의미 분석을 위한 말뭉치 기반의 온톨로지 학습†

(Corpus-Based Ontology Learning for Semantic Analysis)

강 신 재*
(Sin-Jae Kang)

요 약 본 논문은 한국어정보처리에서 단어의 의미를 결정하기 위한 말뭉치 기반의 온톨로지 학습 방법을 제시하고 있다. 먼저 이미 확보된 전자사전의 정보를 이용하여 단어의 확실한 의미를 우선 결정한 후, 아직 결정하지 못한 단어의 의미는 온톨로지를 이용하여 최종 결정하는 절차를 거친다. 온톨로지를 단어 의미 중의성 해소를 위한 지식베이스로 사용하기 위해서는, 온톨로지 내 개념들간의 상호정보가 말뭉치의 통계 정보에 근거하여 미리 계산된다. 계산된 상호정보 값을 가중치로 간주하면 온톨로지는 가중치 그래프로 생각할 수 있으므로, 개념간 최소 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.

핵심주제어 : 온톨로지, 말뭉치분석, 단어 의미 중의성 해소, 기계번역

Abstract This paper proposes to determine word senses in Korean language processing by corpus-based ontology learning. Our approach is a hybrid method. First, 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 least 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.

Key Words : Ontology, Corpus Analysis, Word Sense Disambiguation, Machine Translation

1. 서 론

한국어정보처리에서 형태소분석/구문분석 이상의 처리를 하기 위해서는 온톨로지(ontology)라 불리는 의미지식베이스가 반드시 필요하다.

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 색인표목의 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[1]. 자연어처리에서의 시소러스는 상하위 개념어, 동형이의어, 다의어, 반의어, 부분-

† 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.
* 대구대학교 컴퓨터·IT공학부 교수

집합어, 관련어 등으로 구성되며, 대표적인 시소러스로는 WordNet[2], Goi-Taikei[3] 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[4, 5]. 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 시소러스(thesaurus)와 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. 여기에 대해서는 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, 신경망을 이용한 방법으로 분류할 수 있다[6]. 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 대체로 좋은 결과를 보이고 있다.

본 연구에서는 온톨로지를 의미 분석 단계에서의 핵심 작업인 단어 의미 중의성 해소에 사용하여 기존 성능을 향상시키고자 한다.

2. 온톨로지 구축

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

첫째, 가도카와 시소러스[7]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 이용 시 별도의 사전 작업없이 온톨로지의 활용 및 평가가 가능하다. 이러한 접근법은 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[8].

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[9]과 마이크로 코스모스 온톨로지[4]를 주

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 기계번역 시스템이다.

로 참고하여 총 30개의 의미 관계를 정의하였다<표 1>.

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

계층 관계 외 다른 다양한 의미 관계를 얻기 위해서는 두 가지 접근방법을 사용하였는데, 기존 전자사전에 포함되어 있는 의미 정보의 활용과 말뭉치의 반자동 분석[8]이 그것이다. 개념간의 격 관계는 주로 세종 전자사전과 COBALT-J/K, COBALT-K/J와 같은 기계번역사전에 포함되어 있는 의미 정보를 변환하여 얻을 수 있으며, 의미 태깅된 말뭉치 분석의 결과인 개념 공기정보(concept co-occurrence information)를 통하여 온톨로지에 추가될 의미 관계를 추출할 수 있다.

<표 1> 온톨로지에 포함된 의미관계 유형

대범주	소범주
계층 관계	is-a
격 관계	agent, theme, experiencer, companion, instrument, location, source, destination, reason, appraisee, criterion, degree, recipie
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

