말뭉치와 개념정보를 이용한 명사 중의성 해소 방법

Noun Sense Disambiguation Based-on Corpus and Conceptual Information

이 휘 봉 허 남 원 문 경 희 이 종 혁
(Hui-Feng Li) (Nam-Won Heo) (Kyong-Hi Moon) (Jong-Hyeok Lee)

요 약 본 논문에서는 말뭉치와 개념정보에 기반한 명사 중의성 해소 방법을 제안한다. 기존의 연구에서는 대부분 어휘의 공기정보를 이용하고 있으나, 이러한 방법은 많은 저장공간이 필요하고, 적용률이 크지 않다는 단점이 있다. 본 논문에서는 자동으로 의미 태깅된 한국어 말뭉치에서 추출된 공기 개념정보를 이용하여 명사 중의성을 해소하는 방법을 제안한다. 제안한 방법의 평가 실험에서 기본의미를 정하는 것보다 14.6% 높은 평균 82.4%의 정확률을 보였다. 실험 문장들이 학습문장과 다른 것을 고려하면, 제안된 방법이 어휘 중의성 해소에 유용함을 보여주는 결과라고 할 수 있다.

주제어 통계정보, 개념정보, 명사 중의성 해소

Abstract This paper proposes a noun sense disambiguation method based-on corpus and conceptual information. Previous research has restricted the use of linguistic knowledge to the lexical level. Since knowledge extracted from corpus is stored in words themselves, the methods requires a large amount of space for the knowledge with low recall rate. On the contrary, we resolve noun sense ambiguity by using concept co-occurrence information extracted from an automatically sense-tagged corpus. In one experimental evaluation it achieved, on average, a precision of 82.4%, which is an improvement of the baseline by 14.6%. Considering that the test corpus is completely irrelevant to the learning corpus, this is a promising result.

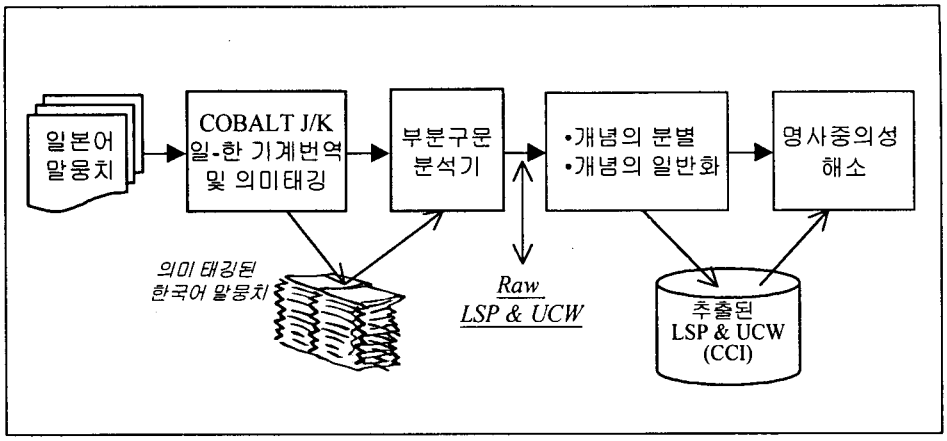
Keywords Statistical information, Conceptual information, Noun sense disambiguation.

1. 서론

본 논문에서는 자연언어 처리에서 반드시 해결해야 할 문제인 한국어 명사 중의성 해소 방법을 제안한다. 한국어는 표음어임으로 하나의 단어가 여러 개의 의미에 대응되는 경우가 자주 발생하며, 특히 명사는 선택제약 (selectional restriction) 등 이용할 수 있는 어휘지식이 미약하기 때문에 중의성 해소에 많은 어려움이 따른다.

어휘중의성 해소를 위한 기존 연구들은 그 이용하는 지식원에 따라 크게 규칙 기반 방법(Rule-based approach), 기계 가독형 사전(Machine Readable Dictionary, MRD)을 이용하는 방법(MRD-based approach), 말뭉치를 이용하는 방법(Corpus-based approach)으로 분류할 수 있다. 규칙 기반 방법은 단어 의미 중의성 해결에 필요한 지식을 수작업을 통해서 구축하고 이를 이용하여 단어 의미 중의성을 해결하는 방법이다[5, 6]. 그러나 이러한 규칙 기반 방법들은 수작업을 통한 규칙 생성의 노력 때문에 어휘 중의성 해소를 위한 지식의 획득에 어려움을 겪게 된다. 기계 가독형 사전을 이용하는 방법은 어휘 중의성을 해결하고자 하는 단어의 사전에서의 정의나 기술, 주제 코드를 이용하는 방법이다[3]. 그러나 이러한 방법들은 사전의 정의나 기술에서 사용되는 어휘가 제한적이기 때문에 무제한의 어휘를 갖는 실제 문장에 적용하는 데에는 한계를 갖는다. 그리고 이 방법이 높은 정확성을 갖기 위해서는 적용되는 도메인에 따라서 잘 정의된 사전의 구축이 필요하다는 문제점이 있다.

