# WordNet을 이용한 한국어 시소러스 자동 구축*

이 창 기, 이 근 배
포항공과대학교 컴퓨터공학과
경북 포항시 남구 효자동 산31번지
우: 790-784
leeck@nlp.postech.ac.kr, gblee@postech.ac.kr

# Using WordNet for the Automatic Construction of Korean Thesaurus

Changki Lee, Geunbae Lee
Natural Language Processing Lab.,
Dept. of Computer Science and Engineering, POSTECH.

**요 약**

최근의 자연어 처리 분야의 연구들에서 광범위하고 완전한 어휘 지식 베이스의 필요성이 입증되었다. 영어권의 경우, 이에 대한 연구가 오래 전부터 있어 왔고, 그 결과로 현재 주로 사용되고 있는 개념체계에는 Roget's Thesaurus와 WordNet 등이 있다. 이러한 개념체계들은 자연어 처리의 여러 응용 분야에서 중요한 역할을 담당하고 있지만, 다른 언어의 경우 널리 사용되고 있는 개념체계가 없는 실정이다.

본 논문에서는 Princeton 대학의 WordNet을 기반으로 한영 사전과 국어 사전을 이용하여 한국어 명사의 개념체계를 자동으로 구축함으로써, 이미 구축되어진 다른 언어의 개념체계를 이용하여 새로운 언어의 개념체계를 자동으로 구축할 수 있음을 보인다. 먼저 한영 사전과 국어 사전으로부터 뽑아낸 한국어 단어 일부의 의미를 다양한 WSD(Word Sense Disambiguation) 방법을 적용시켜 WordNet의 synset에 자동으로 연결시킬 수 있음을 보인다. 그리고 각각의 자동변환으로 나온 결과들에 대해서 적용율과 정확도를 비교하도록 한다.

## 1 서론

사람이 자연어 문장을 이해하는 데에는 각자가 가진 상식이나 지식, 단어 개념 등의 지식 베이스를 이용한다. 이와 같이 일반적인 사람이 가진 지식 베이스를 컴퓨터에 도입하려는 많은 노력이 있었다. 이러한 지식 베이스를 개념체계(ontology) 혹은 시소러스 (thesaurus) 라고 한다. 최근의 자연어 처리 분야의 여러 연구들에서 광범위하고 완전한 어휘 지식 베이스의 필요성이 입증되었다. 영어권의 경우, 이에 대한 연구가 오래 전부터 있어 왔고, 그 결과로 현재 주로 사용되고 있는 개념체계에는 Roget's Thesaurus와 WordNet 등이 있다. 이러한 개념체계들은 자연어 처리의 여러 응용 분야에서 중요한 역할을 담당하고 있지만, 다른 언어의 경우 널리 사용되고 있는 개념체계가 없는 실정이다.

개념체계를 구축할 경우, 가장 확실하고 정확한 방법은 수동으로 구축하는 것이다. 그러나 이것은 많은 비용과 시간이 필요하다는 단점을 가지고 있다. 또한 어플리케이션에 따라 수동으로 구축된 것만큼의 정확도가 필요하지 않은 경우도 있다. 이러한 이유로 이미 구축되어진 많은 어휘 정보들을 이용하여 대량의 어휘 지식을 자동 혹은 반자동으로 얻어내려는 많은 연구들이 있었다.

본 논문에서는 영어권에서 사실상 표준이 되어가고 있는 Princeton 대학의 WordNet을 기반으로 한영 사전과 국어 사전을 이용하여 한국어 명사의 개념체계를 자동으로 구축함으로써, 이미 구축되어진 다른 언어의 개념체계를 이용하여 새로운 언어의 개념체계를 자동으로 구축할 수 있음을 보인다. 먼저 한영 사전과 국어 사전으로부터

* 본 연구는 과학재단 특정기초(1997.9 - 2000.8) 연구비 지원으로 이루어진 것임.

추출한 한국어 단어의 의미를 다양한 방법을 적용시켜 WordNet의 synset에 완전 자동으로 연결시킬 수 있음을 보인다. 그리고 각각의 자동변환으로 나온 결과들에 대해서 적용율과 정확도를 비교하도록 한다.