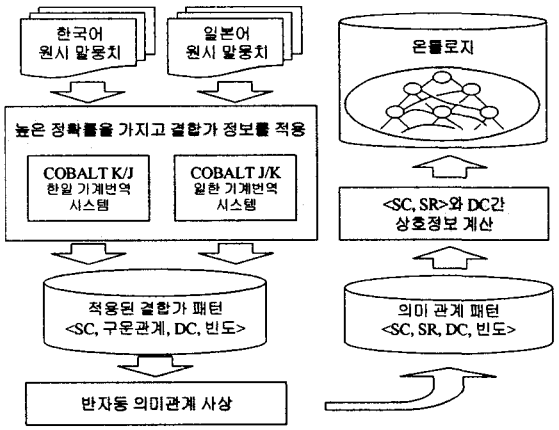
3. 온톨로지 학습

온톨로지를 이용해 추론하기 위해서는 지배소(governor)와 의존소(dependent) 개념간 온톨로지 내에서의 의미 제약을 얼마나 잘 만족하는가를 평가할 수 있는 방법이 필요하다. 본 연구에서는 개념간 연관도를 측정하기 위해 상호정보(mutual information)를 사용하고자 한다. 상호정보는 두 랜덤변수 간 의존도를 측정하는 방법으로 Church & Hanks[10]가 제안하였다. Resnik[11]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.

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

만약, <SC, SR>과 DC간에 깊은 연관이 있다면, 확률 P(<SC, SR>, DC)는 확률의 곱 P(<SC, SR>)P(DC)보다 클 것이므로, I(<SC, SR>, DC) >> 1이 된다. 만약, <SC, SR>과 DC가 특별한 관계가 없다면, P(<SC, SR>, DC) ≈ P(<SC, SR>)P(DC)이 되므로, I(<SC, SR>, DC) ≈ 1이 된다.

온톨로지를 단어 의미 중의성 해소에 사용하기 위해서는 온톨로지에 존재하는 모든 개념간 상호정보 값을 미리 확보하고 있어야 한다. <그림 1>은 <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, '94-97) 한국어 원시 말뭉치와 81만 문장의 일본어 원시 말뭉치를 분석하여 의미 태깅된 결합가 정보 패턴을 추출



<그림 1> 온톨로지 학습 데이터의 생성

하였다. 위의 기계번역 시스템에서는 번역결과를 얻기 위해 단어들의 의미 중의성 해소 과정을 거치게 되는데[8], 이때 내부적으로 적용된 결합가 정보와 가도카와 의미코드를 출력하는 방법으로 의미가 태깅된 결합가 패턴을 얻을 수 있게 된다. 추출된 결합가 정보 패턴 중 구문관계 정보를 규칙과 사람의 직관에 의해, 온톨로지에 정의된 의미 관계로 변환하면, 빈도수를 가진 <SC, SR, DC> 패턴을 얻을 수 있다. 이 결과를 가지고 온톨로지 내 개념간 상호정보를 계산하게 된다.

또, 이 과정에서 얻어진 의미 관계 패턴들 역시 온톨로지에 추가되는데, 본 장에서 생성한 의미 관계 패턴은 2장에서 생성하였던 의미 관계 패턴과 약 20.7% 정도의 중복된 결과를 보였으나, 이 가운데에서 중복된 패턴은 한번만 온톨로지에 추가되었다. 2장에서 추출된 의미 관계 패턴들은 가도카와 시소러스의 계층 구조에서 주로 L10, L100 단계에 추가된 것이고, 본 학습 과정에서 생성된 의미 관계 패턴들은 L1000단계에 주로 추가되게 된다. L1000 단계에 추가된 의미 관계 패턴들은 L10, L100 단계에 추가된 패턴들보다 더 세밀하고 실제적인 선택제약을 나타낸다고 할 수 있기 때문에, 개념 연관도 측정 시 더 유용한 정보를 제공하게 된다. <표 2>는 온톨로지에 추가된 의미 관계 패턴의 최종적인 수를 보여주고 있다.

<표 2> 온톨로지에 추가된 의미관계 패턴의 수

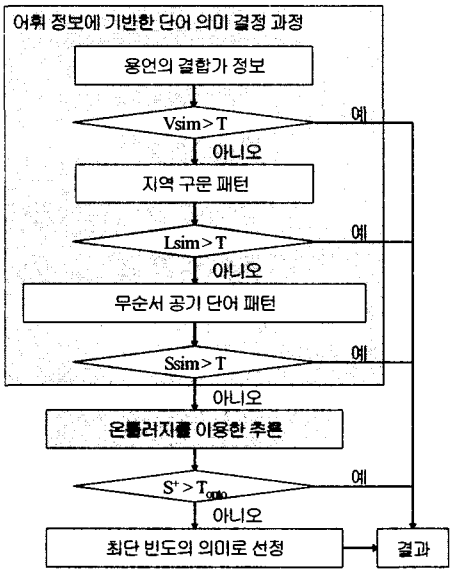
의미관계 유형	패턴 수
계층 관계	1,100
격 관계	112,746
기타 의미관계	2,093
계	115,939