* 포항공과대학교 컴퓨터공학과
Dept. of Computer Science and Engineering
Pohang University of Science and Technology
E-mail: {hflee,nwheo,khmoon,jhlee}@kle.postech.ac.kr



(그림1) 어휘중의성 해소 시스템 구성도

말뭉치를 이용하는 방법은 크게 사용하는 말뭉치의 성질에 따라서 비교사 학습(Unsupervised learning) 방법과 교사 학습(Supervised learning) 방법으로 구분할 수 있다. Yarowsky가 제안한 방법[11]은 의미 태깅(tagging)되지 않은 영어 원시 말뭉치로부터 각 담화에서 단어는 하나의 의미를 갖는다는 성질과 각 연어(Collocation)에서 단어는 하나의 의미를 갖는다는 제약 조건을 이용하여 비교사 학습에 의해 어휘중의성 해소 방법을 제안하였다. 이 방법은 연어 단어의 가중치에 의해서 의미를 판별하는 방법으로 다른 말뭉치 기반의 어휘중의성 해소 방법들과 같은 자료 부족 문제를 겪게 된다. 또한 추출된 지식을 개념이 아닌 실제 어휘로 저장하기 때문에 지식의 저장 공간이 크고 적용률이 저하된다는 문제점이 있다. [1]에서 제안한 방법은 국소 문맥과 공기정보를 이용한 명사중의성 해소 비교사 학습인데, 의미 태깅되지 않은 말뭉치를 이용하기 때문에 공기하는 단어들의 의미 분별력이 약하고, 또한 추출된 공기 정보를 실제 어휘로 저장하기 때문에 적용률이 저하된다는 문제점을 안고 있다.

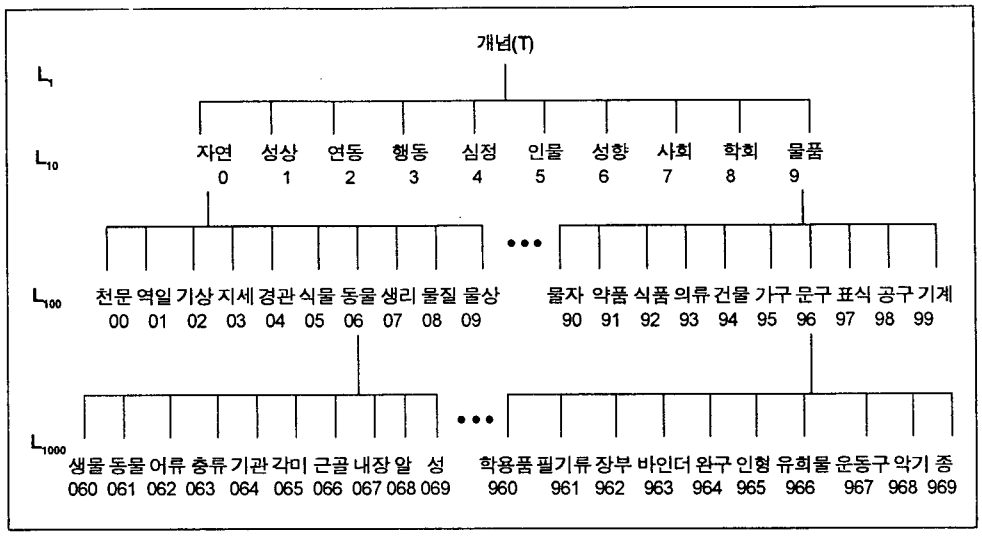
교사 학습에 의한 방법은 의미태깅된 말뭉치를 이용하여 어휘중의성 해소에 필요한 지식을 획득하는 방법이다. Hwee가 제기한 방법[4]은 영어를 대상으로 의미 태깅된 말뭉치에서 주위 문맥 단어, 형태소 정보, 주위 단어의 품사 정보, 공기 정보, 동-목적어 구문 관계 정보 등 다양한 정보들을 추출하여 이용하였다. 따라서 Hwee의 방법은 기존의 말뭉치 기반 방법들에 비해서 상대적으로 높은 정확률을 나타내었다. 그러나 이 방법은 수작업으로 태깅된 말뭉치를 사용하고 있기 때문에 많은 인력이 필요하고, 추출된 지식들은 개념이 아닌 실제 어휘로 저장하기 때문에 저장공간이 크고 적용률 향상이 어렵다.

본 논문에서는 이러한 기존 방법들의 문제점을 고려하여 의미 구분된 말뭉치에서 공기 개념정보(Co-occuring Conceptual Information, CCI)를 추출하여 이용한 명사 중의성 해소 방법을 제안한다. 개념정보의 추출과정은 (그림1)에서 나타낸 것과 같다. 2장에서는 공기 개념정보의 자동 추출에 관해서 논의하고, 3장에서는 개념지식을 이용한 명사중의성 해소 알고리즘을 소개하고, 4장에서는 한-일 번역에서의 명사중의성 해소 실험을 통하여 제안한 방법의 효율성을 검증하고, 5장에서는 결론을 맺는다.

