨로지를 구축하기 위해서는 필수적이라고 할 수 있다. 게다가 가도카와 시소러스는 COBALT-J/K와 COBALT-K/J 기계번역 시스템에서 어휘 중의성 해소의 효과가 이미 입증된 상태이다[14].

두 번째 전략은 가도카와 시소러스의 계층구조에 다른 다양한 의미 관계를 추가하는 것이다. 추가될 의미 관계는 격 관계와 기타 의미 관계로 나눌 수 있는데, 격 관계는 결합가 정보와 격틀의 형태로 기존 연구에서 어휘 중의성 해소에 많이 사용되어 왔으나, 기타 의미 관계는 다른 의미 관계들과의 구별이 용이하지 않아서 그다지 사용되지 못했었다. 이를 위해서 세종 전자사전[1]과 마이크로 코스모스 온톨로지[19]를 주로 참고하여 총 30개의 의미 관계를 정의하였다 (<표 1>).

〈표 1〉 온톨로지에 포함된 의미관계 유형

대범주	소 범 주
계층 관계	is-a
격 관계	agent, theme, experiencer, companion, instrument, location, source, destination, reason, appraisee, criterion, degree, recipient
기타 의미관계	has-member, has-element, contains, material-of, headed-by, operated-by, controls, owner-of, represents, symbol-of, name-of, producer-of, composer-of, inventor-of, make, measured-in

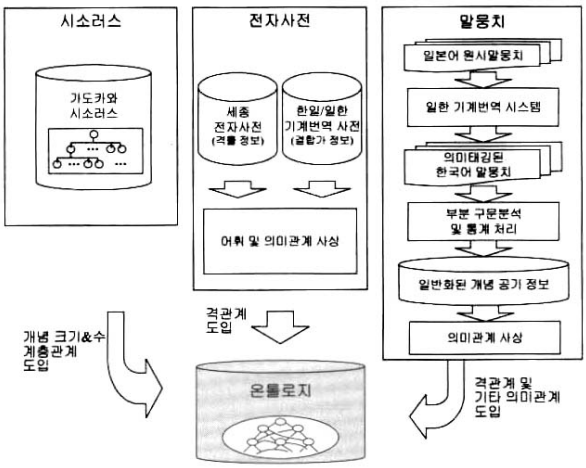
물론 이 30개의 의미 관계만으로 개념간에 존재하는 모든 의미 관계 유형들을 나타낼 수는 없으나, 단어의 의미 중의성 해소에 도움을 줄 수 있는 것들을 우선적으로 선택하여 사용하였다.

온톨로지 구축 절차를 전체적인 그림으로 나타내면 (그림

1) 루트 노드(root node)는 더미 노드(dummy node)이며, 명사와 동사의 분류는 하나의 계층구조에 공존한다. 동사의 의미 부류는 주로 L1000 레벨의 의미 코드 2xx, 3xx, 4xx에서 나타난다.

2) 포항공과대학교 지식 및 언어공학 연구실에서 개발한 일한/한일 기계번역 시스템이다.

1)과 같다.



(그림 1) 온톨로지 구축 절차

계층 관계 외 다른 다양한 의미 관계를 얻기 위해서는 두 가지 방법을 사용하였는데, 기존 전자사전에 포함되어 있는 의미 정보의 활용과 말뭉치의 반자동 분석[14]이 그것이다. 개념간의 격 관계는 주로 세종전자사전과 COBALT-J/K, COBALT-K/J와 같은 기계번역사전에 포함되어 있는 격틀 정보(case frame)와 결합가 정보(valency information)를 변환하여 얻을 수 있으며, 의미 태깅된 말뭉치 분석의 결과인 개념 공기정보를 통하여 온톨로지에 추가될 의미 관계를 추출할 수 있다.

세종 전자사전[1]은 품사에 따라 여러 개의 하위 사전으로 구성되어 있는데, 본 연구에서는 동사사전과 형용사사전을 이용하였다. 세종 동사/형용사 사전으로부터 7,526개의 격틀 정보를 추출한 후, 격틀 정보에 포함되어 있는 어휘들의 가도카와 의미코드를 COBALT-K/J의 기계번역사전을 이용하여 자동으로 변환하여, 총 6,224개의 격 관계패턴을 추출하였다. 숫자가 줄어든 이유는 중복된 패턴이 있기 때문이다.