온톨로지 내 각 개념(노드)에 연결된 의미 관계 인스턴스(instance)의 평균수, 즉 링크의 평균수(degree)는 103.4개이며, 의미 코드 "379(possession)"는 1,398개의 인스턴스가 연결되어서 최대치를 기록했다. 또, 각 개념에 연결된 의미 관계 유형(type)의 평균수는 7.4개이며, 의미 코드 "713(group)"은 20개의 유형이 연결되어서 최대치를 기록했다. 이와 같은 통계 자료를 분석해 볼 때, 본 연구에서 구축한 온톨로지는 상당한 양의 의미 정보(의미 관계)를 포함하고 있음을 알 수 있다. 한편, 개념에 연결된 링크의 수가 많다는 것은 해당 개념이 내포하고 있는 의미 범주가 크기 때문에 말뭉치에서 그 사용이 빈번했다는 것을 의미한다. 즉, 해당 개념 부류가 다의성을 가지고 있다는 뜻도 된다. 그러므로, 분석된 통계 자료는 추후 온톨로지 개념 계층 구조의 확장 시, 분할 대상 개념의 선정을 위한 힌트로 사용될 수 있다. <그림 2>는 구축된 온톨로지의 일부분을 보여주고 있다.

<그림 2> 구축된 온톨로지의 일부분

4. 단어 의미 중의성 해소

<그림 3>은 온톨로지를 사용한 전체적인 단어 의미 중의성 해소 방법을 보이고 있다. 먼저, 의미 중의성이



<그림 3> 제안하는 단어 의미 중의성 해소 방법

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 이용하여 선택제약을 얼마나 잘 만족하는지를 검사한다. 만약, 여기에서도 선택제약의 만족도가 높게 나오지 않았다면, 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), (3), (4), (5)에 의해 구현된다. 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. Ci와 Pj는 개념 유형을 표현하고, Sk는 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 Sk의 j번째 구문관계 패턴의 개념코드 개수이며, 'r'은 의미 Sk의 무순서 공기 단어 패턴의 개념코드 개수이다.

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

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

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

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

식(2)의 Csim(Ci, Pj)는 가도카와 시소러스에 기반하여 개념 Ci와 Pj 사이의 유사도를 계산하는 식이다. 식(2)에서 weight는 개념의 가중치를 의미하며, 유사도 계산시 개념 Ci의 부모(parent) 개념이 형제(sibling) 개념보다 유사한 특징을 더 많이 가지고 있으며, 이러한 관계를 중요시 한다는 것을 뜻한다. 즉, 개념 Ci가 Pj의 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.

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 절대로 작성할 수 없다. 그러므로, 상호정보 값은 식(6)에 의해 페널티(penalty)값으로 바뀌게 되어, 개념간 기피도를 나타내는 수치로서의 역할을 하게 된다.

$$Pe(<SC, SR>, DC) = const - I(<SC, SR>, DC) \quad (6)$$

본 연구에서 사용한 상호정보 수식(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.

개념간 기피도를 측정하기 위한 알고리즘은 Floyd-Warshall의 알고리즘[12]과 유사한데, 식(7)과 같이 정의하였다.

$$S^+(C_i, C_j) = \begin{cases} 0 & if\ C_i = C_j, \\ \min_p \left(Pe(<C_i, R_p>, C_j)\right) & \\ \quad if\ C_i \neq C_j\ and\ C_i \xrightarrow{R_p} C_j, & \\ \min_{C_k \in \{C_i \to C_j\}} \left(S(C_i, C_k) + S(C_k, C_j)\right) & \\ \quad if\ C_i \neq C_j\ and\ C_k \xrightarrow{R_p} C_j. & \end{cases} \quad (7)$$

C와 R은 각각 개념(concept)과 의미 관계(semantic

relation)를 나타낸다. 만약 Ci와 Cj가 동일한 개념을 나타낸다면 페널티가 없고, 만약 Ci와 Cj가 동일한 개념이 아니면서 직접적인 의미 관계(또는, 선택제약 정보)가 있다면 Ci와 Cj 사이에 존재하는 모든 의미 관계 중 페널티 값이 최소인 것이 선택된다. 마지막으로 Ci와 Cj가 동일한 개념이 아니면서 직접적인 의미 관계가 없는 경우, 식(7)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)은 두 개념간 선택제약이 얼마나 잘 만족되었는가를 측정할 수 있게 해주는 역할을 한다.