## 2 관련 연구

개념체계를 구축하기 위한 방법은 여러 가지가 있으나 가장 정확하고 신뢰할 수 있는 방법은 수동으로 구축하는 것이다. 그러나 이것은 언어학자나 심리학자의 도움이 필요한 매우 어려운 작업이고 시간과 비용이 많이 걸리는 단점이 있다. 이러한 이유로 기존에 존재하는 어휘 지식 정보를 가지고 자동이나 반자동으로 개념체계를 구축하려는 많은 연구들이 있었다. 이러한 방법 중에 하나가 기존의 사전을 이용하여, 단어의 상의어를 뽑아내어 Bottom-up 방식으로 개념체계를 구축하는 것이다. 이러한 방법은 사전의 정의문으로부터 단어의 상의어를 자동 혹은 반자동으로 추출하고 이 상의어들을 연결함으로써 개념체계를 자동 혹은 반자동으로 구축하게 된다. 이러한 방법을 사용하기 위해서는 사전으로부터 추출한 상의어의 올바른 의미를 선택하는 작업(Genus Disambiguation)이 필요한데, 이를 위해서는 WSD(Word Sense Disambiguation) 작업이 필요하다. 그러나 이러한 방법은 사전에 따라 상의어가 다를 수 있고, 정의문 안에 정확한 상의어가 없거나 정의에 loop가 존재할 수 있는 문제가 있다. 또한 상의어의 올바른 의미를 선택할 때의 정확도가 문제가 된다.

[3]에서는 사전으로부터 상의어를 자동으로 추출한 후, 상의어의 올바른 의미를 선택하기 위해서 [2]에서 사용했던 다양한 휴리스틱의 조합을 이용한 Genus Disambiguation 기법을 이용하여, 자동으로 개념체계를 구축하였다.

[6]에서는 사전에서 상의어를 자동으로 추출하고, 다의어의 의미를 a, b, c, ... 등으로 구분하며 사람이 개입하여 상의어의 올바른 의미를 선택하게 하여 한국어 명사의 개념체계를 구축하였다.

위에서 설명한 방법과는 다르게 대역어 사전(bilingual dictionary)을 가지고 이미 구축되어진 다른 언어의 개념체계를 이용하여 새로운 언어의 개념체계를 구축하려는 연구들이 있었다. 이러한 방법은 대역어 사전을 이용하여 새로운 언어의 단어의 의미를 다른 언어의 개념체계에 연결시키는 방법을 사용하였다. 이 방법에서는 대역어의 의미가 여러 개 있을 경우, 대역어의 올바른 의미를 선택하기 위해서 WSD 작업이 필요하다. 이러한 방법은 사용하는 언어와 문화의 차이에 따른 문제점이 있지만, 비교적 쉽고 빠르게 구축할 수 있고, 다국어 개념체계를 구축할 수 있다는 장점을 가지고 있다.

[4]에서는 영어 WordNet과 대역어 사전을 이용하여 스페인어 WordNet을 자동으로 구축하였다. 여기에서 사용한 방법은 Class methods와 Structural methods, 그리고 Conceptual Distance methods이다. Class methods는 스페인어와 영어 대역어의 관계에 따라 여러 클래스로 분류하여 스페인어의 의미를 WordNet의 synset에 연결하는 방법이다. 여러 클래스 중에서 실제로 사용한 것은 스페인어의 영어 대역어의 의미 수가 1인 경우(Monosemic Criteria)와 영어 대역어들의 의미가 같은 경우(Hybrid Criteria)이다. 나머지 클래스는 정확도가 낮아 사용하지 않았다. 이 Class methods는 비교적 높은 정확도를 보이지만, 영어 대역어의 의미 수가 1 이거나 영어 대역어들의 의미 중에 같은 것이 있어야 적용할 수 있으므로 적용율이 매우 낮다는 단점이 있다. Structural methods는 영어 WordNet의 계층 구조를 이용하여, 스페인어의 영어 대역어의 의미들 중에서 Intersection criterion, Parent criterion, Brother criterion, Distant hyperonymy criterion 등의 조건을 만족하는 의미를 선택하는 방법이다. 그러나 이 방법은 정해진 패턴에 맞는 경우에만 적용할 수 있어서 적용율이 매우 낮고, 정확도도 낮아서, 실제로 스페인어 WordNet을 구축할 때는 사용하지 않았다. Conceptual Distance methods 역시 정확도가 낮기 때문에 실제로는 사용하지 않았다. 그러나, 정확도가 낮은 방법들의 결과를 수동으로 조합해본 결과 정확도가 높아지는 것이 있어서, 이러한 것들을 스페인어 WordNet에 포함시켰다.

[7]에서는 국어사전으로부터 명사의 상의어를 자동으로 추출하여 상의어 사전을 만들고, 한영사전과 영한사전을 이용하여 영어 WordNet을 번역한 후, 상의어 사전과 번역된 영어 WordNet을 합성하였다. 마지막으로 이 합성된 한국어 명사 WordNet의 예비 리스트의 pruning 작업을 하여 한국어 명사 WordNet을 구축하였다. 이 방법의 pruning 작업은 무척 까다롭고 시간이 많이 걸리는 작업으로 상식과 언어학자의 어휘 개념을 참조하여 수작업으로 수행하였다.