한-일/일-한 기계번역 사전으로부터는 20,580개의 동사와 형용사 표제어에서 16,567개의 결합가 정보를 추출한 후, 결합가 정보에 포함되어 있는 구문관계를 사상 규칙(mapping rule)과 작업자의 직관에 의해 <표 1>에서 정의한 격 관계로 사상하여, 총 15,956개의 패턴을 얻었다[2].

온톨로지에 추가할 의미 관계 패턴들을 얻기 위한 또 다른 방법으로 의미 태깅된 말뭉치의 분석을 통해 개념 공기정보(concept co-occurrence information, CCI)를 추출하는 것을 생각해 볼 수 있다. 일한 기계번역 시스템(COBALT-J/K)[17]을 사용하여 의미 태깅된 한국어 말뭉치를 생성하였는데, 이 시스템은 포항제철에서 철강에 관한 일본어 특허 문서를 번역하는 데에 성공적으로 적용된 바 있다. COBALT-J/K가 번역을 수행할 때 내부적으로 중의성 해소를 위해 사용하는 가도카와 시소러스의 코드들을 한국어 어휘의 생성시 같이 출력되게 하였으며, 총 25만 일본어 문장을 분석하여 의미 태깅된 한국어 말뭉치를 생성하였다. 중의성을 가지는 명사와 의미적 제약을 가지는 공기정보들은 일정한 구문관계를 가지고 한 문장 속에 출현하기도 하고, 특정 구문관계를 갖지는 않지만 의미적으로 제약을 주는 것도 있다. 따라서, 의미 태깅된 말뭉치를 부분 구문분석한 후, 스캐닝(scanning)하는 과정을 통하여 이러한 패턴들을 얻은 후, 잡음(noise)을 제거하여 변별력이 좋은 일반화된 개념 공기정보를 얻게 된다[14]. 결과적으로 9,245개의 개념 공기정보로부터 3,701개의 격 관계와 1,650개의 기타 의미관계를 생성하였다.

위와 같은 과정을 거쳐 추출한 의미관계 패턴들은 여러 번 중복되어 나타나기도 하지만, 각 패턴들은 온톨로지에 유일하게 한번만 추가되었다. 온톨로지에 최종적으로 포함된 의미관계의 수는 <표 2>와 같다.

〈표 2〉 온톨로지에 추가된 의미관계 패턴의 수

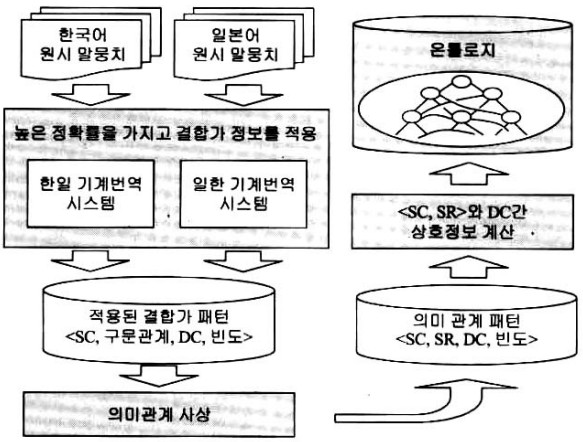
의미관계 유형	패턴 수
계층 관계	1,100
격 관계	112,746
기타 의미관계	2,093
계	115,939

4. 온톨로지 학습

이전의 연구들은 온톨로지를 어떻게 구축할 것인가에만 초점이 맞추어져 있는 반면, 그 활용에 관한 연구는 미진하였다. 온톨로지를 이용해 추론하기 위해서는 지배소(governor)와 의존소(dependent)의 역할을 하는 개념간에 온톨로지 내에서 의미 제약을 얼마나 잘 만족하는가를 평가할 수 있는 방법이 필요하다. 본 연구에서는 개념간 연관도를 측정하기 위해 상호정보(mutual information)를 사용하였다. 상호정보는 두 랜덤변수 간 의존도를 측정하는 방법으로 Church & Hanks[16]가 제안하였다. Resnik[21]은 IS-A 계층구조에서 상호정보에 기반을 둔 의미 유사도의 측정법을 제시하였는데, 본 연구는 IS-A 관계뿐 아니라 다른 의미 관계들도 유사도 측정에 사용한다는 점에서 Resnik의 연구와 구별된다고 할 수 있다. 먼저, 상호정보를 의미 연관도 측정에 사용하기 위해서, 지배소 개념(governor, source concept : SC)과 의미 관계(semantic relation : SR)를 하나의 단위로 묶고, 의존소 개념(dependent, destination concept : DC)을 독립적으로 하나의 단위로 간주하여 사용하였다. 왜냐하면, 의미 관계는 의존소 개념보다 지배소 개념에 의해 더 많은 영향을 받기 때문이다. 그러므로, <SC, SR>과 DC의 확률이 각각 P(<SC, SR>)과 P(DC)라 가정하면, <SC, SR>과 DC의 상호정보 I(<SC, SR>, DC)는 다음과 같이 정의할 수 있다.