5. 실 험

본 논문에서 제안한 온톨로지의 학습을 통한 추론의 단어 의미 중의성 해소에서의 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 포함되어 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>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.

<표 3> 한국어에서의 WSD 실험 결과(%)

품사	단어	의미	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

6. 결 론

본 논문에서는 말뭉치를 이용한 온톨로지의 학습을 통하여 온톨로지의 추론을 가능케 함으로써 자연어 처리에서 중요한 문제 가운데 한 가지인 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. 하지만 이전의 연구들은 온톨로지를 어떻게 구축할 것인가에만 초점이 맞추어져 있는 반면, 그 활용에 관한 연구는 미진하였으므로 본 연구에서는 온톨로지의 학습을 통하여 추론에 활용하였다.

추론 과정은 온톨로지에 존재하는 개념간 연관도를 측정하는 것으로 볼 수 있는데 상호정보(mutual information)

를 이용하여 그 정도를 측정하였으며, 이는 말뭉치의 분석을 통해 얻은 통계정보로부터 구할 수 있다. 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 간주된다. 그래프에서 최소 비용 경로를 찾는 형태로 단어 의미 중의성 해소에서 활용되게 되는데, 이를 통하여 중의성이 있는 단어의 후보 개념간 선택 제약이 얼마나 잘 만족되는가를 평가하게 된다. 실용 기계번역 시스템(COBALT-K/J)에서 단어의미 중의성 해소 실험을 한 결과, 온톨로지를 사용하지 않았을 때보다 한국어 분석에서 9%의 평균 정확률 향상을 얻을 수 있었다.

향후 연구로는 보다 정확한 의미 관계의 추출과 식(7)의 성능을 개선할 수 있는 추론식의 개발 및 온톨로지와 시맨틱 웹(semantic web)[13]의 상호 활용 가능성에 관한 연구를 하고자 한다.

참 고 문 헌

[1] 김영택 외 공저. 2001. *자연언어처리*. 생능출판사.

[2] Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. 1990. WordNet: An On-line Lexical Database, *International Journal of Lexicography*. 3(4), pp. 235-24.

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

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

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

[6] Ide, N. and Veronis, J. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, vol.24, no.1, pp.1-40.

[7] Ohno, S. and Hamanishi, M. 1981. *New Synonyms Dictionary*, Kadokawa Shoten, Tokyo.

[8] Li, H. F., Heo, N. W., Moon, K. H., Lee, J. H., and Lee, G. B. 2000. Lexical Transfer Ambiguity Resolution Using Automatically-Extracted Concept Co-occurrence Information, *International Journal of Computer Processing of Oriental Languages*, World Scientific Pub., 13(1):53-68.

[9] 21세기 세종계획 전자사전 개발분과, 2000년도 연구보고서, 문화관광부.

[10] Church, K. and Hanks, P. 1989. 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.

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

[12] Cormen, T. H., Leiserson, C. E., and Rivest, R. L. 1990. *Introduction to Algorithm*. McGraw-Hill Book Co.

[13] Berners-Lee, T., Hendler, J., and Lasilla, O. 2001. The Semantic Web. *Scientific American*, May.

강 신 재 (Sin-Jae Kang)

1995년 경북대학교 컴퓨터공학과 졸업(학사)

1997년 포항공과대학교 컴퓨터공학과 졸업(공학석사)

2002년 포항공과대학교 컴퓨터공학과 졸업(공학박사)

1997년~1998년 SK Telecom 정보기술연구원 주임연구원

2002년~현재 대구대학교 컴퓨터·IT공학부 조교수

(관심분야 : 문서분류, 정보검색, 기계학습, 기계번역)