## 3 한국어 시소러스 자동 구축

본 논문에서는 한국어 시소러스를 구축하기 위해서 한국어 단어의 의미를 WordNet의 synset에 연결하는 방법을 사용한다. 한국어 단어의 의미를 WordNet에 연결시키기 위해서는 한국어 단어의 의미에 맞는 영어 대역어가 필요한데, 이를 위해서 한영 대역어 사전을 이용한다. 또한 이 방법을 사용하기 위해서는 영어 대역어의 올바른 의미를 선택하기 위해서 WSD 작업이 필요하다. 영어 대역어의 올바른 의미를 선택하기 위한 WSD 작업을 수행하기 위해서 다음과 같은 두 가지의 휴리스틱을 사용한다.

1. 하나의 한국어 단어에 대한 영어 대역어들의 의미는 서로 유사하다.
2. 두 개의 한국어 단어가 상하위 관계를 갖는다면, 그들의 영어 대역어들 중에도 상하위 관계를 갖는 것이 존재한다.

3.1 장에서 첫 번째 휴리스틱을 사용한 방법을 설명하고, 3.2 장에서 두 번째 휴리스틱을 사용한 방법을 설명하도록 한다.

### 3.1 유사도 행렬을 이용한 방법

이 방법에서는 "하나의 한국어 단어에 대한 영어 대역어들의 의미는 서로 유사하다" 라는 휴리스틱을 사용한다. 먼저 한국어 단어의 각 의미에 해당하는 영어 대역어들을 한영 사전을 이용하여 구한다. 다음에 영어 대역어의 가능한 의미(synset) 후보를 구하기 위해서 WordNet을 이용한다. 각 영어 대역어에 대해 그 의미 수가 1인 경우, 그 의미를 선택한다. 각 영어 대역어의 의미 수가 2 이상인 경우, 각 영어 대역어들의 의미 후보들로 구성된 유사도 행렬을 구성하여 올바른 의미를 선택하게 된다. 다음의 알고리즘은 한국어 단어의 의미 수가 1인 경우(monosemous)의 알고리즘이다. 한국어 단어의 의미 수가 2 이상인 경우에는 각각의 의미에 대하여 다음의 알고리즘을 따로 적용시키면 된다.

**Step A.1 : 대역어 1, 대역어의 의미 1**

한국어 단어의 영어 대역어 수가 1 이고, 영어 대역어의 의미 수가 1 이면, 영어 대역어의 의미를 선택한다.

**Step A.2 : 대역어 1, 대역어의 의미 2 이상**

영어 대역어 수가 1 이고, 영어 대역어의 의미 수가 2 이상인 경우, 대역어간의 유사도를 비교할 수 없으므로, 3.2절의 방법으로 처리한다.

**Step A.3 : 대역어 수가 2 이상**

영어 대역어 수가 2 이상이면, 대역어들간의 유사도 행렬을 만든다.

먼저 각각의 영어 대역어들의 의미(synset) 후보들을 구한 후, 이러한 의미 후보들로 이루어진 유사도 행렬을 만든다.

$$S_{ij}(l,m) = \begin{cases} sim(s_{il}, s_{jm}) & \text{if } i \neq j \text{ and } sim(s_{il}, s_{jm}) > \theta \\ 0 & otherwise \end{cases}$$

$$where,\ l \in [1, n_i],\ m \in [1, n_j]$$

**Step B**

각 대역어($ew_i$)의 각 의미($s_{ij}$)가 다른 대역어($ew_j$)로부터 얻는 지지도를 구하고, 가장 큰 지지도 합계를 받는 대역어와 그 의미를 구한다.

$$support(s_{il}, ew_j) = \max_{where,\ m \in [1, n_j]} S_{ij}(l,m)$$

$$(i_{max}, l_{max}) = \arg\max_{i,l} \sum_{j=1}^{k} support(s_{il}, ew_j)$$

$$where,\ l \in [1, n_i],\ k = \text{대역어 수}$$

**Step C**

$ew_i$의 의미를 $l_{max}$로 결정하고, 유사도 행렬을 갱신한다.

$$S_{i_{max}j}(l,m) \leftarrow 0 \quad \text{if } l \neq l_{max}$$

$$where,\ j \in [1, k],\ l \in [1, n_{i_{max}}],\ m \in [1, n_j]$$

**Step D**

모든 대역어들의 의미를 결정할 때까지 Step B 부터 반복한다.

**Step E**

지지도의 평균값이 임계값을 넘는 의미를 한국어 단어의 의미로 결정한다.