$$I(\langle SC, SR\rangle, DC) = \log_2\left(\frac{P(\langle SC, SR\rangle, DC)}{P(\langle SC, SR\rangle)\,P(DC)} + 1\right) \quad (1)$$

온톨로지를 단어 의미 중의성 해소에 사용하기 위해서는 온톨로지에 존재하는 모든 개념간 상호정보 값을 미리 확보하고 있어야 한다. (그림 2)는 <SC, SR, DC, 빈도수> 형태의 학습 데이터를 생성하는 과정을 보이고 있다. 가도카와 시소러스의 의미코드가 태깅된 결합가 정보 패턴을 얻기 위해, 기존 일한/한일 기계번역 시스템을 약간 수정하였다. 7,000만 어절의 KIBS(Korean Information Base System, 1994~1997) 한국어 원시 말뭉치와 81만 문장의 일본어 원시 말뭉치를 분석하여 의미 태깅된 결합가 정보 패턴을 추출하였다. 추출된 결합가 정보 패턴 중 구문관계 정보를 온톨로지에 정의된 의미 관계로 변환하면, 빈도수를 가진 <SC, SR, DC> 패턴을 얻을 수 있다[2]. 이 결과를 가지고 온톨로지 내 개념간 상호정보를 계산하게 된다.

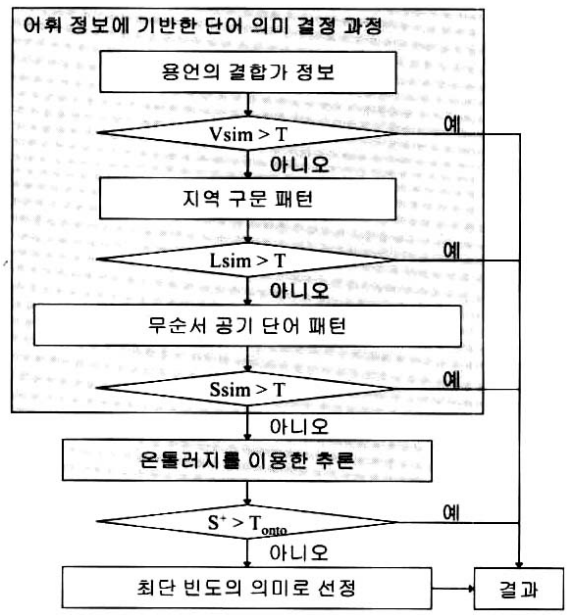


(그림 2) 온톨로지 학습 데이터의 생성

5. 온톨로지 기반 WSD

(그림 3)은 온톨로지를 사용한 전체적인 단어 의미 중의성 해소 방법을 보이고 있다. 먼저, 의미 중의성이 있는 단어들에 대해서, 전자사전에 코딩되어 있는 용언의 결합가 정보, 지역 구문 패턴, 무순서 공기 단어 패턴들을 순서대로 적용시켜 본다. 이 경우 적용된 결과가 정확하다고 추정되는 경우만 단어의 의미로 결정하게 되고, 그렇지 않으면 온톨로지를 적용하는 단계로 넘어가게 된다. 만약, 여기에서도 의미가 결정되지 못한다면, 최후의 선택으로 최다빈도의 의미를 단어의 의미로 선정하게 된다.