2. 공기 개념정보의 추출

2.1. 의미 구분된 한국어 말뭉치의 생성

1단계에서는 먼저 일-한 번역기 COBALT-J/K[8]를 이용하여 학습 자료로 사용될 의미 구분된 한국어 말뭉치를 생성한다. 한국어와 일본어는 언어 계통상 알타이 어족에 속하는 동족 언어로 동일 한자 문화권에 속하며 문법 체계가 비슷하고 언어 유형론적인 측면에서도 많은 유사성을 가지고 있다. 특히 양 언어의 어순이 거의 일치하고 있다. COBALT-J/K는 이러한 언어적 유사성을 최대한 이용한 직접 번역 방식만을 채택하더라도 고품질의 번역 성능을 보여주고 있다. 따라서, 본 논문에서는 의미 태깅된 말뭉치를 생성하기 위하여 COBALT-J/K가 번역을 수행할 때 내부적으로 다의성을 해소하기 위해서 사용하는 가도



(그림2) 가도까와 시소러스 개념 계층 구조

까와 시소러스[7]의 코드들을 번역된 한국어 어휘 뒤에 태깅하여 번역문을 생성하도록 하였다. 가도까와 시소러스는 (그림2)에서 보여주듯이 4계층 구조를 가지고 있고, L_1계층부터 L_{100}계층까지의 각 계층의 개념은 10개의 하위 개념으로 분류되고, 각각의 개념은 중복되지 않는 숫자로 코딩된다. 한국어 어휘들은 4번째 계층인 L_{1000}계층의 개념코드로 의미 태깅된다.

2.2. 구문관계패턴과 문맥정보의 추출

중의성 명사와 의미적 제약을 가지는 공기정보들은 일정한 구문관계를 가지고 한 문장 속에 출현하기도 하고 특정 구문관계를 갖지 않지만 의미적으로 제약을 주는 것도 있다. 따라서, 본 논문에서는 공기 개념정보(Conceptual Co-occurrence Information, CCI)로서 개념으로 표현된 구문관계패턴(Local Syntactic Pattern, LSP)과 일정한 구문관계를 갖지 않지만 의미제약을 주는 문맥정보(Unordered Co-occuring Words, UCW)를 말뭉치로부터 추출하여 사용하였다. 구문관계패턴은 한국어 말뭉치에서 조사의 분포를 고려하여 명사가 다른 명사나 용언과 수식 또는 피수식의 관계 등으로 사용되는 표현들을 어순, 구문 관계, 품사, 형태적 특성을 반영하여 중요도에 따라 구분 정리한 상위 10개의 구문관계패턴(Local Syntactic Pattern, LSP)들이며, 이를 〈표1〉

〈표1〉 구문관계패턴 (LSPs)

패턴	패턴구조	예문
$type_1$	**noun** + noun	**부자** (father and child) 관계
$type_2$	noun + 의 + **noun**	노래의 **가사** (words)
$type_3$	**noun** + 기타조사 + noun	**눈** (eye)-과 귀
$type_4$	**noun** + 로/으로 + verb	**구두** (verbally)-로 설명하다
$type_5$	**noun** + 에 + verb	**배** (ship)-에 타다
$type_6$	**noun** + 에게 + verb	**철수** (human name)-에게 주다
$type_7$	**noun** + 에서 + verb	**지도** (map)-에서 찾다
$type_8$	**noun** + 을/를 + verb	**배** (pear)-를 먹다
$type_9$	**noun** + 이/가 + verb	**배** (ship)-가 떠나다
$type_{10}$	verb + 관계화소 + **noun**	흰 **눈** (snow)

에 나타내었다. 이들 수식 관계 쌍들은 의미적으로 빈번히 공기하면서 서로 강한 의미적인 제약을 가진다. 또한 문장에서 구문의존관계와 상관없이 중의성 단어와 자주 같이 나타나는 단어들을 문맥정보로 정의하였다. 이러한 단어들은 비록 중의성 단어와 구문관계는 존재하지 않지만 의미해소에 도움이 되는 어휘들이다.

한국어는 영어와 같은 어순이 고정적인 언어와 달리 어순이 비교적 자유롭기 때문에 단어들의 문장에서의 상호위치로는 구문관계패턴을 정확하게 추출하기 어렵다. 따라서 의미 태깅된 한국어 말뭉치에 대해서는 부분적인 구문분석과 단어들간의 연관관계를 이용하여 구문관계패턴과 문맥정보를 추출한다. 〈표2〉에는 추출된 지역 및 문맥정보의 예문들이다. 'n024', 'n114', 'v312' 등은 가도까와 시소러스[7]의 개념 코드들인데, 'n으로 시작하는 코드는 명사성 개념을 뜻하며, 'v'로 시작하는 코드는 동사성 개념을 의미한다. 지식의 효율적인 적용을 위하여 구문관계패턴과 문맥정보에 나타난 개념들에 대하여 다음절에 기술할 통계적 방식으로 의미제약 변별력에 기반한 선별을 진행한다.