전자사전에 코딩되어 있는 용언의 결합가 정보, 지역 구문 패턴, 무순서 공기 단어 패턴을 이용한 중의성 해소는 식 (2)~식 (5)에 의해 구현된다[14]. S(W)는 중의성 명사 W의 의미 집합이며, SR(V)는 W와 입력 문장에서 같이 나타나는 동사 V의 선택제약 집합이며, LSP(W)는 구문관계 패턴 정보를, 그리고, UCW(W)는 W의 무순서 공기 단어 패턴 정보를 의미한다. C_i와 P_j는 개념 유형을 표현하고, S_k는 W의 k번째 의미를 뜻한다. 또, 'n'은 어휘 W의 의미 개수, 'm'은 문장에서 동사 V의 W의 격에 해당하는 선택제약의 개념코드 수, 'w'는 의미 S_k의 j번째 구문관계패턴의 개념코드 개수이며, 'r'은 의미 S_k의 무순서 공기 단어 패턴의 개념코드 개수이다.



(그림 3) 제안하는 단어 의미 중의성 해소 방법

$$Csim(C_i, P_j) = \frac{2 * level(MSCA(C_i, P_j))}{level(C_i) + level(P_j)} * weight \quad (2)$$

$$Vsim(S(W), SR(V)) = \max_i(Csim(C_i, P_j)),\quad 1 \le i \le n; 1 \le j \le m; C_i \in S(W); P_j \in SR(V) \quad (3)$$

$$Lsim(S(W), LSP(W)) = \max_k(Csim(C_i, P_{k,j,l})),\quad 1 \le i, k \le n; 1 \le j \le 10; 1 \le l \le w; P_{k,j,l} \in LSP_j(S_k) \quad (4)$$

$$Ssim_i(S(W), UCW(W)) = \max_k(Csim(C_i, P_{k,j})),\quad 1 \le i, k \le n; 1 \le j \le r; P_{k,j} \in UCW(S_k) \quad (5)$$

식 (2)의 $Csim(C_i, P_j)$는 가도카와 시소러스에 기반하여 개념 C_i와 P_j 사이의 유사도를 계산하는 식이다. 식 (2)에서 weight는 개념의 가중치를 의미하며, 유사도 계산시 개념 C_i의 부모(parent) 개념이 형제(sibling) 개념보다 유사한 특징을 더 많이 가지고 있으며, 이러한 관계를 중요시 한다는 것을 뜻한다. 즉, 개념 C_i가 P_j의 하위 개념이면 weight를 1로 지정하고, 그렇지 않으면 0.5의 값을 지정하여 유사도 값을 감소시킨다. 또, 식 (2)의 MSCA(Most Specific Common Ancestor)는 두 개념이 공유하고 있는 가장 가까운 상위 개념을 가리킨다. 용언의 결합가 정보와의 일치 여부를 식 (2)로 계산한 후, 성공여부를 결정하는 임계치는 실험에 의해

0.3으로 설정하였다.

중의성을 지닌 단어 W에 대한 의미 결정과정을 위의 식들을 이용하여 자세히 설명하면 다음과 같다.

(단계 1) 단어 W가 동사이거나 문장에 존재하는 동사 V의 논항으로 사용된 경우, 식 (3)을 이용하여 선택제약(Selectional Restriction, SR) 만족도 Vsim(S(W), SR(V))를 계산한다. 계산된 값이 임계치 T보다 크면 Csim 값을 최대로 하는 의미 C_i를 W의 의미로 결정하고, 그렇지 않으면 다음 단계로 넘어 간다.

(단계 2) W의 구문관계 패턴을 적용하는 식 (4)를 이용하여 Lsim(S(W), LSP(W)) 값을 계산한다. 계산된 값이 임계치 T보다 크면 Lsim 값을 최대로 하는 의미 C_i를 W의 의미로 결정하고, 그렇지 않으면 다음 단계로 넘어 간다.

(단계 3) W의 무순서 공기 정보를 적용하는 식 (5)를 이용하여 Ssim(S(W), UCW(W)) 값을 계산한다. 계산된 값이 임계치 T보다 크면 Ssim 값을 최대로 하는 의미 C_i를 W의 의미로 결정하고, 그렇지 않으면 다음 단계로 넘어 간다.

(단계 4) 온톨로지에서 제공하는 확장된 선택제약을 얼마나 잘 만족하는지를 검사한다. 만약, 여기에서도 계산된 값 S+이 미리 정의된 임계치 T_{onto}보다 높게 나오지 않았다면, 다음 단계로 넘어 간다.

(단계 5) 최후의 선택으로 최다빈도의 의미를 단어 W의 의미로 선정한다.

위의 (단계 4)를 적용하는 과정을 자세히 설명하면 다음과 같다. 온톨로지의 개념을 노드(node)로, 개념간 의미 관계를 링크(link)로, 상호정보는 온톨로지 개념간의 가중치(weight)로 보면, 온톨로지는 사이클(cycle)이 있는 가중치 그래프(weighted graph)로 간주할 수 있다. 그러나, 상호정보 값을 그대로 가중치로 사용하는 경우, 개념간의 의미 연관도가 높은지를 평가하기 위해서는 최대 가중치 경로를 찾는 알고리즘이 필요하게 되는데, 대상 그래프에 사이클이 존재하기 때문에 이러한 알고리즘은 작성할 수 없다. 그러므로, 상호정보 값은 식 (6)에 의해 페널티(penalty)값으로 바뀌게 되어, 개념간 기피도를 나타내는 수치로서의 역할을 하게 된다. 본 연구에서 사용한 상호정보 식 (1)은 계산된 결과가 1보다 큰지 적은지에 따라 서로 긍정적인 연관이 있는지 혹은 부정적인 연관이 있는지를 의미하는데, 이를 페널티로 바꾼 값의 의미도 이와 유사하게 값이 적을수록 상호 연관이 크고, 클수록 상호 연관이 적다는 것을 의미한다고 할 수 있다. *const*는 상호정보를 갖는 모든 개념쌍 중에서 최대 상호정보를 가리키는 상수이다.

$$Pe(\langle SC, SR\rangle, DC) = const - I(\langle SC, SR\rangle, DC) \quad (6)$$

개념간 기피도를 측정하기 위한 알고리즘은 Floyd-Warshall의 알고리즘[26]과 유사한데, 식 (7), 식 (8)과 같이 정의하였다.

$$S^*(C_i, C_j) = \begin{cases} 0 & if\ C_i = C_j, \\ \min_p (Pe(\langle C_i, R_p\rangle, C_j)) & \\ \quad if\ C_i \neq C_j \text{ and } C_i \xrightarrow{R_p} C_j, & \\ \min_{C_k \in \{C_i \rightarrow C_j\}} (S(C_i, C_k) * S(C_k, C_j)) & \\ \quad if\ C_i \neq C_j \text{ and } C_k \xrightarrow{R_p} C_j. & \end{cases} \quad (7)$$

$$S^+(C_i, C_j) = \begin{cases} 0 & if\ C_i = C_j, \\ \min_p (Pe(\langle C_i, R_p\rangle, C_j)) & \\ \quad if\ C_i \neq C_j \text{ and } C_i \xrightarrow{R_p} C_j, & \\ \min_{C_k \in \{C_i \rightarrow C_j\}} (S(C_i, C_k) + S(C_k, C_j)) & \\ \quad if\ C_i \neq C_j \text{ and } C_k \xrightarrow{R_p} C_j. & \end{cases} \quad (8)$$

C와 R은 각각 개념(concept)과 의미 관계(semantic relation)를 나타낸다. 만약 C_i와 C_j가 동일한 개념을 나타낸다면 페널티가 없고, 만약 C_i와 C_j가 동일한 개념이 아니면서 직접적인 의미 관계(또는, 선택제약 정보)가 있다면 C_i와 C_j 사이에 존재하는 모든 의미 관계 중 페널티 값이 최소인 것이 선택된다. 마지막으로 C_i와 C_j가 동일한 개념이 아니면서 직접적인 의미 관계가 없는 경우, 식 (7), 식 (8)은 최소 페널티를 갖는 경로를 찾아주게 되는데, 이는 두 개념간 존재하는 최적의 의미 연관도를 나타내게 된다. 이러한 특성은 은유나 환유 같은 표현을 해결하는데 도움을 줄 수 있다. 다시 말하면, 식 (7), 식 (8)은 두 개념간 선택제약이 얼마나 잘 만족되었는가를 측정할 수 있게 해주는 역할을 한다. 실험에 의해 위 두 식은 성능에 차이가 없음이 확인되었으므로 계산 속도를 고려하여 식 (8)을 이용하여 실험을 하였다.

6. 실 험

본 논문에서 제안한 단어 의미 중의성 해소 방법의 성능 평가를 위해서 8개의 명사와 4개의 동사를 선정하고 대상 단어가 나타나는 총 604개의 실험 문장을 선택하였다. 실험 문장은 원시 말뭉치에서 임의로 선정하였으며, 중의성을 갖는 단어의 여러 의미 중에서 가장 많이 사용되는 두, 세 가지의 의미만 고려하였다.