2.3. 개념정보의 구분

〈표2〉에서 눈(eye, snow, bud)의 $type_9$의 두 공기패턴에 크다('v243')의 개념이 같이 나타나고, 눈의 eye와 snow의 의미에 대한 문맥정보(UCW)에 사람('n507')이라는 개념이 같이 공기하고 있는 것을 관찰할 수 있다. 즉, 하나의 개념이 어휘의 여러 의미와 공기할 수 있기에 한 어휘의 여러 의미와 공기하는 개념의 의미 결정 변별력을 구할 필요가 있다. 어휘 W의 두 가지 의미 S_1, S_2에 대하여 S_1과 공기하는 개념과 개수는 $\{C_1(2), C_2(10), C_5(23)\}$이고, S_2와 공기하는 개념과 개수는 $\{C_1(20), C_3(4), C_4(12)\}$이라고 가정할 때, 개념 C_1은 의미 S_1과 S_2의 공기정보에 모두 존재한다. 그러나 C_1은 S_2와의 공기빈도가 크기 때문에 S_2의 의미결정에 크게 기여한다. 따라서 개념의 변별 과정을 거쳐 C_1을 S_2의 공기정보에 속하게 한다.

본 논문에서는 Shannon의 정보 이론에 기반한 분류 정보[2, 9]를 이용하여 중의성 단어에 대한 개념들의 변별력을 구하는 수식(1)과 수식(2)을 정의하였다. 개념코드 C_k가 어휘 의미 S_i와 공기할 수 있는 조건확률을 정규화하여 C_k가 의미 S_i에 전달하는 잡음(noise)을 엔트로피 공식으로 나타내면 수식(1)와 같다. 이 때, $noise_k$의 값은 0에서 $\log_2 n$사이의 값을 가지므로 수식(1)을 이용하여 정보 C_k가 갖는 분별값(Discrimination Score, DS)을 정의한다. 개념 C_k의 S_i에 대한 $noise_k$의 값이 크면 의미 S_i에 대한 결정력이 약해지고, 변별력 DS_k값이 작아진다. 개념

〈표2〉 "눈"(snow, eye, bud)에 대한 구문관계패턴 및 문맥정보

유형	LCP 유형	다의어	공기정보
LSP	$type_1$	눈(n024: snow)	송이(n114), 보라(n024), ...
	$type_2$	눈(n613: eye)	자기(n501), 동물(n061),...
	$type_3$	눈(n613: eye)	귀(n613), 코(n613), ...
	$type_5$	눈(n613: eye)	보이다(v312), 띄다(v299),...
	$type_9$	눈(n613: eye)	번뜩이다(v095), 크다(v243), ...
	$type_9$	눈(n053: bud)	크다(v243),.
	$type_{10}$	눈(n613: eye)	충혈된(v076), 감긴(v236), ...
UCW	다의어	문맥정보	
	눈(n613: eye)	고양(n061: cat), 인간(n507, n660: human), 모습(n110, n620: appearance), 기형(n110, n609: deformation), 사람(n507: human)	
	눈(n024: snow)	산(n032: mountain), 북부(n109: north), 전선(n102: battle line), 지구(n700: area), 영하 (n126: below zero), 가운데(n157, n192: among), 사람(n507: human)	

C_k가 한 단어의 여러 의미와 공기하면 그중에서 DS_k 값이 0.7보다 큰 의미 S_i를 선택하여 공기하게 한다. 만약 C_k가 S_i와 S_j간의 DS_k값이 같으면 의미결정 변별력이 약하기 때문에 0.7보다 값이 커도 사용하지 않는다. 아래의 수식에서는 단어 W의 의미를 n개로 가정하였다.

$$noise_k = -\sum_{i=1}^{n} \frac{p(C_k|S_i)}{\sum_{j=1}^{n} p(C_k|S_j)} \log_2 \frac{p(C_k|S_i)}{\sum_{j=1}^{n} p(C_k|S_j)} \quad (1)$$

$$DS_k = \frac{\log_2 n - noise_k}{\log_2 n} \quad (2)$$

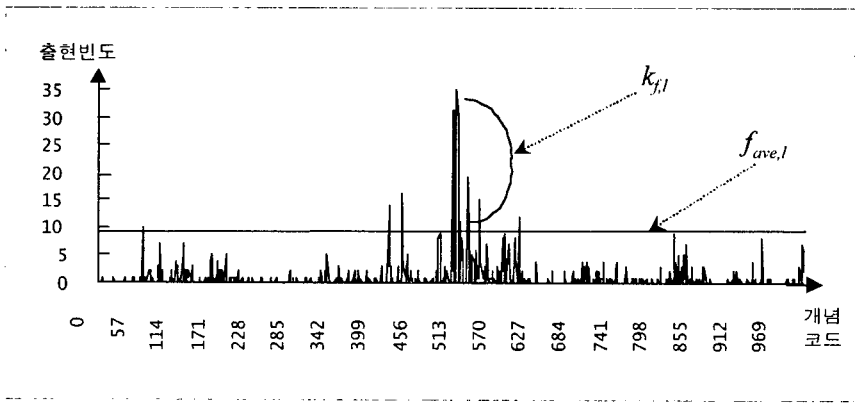
공기하는 개념들에 대하여 이와 같이 변별력을 구하여 의미에 따라 구분 및 선별한 후에도 어휘의 각 의미에 대한 구문관계패턴과 문맥정보에는 개념들이 많이 존재할 수 있다. 다음 절에는 이러한 개념 중에서 중의성 해소에 기여도가 가장 높은 대표적인 개념들을 추출하여 효율적인 지식원으로 사용하는 과정에 대하여 논의한다.