실험은 세 가지 형태로 이루어졌는데, 첫 번째 실험인 "BASE"는 최다빈도의 의미로만 단어의 의미를 선정한 경우인데, 이는 본 실험의 베이스라인(baseline), 즉 최소한 이 정도의 성능보다는 좋아야 한다는 가이드라인을 제시해 주는 역할을 한다. 두 번째 실험인 "LEX"은 용언의 결합가 정보, 지역 구문 패턴, 무순서 공기 단어 패턴과 같은 어휘정보가 포함

되어 있는 전자사전 정보만을 사용한 경우이다. 이것은 온톨로지를 사용하지 않은 일반적인 방법이라 할 수 있다. 세 번째 실험 "ONTO"는 본 연구에서 제안한 알고리즘 (그림 3)에 따라 온톨로지를 단어 의미 중의성 해소에 활용한 경우이다.

기계번역 시스템에서의 WSD 실험 결과는 <표 3>에 나타나 있다. 결과적으로 "ONTO" 실험은 "LEX" 실험보다 평균 정확률 9%의 성능 향상을 보였다.

〈표 3〉 한국어에서의 WSD 실험 결과(%)

품사	단어	의 미	BASE	LEX	ONTO
명사	부자	father & child / rich man	65.3	69.2	86.0
	간장	liver / soy sauce	66.0	87.8	91.8
	가사	housework / words of song	48.0	88.5	96.1
	구두	shoe / word of mouth	78.0	85.7	95.9
	눈	eye / snow	82.0	96.0	94.0
	용기	courage / container	62.0	74.0	82.0
	경비	expenses / defense	74.5	78.4	90.2
	경기	times / match	52.9	80.4	93.2
동사	내리다	get off / draw	42.0	72.0	88.0
	세우다	make(a plan) / build	54.0	88.0	95.4
	쓰다	use / write / put on(a hat)	46.0	86.0	96.0
	태우다	burn / give a ride	50.0	86.0	92.0
평균 정확률			60.1	82.7	91.7

<표 3>에 제시된 실험결과 중 "눈"에 대한 결과를 보면 LEX의 결과가 ONTO의 결과보다 좋은 것을 볼 수 있다. 이는 표제어 "눈"이 가지는 개념에 대한 의미 관계 패턴의 정보가 온톨로지 학습과정에서 제대로 추출되지 않았기 때문으로 볼 수 있다.

제시한 WSD 알고리즘의 각 단계별 적용률과 정확률은 <표 4>와 같다. 온톨로지는 WSD 작업에서 약 18% 정도의 역할을 담당하고 있는데, 구체적인 어휘 정보를 적용하는 전 단계에 비해 정확률이 낮게 나오는 것은 당연한 결과이며, 최다빈도 의미 적용단계보다 정확률이 높게 나온 부분이 본 연구를 통해 얻은 성과라 분석해 볼 수 있다.

〈표 4〉 WSD 알고리즘의 각 단계별 실험결과 분석

적 용 단 계	적용률(%)	정확률(%)
용언의 결합가 정보	34.8	91.6
지역 구문 패턴	9.8	91.4
무순서 공기 단어 패턴	28.2	92.3
온톨로지 추론	18.1	86.4
최다빈도 의미	9.1	74.2
합계 / 적용률을 고려한 평균 정확률	100	89.2

본 연구에서의 실험결과를 객관적으로 평가하기 위해서는 다른 기존 연구와의 비교가 필요하나, 실험대상 단어 및 후보 의미의 수, 실험 문장 등 동일한 조건 하에서의 실험이 사실상 불가능하기 때문에 실험결과 수치의 직접적인 비교는 의미가 없다고 볼 수 있다. 하지만, 대략적인 비교를 위해 동일한 실험단어를 대상으로 정리한 내용을 <표 5>에 제시하였다. 제시된 기존 방법 중 후보 의미가 많은 것은 다의어의 중의성 해소를 시도한 것으로 볼 수 있다. 그러나, 후보 의미의 수에 비례하여 문제도 어려워지는 것은 아니다. 왜냐하면, 자주 사용되지 않는 의미는 특정한 어휘와 같이 사용되거나 특정 문맥에서만 나타나는 경우가 많기 때문에 자주 사용되는 의미에 비하여 의미를 결정하기가 오히려 쉬워지는 경향이 있기 때문이다. 이러한 관점에서 자주 사용되는 서너 가지의 의미간 중의성 해소가 오히려 더 중요하고 어렵다고 할 수 있다.

〈표 5〉 기존 WSD 연구와의 실험결과 비교

기존 연구	실험단어	후보 의미수		정확률(%)	
		기존연구	ONTO	기존연구	ONTO
이승우[5]	가사	5	2	87.7	96.1
	경기	4	2	78.3	93.2
조정미[6]	쓰다	23	3	87.3	96.0
허정[7]	눈	3	2	80.0	94.0

7. 결 론

본 논문에서는 시소러스, 기계번역사전, 세종전자사전, 말뭉치 등 기존 언어자원들을 최대한 활용하여 온톨로지를 구축하고, 온톨로지의 학습을 통하여 추론이 가능케 함으로써 자연어 처리에서 중요한 문제 가운데 하나인 단어 의미 중의성 문제를 해결하고자 하였다.

온톨로지는 실세계에 존재하는 모든 개념들(concepts)과 그 개념들의 속성들(properties), 그리고 개념들이 상호간 의미적으로 어떻게 연결되어 있는가(semantic relations)에 대한 정보를 가지고 있는 지식베이스(knowledge base)로 여러 응용 분야에서 활용될 수 있으나, 특히 언어를 전산처리하기 위하여 의미 분석하고자 할 때 반드시 필요한 자원이라 할 수 있다.

이러한 온톨로지를 실용적으로 구축하기 위해서 가도카와 시소러스의 개념 체계에 격 관계와 기타 의미관계와 같은 다른 의미관계를 추가하여 확장하는 방법을 사용하였으며, 계층 관계 외 다른 다양한 의미관계를 얻기 위해서, 기계번역사전과 세종전자사전에 각각 포함되어 있는 결합가 정보와 격틀정보를 활용하였고 또한 의미 태깅된 말뭉치를 분석하여 얻은 개념공기정보도 사용하였다.

온톨로지의 추론 과정은 온톨로지에 존재하는 개념간 연관도를 측정하는 것으로 볼 수 있는데 상호정보(mutual in-

formation)를 이용하여 그 정도를 측정하였다. 온톨로지의 개념을 노드(node)로, 개념간 의미 관계를 링크(link)로, 상호 정보는 온톨로지 개념간의 가중치(weight)로 보면, 온톨로지는 사이클(cycle)이 있는 가중치 그래프(weighted graph)가 된다. 온톨로지 그래프에서 최소 비용 경로를 찾는 형태로 단어 의미 중의성 해소에서 활용되게 되는데, 이를 통하여 중의성이 있는 단어의 후보 개념간 선택 제약이 얼마나 잘 만족되는지를 평가하게 된다. 기계번역 시스템(COBALT-K/J)에서 실험한 결과, 온톨로지를 사용하지 않았을 때보다 한국어 분석에서 9%의 평균 정확률 향상을 얻을 수 있었다.

향후 연구로는 온톨로지에 추가될 의미 관계를 보다 쉽고 정확하게 추출하는 방법과 온톨로지 개념간 선택 제약 정도를 계산하는 식 (8)의 성능 개선 및 온톨로지와 시맨틱 웹(semantic web)[25]의 상호 활용 가능성에 관한 연구를 하고자 한다.

참 고 문 헌

[1] 21세기 세종계획 전자사전 개발분과, '2000년도 연구보고서', 문화관광부, 2000.

[2] 강신재, 박정혜, "대규모 말뭉치와 전산 언어 사전을 이용한 의미역 결정 규칙의 구축", 정보처리학회논문지B, 제10-B권 제2호, pp.219-228, 2003.

[3] 김영택 외 공저, '자연언어처리', 생능출판사, 2001.

[4] 서희철, 이 호, 백대호, 임해창, "유사어를 이용한 단어 의미 중의성 해결", 제11회 한글 및 한국어 정보처리 학술대회, pp.304-309, 1999.

[5] 이승우, 이근배, "국소 문맥과 공기 정보를 이용한 비교사 학습 방식의 명사 의미 중의성 해소", 정보과학회논문지, 제27권 제7호, pp.769-783, 2000.

[6] 조정미, 코퍼스와 사전을 이용한 동사 의미 분별, 한국과학기술원 전산학과 박사학위논문, 1998.

[7] 허정, 옥철영, "사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 정보과학회논문지, 제28권 제9호, pp.688-698, 2001.