2.4. 개념의 일반화

앞절에서 기술한 개념의 선별과정을 거쳐 중의성 어휘에 관한 구문관계패턴과 문맥정보에는 패턴간의 중복된 개념들이 어휘 중의성 해소 중요도에 따라 선별 및 분리되었다. 앞에서 언급한 바와 같이 일본어 말뭉치의 번역문인 의미 태깅된 한국어 말뭉치에는 어휘들이 가도까와 시소러스의 L_{1000}계층의 의미코드로 표현된다. 따라서 정리된 개념정보에는 한 패턴안에 많으면 1000개의 개념이 존재할 수 있으므로, 이 중에서 중의성 해소에 기여도가 가장 높은 대표적인 개념들을 추출하여 효율적인 지식원으로 사용하는 것이 필요하다. (그림3)은 구문관계패턴의 개념코드와 출현빈도의 관계를 히스토그램 (Histogram)으로 표현한 것이다.

개념 코드 '514'처럼 공기빈도가 높은 개념들은 L_{1000}계층의 코드 그대로 사용하고, 비교적 적게 나타나지만 코드 '855' 주변에 가까이 있는 밀집한 공기 개념들에 대해서는 L_{1000}계층의 상위 계층 L_{100}의 코드로 표현한다. 공기 빈도가 높은 개념들은 중의성 어휘의 의미결정에 크게 영향을 미칠 수 있기 때문이다.(그림3)에 코드 '514'와 같은 빈도수가 높은 개념들을 선택한 후, 그 코드를 제거하면 개념코드 '500'에서 '600'사이에 남는 개념들 역시 밀집한 공기 개념들을 구성한다. 이러한 개념들에 대해서도 개념과 출현빈도관계를 L_{100}의 코드로 표현하고 지식을 추출한다. 이러한 과정을 본 논문에서는 개념의 필터링 (filtering)을 위한 일반화 과정으로 정의한다.

개념의 일반화는 개념 계층 L_{1000}과 L_{100}에서 실행된다. 개념의 일반화를 위하여 먼저 개념의 출현 빈도 분포를 고찰하여야 한다. (그림3)과 같은 히스토그램에서 모든 개념이 어휘의 의미와 고르게 공기하면 그래프는 평탄한 모양을 나타낸다. 이런 경우에는 특정한 자주 어울리는 개념이 존재한다고 판단하기 어렵다. 반면, 히스토그램에 개념 '514'와 같은 정점 (peak)모양을 가진 개념이 존재하면 그러한 개념을 추출하면 된다. 본 논문에서는 개념의 출현 빈도의 이와 같은 분포를 자동으로 고찰하고 개념들을 일반화하기 위하여 개념계층에서의 출현빈도 분포에 관한 표준편차 σ_l, 그리고 출현빈도의 상대적 크기를 나타



(그림3) 공기개념과 빈도관계를 나타내는 히스토그램

내는 $k_{f,l}$ 을 [10]을 참조하여 다음과 같이 정의한다.

$$\sigma_l=\sqrt{\frac{\sum_{k=1}^{n_l}(f_{k,l}-f_{ave,l})^2}{n_l-1}} \quad (3)$$

$$k_{f_{k,l},l}=\frac{f_{k,l}-f_{ave,l}}{\sigma_l} \quad (4)$$

수식에서 $f_{k,l}$은 개념 C_k의 가도까와 시소러스 계층 l에서의 출현빈도를 표현하고, $f_{ave,l}$은 전체개념(즉, L_{1000}에서의 1000개 개념)의 평균 출현빈도를 나타내고, n_l은 개념 계층[7] L_l에서의 개념노드의 개수이다.

위의 수식을 이용하여 일반화하는 과정은 다음과 같다. 우선 중의성 단어 W의 의미 S_i에 관한 공기패턴의 개념들의 출현빈도 f_i를 수식(3)을 이용하여 분포모양을 분석한다. 만약 표준편차의 값이 사전(previously)에 정의한 임계치(threshold) $\sigma_{0,l}$보다 크면 공기 개념의 분포에 정점 코드가 존재한다고 판정하고, 수식 (4)를 이용하여 그 정점 코드의 빈도의 상대 크기 $k_{f,l}$를 임계치 $k_{0,l}$와 비교하여 이보다 크면 해당 코드를 선택하고, 선택된 그 개념의 빈도는 0으로 지정하여 상위 계층의 처리과정에서 다시 고려하지 않는다. 다음으로는 개념 계층 L_l에서 일반화되지 못한 나머지의 값들을 이용하여 상위 계층에서 일반화를 시도한다. 예를 들면, 계층 L_{1000}에서 일반화 과정을 거친후 상위 계층인 L_{100}에서 고찰하기 위하여, L_{1000}의 개념 코드 '100'부터 '109'까지의 유사한 개념의 빈도를 합산하여 상위 개념인 '10'의 출현 빈도로 지정한다.