[8] C. Leacock and M. Chodorow, "Using Corpus Statistics and WordNet Relations for Sense Identification," Computational Linguistics, Vol.24, No.1, pp.147-165, 1998.

[9] D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt and M. Shepherd, "Cyc : toward programs with common sense," Communications of the ACM, Vol.33, No.8, pp.30-49, 1999.

[10] D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics(ACL'95), Cambridge, MA, pp.189-196, 1995.

[11] D. Yarowsky, "Word sense disambiguation using statistical models of Roget's categories trained on large corpora," The 14th International Conference on Computational Linguistics, Nantes, France, pp.454-460, 1992.

[12] E. Agirre and G. Rigau, "Word-Sense Disambiguation Using Conceptual Density," In Proceedings of the 16th International Conference on Computational Linguistics, Somerset, NJ, Association for Computational Linguistics, 1996.

[13] G. A. Miller, M. Chodorow, S. Landes, C. Leacock and R. G. Thomas, "WordNet : An On-line Lexical Database," International Journal of Lexicography, Vol.3, No.4, pp.235-244, 1990.

[14] H. F. Li, N. W. Heo, K. H. Moon, J. H. Lee and G. B. Lee, "Lexical Transfer Ambiguity Resolution Using Automatically-Extracted Concept Co-occurrence Information," International Journal of Computer Processing of Oriental Languages, World Scientific Pub., Vol.13, No.1, pp.53-68, 2000.

[15] Japan Electronic Dictionary Research Institute, LTD., 'EDR Electronic Dictionary Version 1.5 Technical Guide,' 1995.

[16] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, pp.76-83, 1989.

[17] K. H. Moon and J. H. Lee, "Representation and Recognition Method for Multi-Word Translation Units in Korean-to-Japanese MT System," In the 18th International Conference on Computational Linguistics, Germany, pp.544-550, 2000.

[18] K. Knight and S. K. Luk, "Building a Large Knowledge Base for Machine Translation," Proceedings of the American Association of Artificial Intelligence Conference AAAI-94, Seattle, WA, 1994.

[19] K. Mahesh and S. Nirenburg, "Knowledge-based systems for Natural Language Processing," Memoranda in Computer and Cognitive Science. NMSU CRL Technical Report, MCCS-96-296, 1996.

[20] N. Ide and J. Veronis, "Introduction to the special issue on word sense disambiguation : the state of the art," Computational Linguistics, Vol.24, No.1, pp.1-40, 1998.

[21] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," In Proceedings of IJCAI-95, Montreal, Canada, pp.448-453, 1995.

[22] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama and Y. Hayashi, 'Goi-Taikei : A Japanese Lexicon,' Iwanami Shoten, Tokyo, 5 volumes/CDROM, 1997.

[23] S. Nirenburg, J. Carbonell, M. Tomita, and K. Goodman, 'Machine Translation : A Knowledge-Based Approach,' Morgan Kaufmann Pub., San Mateo, California, 1992.

[24] S. Ohno and M. Hamanishi, 'New Synonyms Dictionary,'

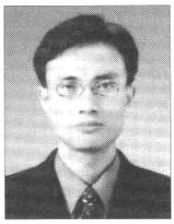
Kadokawa Shoten, Tokyo, 1981.

[25] T. Berners-Lee, J. Hendler and O. Lasilla, 'The Semantic Web,' Scientific American, May, 2001.

[26] T. H. Cormen, C. E. Leiserson and R. L. Rivest, 'Introduction to Algorithm,' McGraw-Hill Book Co., 1990.

[27] X. Li, S. Szpakowicz and S. Matwin, S., "A WordNet -based algorithm for word sense disambiguation," in IJCAI'95, pp.1368-1374, 1995.

[28] Z. Dong and Q. Dong, HowNet. http://www.keenage.com/zhiwang/e_zhiwang.html, 1999.



강 신 재

e-mail : sjkang@daegu.ac.kr

1995년 경북대학교 컴퓨터공학과(공학사)

1997년 포항공과대학교 대학원 컴퓨터공학과 (공학석사)

2002년 포항공과대학교 대학원 컴퓨터공학과 (공학박사)

1997년~1998년 SK Telecom 정보기술연구원 주임연구원

2002년~현재 대구대학교 정보통신공학부 조교수

관심분야 : 온톨로지, 시맨틱 웹, 자연어처리, 정보검색, 기계학습 등