정확한 개념 코드를 찾아내기 위하여 표준편차의 임계치 (threshold) $\sigma_{0,l}$와 코드 출현 빈도의 임계치 $k_{0,l}$를 실험을 통하여 〈표3〉과 같이 지정한다. $\sigma_{0,l}$와 $k_{0,l}$의 값을 작게 지정하면 계층 L_l에서 더 많은 개념 코드가 추출되므로 의미 중의성 해소의 정확성은 높아질 수 있으나 추출된 지식들이 많아지므로 저장공간이 크게 되고 일반적인 적용률이 낮아진다. 중의성 해소 시스템의 성능을 고려하여 $\sigma_{0,l}$와 $k_{0,l}$의 값은 낮은(low) 계층에서 뽑은 개념들의 수와 좀 더 높은 계층에서 뽑는 개념의 수 사이의 균형을 이룬다. 낮은 계층에 있는 개념 코드는 많이 사용되는 특정한 공기 개념들을 중요시하고, 높은 계층의 개념들은 일반적인 개념들이어서 적용률을 향상시킨다.

〈표3〉 일반화 모듈의 임계치 설정

Level	표준편차의 σ_l의 임계치	출현빈도 크기 $k_{f,l}$의 임계치
L_{1000}	$\sigma_{0,1000}$=1.5	$k_{0,1000}$=6.0
L_{100}	$\sigma_{0,100}$=4	$k_{0,100}$=3.0

〈표4〉는 개념의 구분과정을 거친 중의성 단어 "눈"에 관한 $type_2$의 공기 개념과 출현 빈도이다. 〈표4〉에 등록된 내용들은 해당 코드가 4번 이상 나타난 개념 코드와 출현 빈도이다. 출현 빈도가 4보다 적은 코드들은 상대적으로 중요도가 약하므로, 이들의 빈도 합

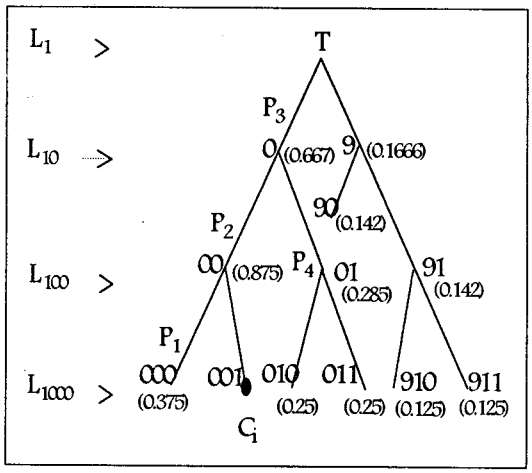
〈표4〉 "눈"(eye)의 구문관계패턴 $type_2$의 공기개념과 빈도

Code	Freq.	Code	Freq.	Code	Freq.	Code	Freq.	Code	Freq.	Code	Freq.
28	12	61	7	103	4	107	8	121	7	126	4
143	8	160	5	179	7	277	4	320	8	331	6
416	7	419	12	433	4	501	13	503	10	504	11
505	6	507	12	508	27	513	5	530	6	538	16
552	4	557	7	573	5	709	5	718	5	719	4
733	5	819	4	834	4	966	4	987	9	other	210

은 other에 지정한다. 개념의 일반화를 통해 최종 추출된 구문관계패턴은 ({'n028', 'n419', 'n501', 'n504', 'n507', 'n508', 'n538', 'n50'}, $type_2$, "눈"(eye)) 이다. 이러한 정보는 다음 절에서 서술할 어휘중의성 해소 알고리즘에서 사용된다.

3. 명사 중의성 해소 방법

명사 중의성 해소는 동사의 선택제약, 명사의 구문관계패턴, 문맥정보 및 어휘출현 빈도 등 지식들을 사용한다. 동사의 선택제약은 한-일 번역기 개발과정에서 중의성 동사에 대해 수작업으로 작성된 공기패턴이다. 명사 W에 대한 위에 언급한 지식을 이용한 중의성 해소 알고리즘은 수식 (5), (6), (7), (8) 등을 통하여 구현된다. 여기에서 S(W)는 중의성 명사 W의 의미 집합이며, P(V)는 W와 입력 문장에서 같이 나타나는 동사 V와의 해당하는 구문관계의 선택제약 집합이고, LSP(W)는 구문관계패턴 정보를, 그리고 USW(W)는 W의 문맥정보를 표현한다. C_i와 P_j는 개념 유형을 표현하고, S_k는 W의 k번째 의미를 뜻한다. 수식 (8)의 $Csim(C_i, P_j)$는 가도까와 시소러스에 기반한 개념 C_i 와 P_j사이의 유사도를 계산하는 수식이다. 수식(8)에서 weight는 개념의 가중치를 표현함으로, 유사도 계산시 개념 C_i의 부모(parent) 개념이 형제(sibling) 개념보다 유사한 특징을 더 많이 가지고 있으며, 이러한 관계를 중요시 한다는 것을 뜻한다. 개념 C_i가 P_j의 시소러스상의 하위 개념이면 weight를 1로 지정하고, 그렇지 않으면 0.5의 값을 지정함으로서 유사도 값을 감소시킨다. 수식 (8)의 MSCA(Most Specific Common Ancestor)는 두 개념의 공유하고 있는 가장 가까운 상위 개념을 가리킨다. 이러한 관계를 반영하여 개념간의 유사도를 계산하면 (그림4)에서 표현한 것과 같다. 개념 C_i와 개념 P_1, P_2, P_3간의 유사도는 모두 0.3보다 크고, P_4 및 기타 개념간의 유사도는 0.3보다 작다. 이러한 특성을 고려하여 명사 중의성 알고리즘의 실행에 필요한 임계치 T_1, T_2, T_3을 0.3으로 결정한다.



(그림4) 개념 계층 구조에 기반한 개념 유사도 계산

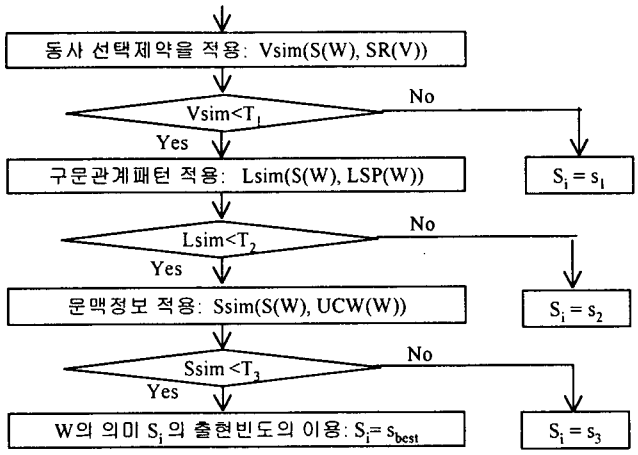
중의성 단어 W에 대하여 의미 결정과정은 다음과 같다. 단어 W가 동사 V의 매개변수로 사용되면 동사 V의 선택제약 (selectional restriction, SR) 만족도 Vsim(S(W),SR(V))를 계산한다. 계산 값이 임계치 값 T_1보다 크면 W의 의미를 이 단계에서 결정한다. 계산값이 T_1보다 작으면 W의 구문관계패턴을 적용하여 지역정보 만족도 Lsim(S(W),LCP(W))를 계산하고 임계치 T_2보다 크면 W의 의미를 이 단계에서 결정한다. 다음으로 문맥정보를 적용하여 유사도 Ssim(S(W),USW(W))을 계산하고 임계치 T_3과의 비교를 거쳐 의미를 결정한다. 위의 지식들을 적용해도 W의 의미를 결정하지 못할 때에는 W의 말뭉치에서의 의미 사용 빈도가 큰 것을 선택한다. 본 논문에서는 임계치 T_1, T_2, T_3의 값들을 (그림4)에서 나타낸 개념간의 유사도 관계를 기반하여 0.3으로 정하였다. 즉, 개념간에 최상위 계층 L_1의 개념을 제외한

$$Vsim(S(W), SR(V)) = \max_i (Csim(C_i, P_j)), 1 \le i \le n; 1 \le j \le m; C_i \in S(W); P_j \in SR(V) \quad (5)$$

$$Lsim(S(W), LSP(W)) = \max_k (Csim(C_i, P_{k,j,l})), 1 \le i, k \le n; 1 \le j \le 10; 1 \le l \le w; P_{k,j,l} \in LSP_j(S_k) \quad (6)$$

$$Ssim_i(S(W), UCW(W)) = \max_k (Csim(C_i, P_{k,j})), 1 \le i, k \le n; 1 \le j \le r; P_{k,j} \in UCW(S_k) \quad (7)$$

$$Csim(C_i, P_j) = \frac{2 * level(MSCA(C_i, P_j))}{level(C_i) + level(P_j)} * weight \quad (8)$$



(그림5) 명사 중의성 해소 알고리즘

개념 중에서 최하위 계층의 친 형제 또는 직속 친족의 관계가 없으면 유사하지 않고 의미제약이 없다고 간주한다. 즉, (그림4)에서 C_i에 대해 P_1, P_2, P_3에 해당하는 의미코드가 임계치 이상의 유사도를 갖는 것으로 간주한다. 수식에서 'n'은 어휘 W의 의미 개수이고, 'm'은 문장에서 동사 V의 W의 격에 해당하는 선택제약의 개념코드 개수이고, 'w'는 의미 S_k의 j번째 구문관계패턴의 개념코드 개수이고, 'r'은 의미 S_k의 문맥정보의 개념코드 개수이다.

4. 실험 및 평가

구문관계패턴과 문맥정보의 추출을 위하여 일본 아사히 신문, 일본 경제 신문 및 EDR 사전 말뭉치의 약 25만 문장으로부터 의미태깅된 한국어 말뭉치를 생성하였다. 본 연구의 성능 평가를 위하여 먼저 명사중의성 해소 알고리즘을 한-일 번역기에 구현하였다. 그리고 자주 나타나는 중의성 명사 8개를 선택하고, 각 의미별 분포가 비슷하도록 국어정보베이스에서 임의로 404개의 문장을 수집한 후 이들에 대한 중의성 해소 결과를 조사하여 보았다. 본 실험에서는 (그림5)에서 나타낸 중의성 해소 과정에서 동사 선택제약(selectional restriction, SR), 구문관계패턴(local syntactic pattern, LSP), 문맥정보(unordered co-occurring word, UCW) 및 어휘의 의미 출현빈도 등 지식의 적용 및 중의성 해소 성공과 실패 여부를 〈표5〉에 나타내었다. 기호 'O'는 중의성이 해소된 것을 의미하며, 'X'는 실패한 것을 의미한다. 실험에서 평균 82.4%의 정확률을 보였으며, 이는 자주 나타나는 의미를 지정하는데 비해 14.6%의 정확도 향상을 보였다. 동사의 선택제약 정확률은 96%이었으나 적용률은 6.9%이었다. 반면 구문관계패턴 정보는 15.8%의 적용률과 84.3%의 정확률을 보였고, 문맥정보는 45.5%의 적용률과 85.3%의 정확성을 나타냈다.

5. 결론

본 논문에서는 말뭉치와 개념정보를 이용한 명사중의성 해소방법을 제안하였으며 실제 한-일 기계번역에 적용하여 중의성 해소의 효율성을 검증하였다. 자동으로 의미태깅된 말뭉치로부터 각 중의성 명사에 대한 구문관계패턴과 문맥정보를 획득하였으며, 이를 통계처리하여 중의성 해소에 이용하였다.

제안한 방법의 평가를 위해 8개의 중의성 명사를 선정하고, 이들 명사를 포함한 404문장에 대해서 실험한 결과 평균 82.4%의 정확률을 보였다. 실험에서 구문관계패턴 및 문맥 정보에 대한 개념 일반화를 수행하여 적용률을 향상할 수 있었고, 자료부족 문제도 일부 해결할 수 있었으며, 저장되는 정보의 크기도 적당한 수준에서 유지할 수 있었다.

실험에서 기존의 일-한 번역기를 사용하여 의미태깅된 한국어 말뭉치를 생성할 때, 한국어의 일부 어휘와 어휘의 일부 의미를 생성하지 못하는 문제점을 관찰 할 수 있었다. 이것은 하나의 일본어 단어에 여러 개의 한국어 단어가 대응될 수 있는데, 일-한 사전에는 그 중의 일부만 기록했기 때문이다. 이러한 번역문에서 나타나지 않은 단어의 개념 공기패턴의 추출은 향후 연구과제로 추진할 것이다.

〈표5〉 중의성 명사 해소의 실험 및 평가

중의성 단어	의미	지식 / 개수	SR		LSP		UCW		의미빈도		계	
			O	X	O	X	O	X	O	X	O	X
부자	father & child	40			4	6	16	2	12		32	8
	rich man	12					4	3		5	4	8
간장	liver	33			3		13	3	14		30	3
	soy sauce	16	1		2		10			3	13	3
가사	housework	25			6		12	1	6		24	1
	words of song	27					22			5	22	5
구두	shoes	39	9		3		16	2	9		37	2
	word of mouth	10					5	2		3	5	5
눈	eye	41	6		8		3		24		41	0
	snow	9	4		1		2	1		1	7	2
용기	courage	31			1	2	17		11		29	2
	container	19	1		6		1	3		8	8	11
경비	expenses	38	3		13		7	6	9		32	6
	defense	13	2			2	6			3	8	5
경기	times	27	1		7		8	1	10		26	1
	match	24		1			15	3		5	15	9
평균 정확률 (%)		404	96		84.3		85.3		74.2		82.4	

참고문헌

[1] 이승우, 이근배, 국소 문맥과 공기 정보를 이용한 비교사 학습 방식의 명사 의미 중의성 해소, HCI99 학술발표대회 논문집, pp839-845,1999.

[2] 이호, 백대호, 임해창, 최소한의 코퍼스 정보를 이용한 단어 의미 중의성 해결 기법, 한국정보과학회 봄 학술발표논문집, 24권 1호, pp.467-470, 1997.

[3] R. Bruce and J. Wiebe, Word-Sense Disambiguation Using Decomposable Models, in *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, pp.139-145, 1994.

[4] Tou Ng Hwee and Hian Beng Lee, Integrating Multiple Knowledge Sources to Disambiguate Word Sense: an Exemplar-Based Approach, in *Proc. of 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, pp.40-47, 1996.

[5] Edward Kelly and Phillip Stone, *Computer Recognition of English Word Senses*, North-Holland, Amsterdam, 1975.

[6] Susan W. McRoy, Using Multiple Knowledge Sources for Word Sense Discrimination, *Computational Linguistics*, Vol.18, No.1, pp.1-30, 1992.

[7] S. Ohno and M. Hamanishi, *New Synonym Dictionary*, Kadokawa Shoten, Tokyo, 1981 (written in Japanese).

[8] Chul-Jae Park, Jong-Hyeok Lee, Geunbae Lee and K. Kakechi, Collocation-Based

Transfer Method in Japanese-Korean Machine Translation, *Transaction of Information Processing Society of Japan*, Vol.38, No.4, pp.707-718, 1997 (written in Japanese).

[9] C. E. Shannon, Prediction and Entropy of Printed English, *Bell System Technical Journal*, pp.50-65, 1951.

[10] Frank Smadja, Retrieving Collocations from Text: Xtract, *Computational Linguistics*, Vol.19, No.1, pp.143-177, 1993.

[11] David Yarowsky, Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in *Proc. of 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Massachusetts, pp.189-196, 1995.