한국어 단어 의미 =

$$\{x \mid x = s_{il}, \frac{1}{k} \times \sum_{j=1}^{k} support(s_{il}, ew_j) > \theta\}$$

위 식에서 사용한 유사도 측정 식은 다음과 같다.

$$sim(a,b) = \frac{2 \times level(MSCA(a,b))}{level(a) + level(b)}$$

위 식에서, MSCA(a,b)는 a와 b의 공통 부모 중에서 가장 구체적인 의미이다.

유사도 행렬을 이용한 방법은 [8]에서 사용한 비교사 학습의 WSD 알고리즘을 이용한 것인데, 이는 Dekang Lin의 아이디어인 "서로 다른 명사가 동일한 문맥 내에 나타나면 그 두 명사는 유사한 의미를 가진다"라는 휴리스틱[1]을 참고하였고, 동일한 문맥 내에 나타난 단어들의 유사도 행렬을 이용하여 단어의 의미를 선택하였다.

스페인어 WordNet을 구축하기 위해 사용했던 방법과 본 유사도 행렬을 이용한 방법을 비교해 보면, Class Methods나 Structural Methods는 특정 조건을 만족해야만 적용할 수 있기 때문에 적용율이 매우 낮다. 또한 여러 대역어들이 있을 때, 이러한 대역어들을 모두 이용하는 것이 아니라 특정한 패턴을 만족하는 대역어만을 이용하게 된다. 그러나 유사도 행렬을 이용한 방법은 특정한 패턴을 사용하지 않고, 전체 대역어간의 유사도 행렬을 이용하기 때문에 두 방법에 비해 높은 정확도와 적용율을 보인다.

## 3.2 상하위 관계를 이용한 방법

이 방법에서는 "두 개의 한국어 단어가 상하위 관계를 갖는다면, 그들의 영어 대역어들 중에도 상하위 관계를 갖는 것이 존재한다"라는 휴리스틱을 사용하여 영어 대역어의 올바른 의미를 선택한다. 본 논문에서는 국어 사전으로부터 단순한 규칙을 이용하여 추출한 상의어만을 사용하였다. 만약 기존에 한국어 개념체계가 존재한다면, 이로부터 상의어뿐만 아니라 하의어 등도 이용할 수 있으므로 더 정확한 결과를 얻을 수 있을 것이다. 이러한 경우 하나의 개념체계를 다른 개념체계와 연결시키는 일도 가능해진다.

이 방법의 알고리즘은 다음과 같다.

**Step A : 한국어 단어의 각 의미에 해당하는 상의어를 국어사전으로부터 추출**

한국어 단어가 다의어인 경우에는 각 의미에 따라 상의어도 달라지게 된다. 예를 들면 "정처"라는 단어가 있을 때, 이 단어의 상의어는 "본처"와 "곳"이 된다. 물론 "본처"와 "곳"이라는 단어가 다의어일 수도 있으나, 이 단계에서는 상의어의 의미는 구분하지 않는다. 상의어의 의미를 구분하는 작업은 Step C 에서 하게 된다.

**Step B : 한국어 단어와 그 상의어의 영어 대역어들을 한영사전을 이용하여 구함**

이 단계에서는 한국어 단어의 각 의미에 해당하는 영어 대역어와, 한국어 단어의 상의어의 영어 대역어를 구한다.

**Step C : 상하위 관계를 이용하여 영어 대역어의 올바른 의미를 선택**

한국어 단어의 각 의미에 해당하는 영어 대역어들과, 한국어 단어의 상의어들의 영어 대역어들

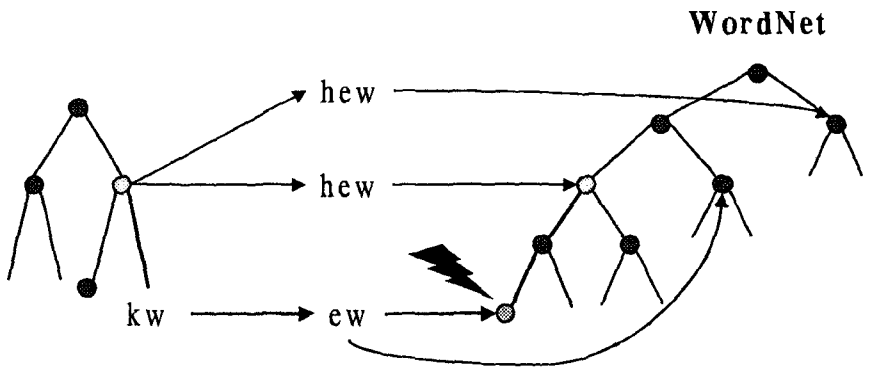


그림 1 : 상하위 관계를 이용한 방법

간에 상하위 관계가 있는지를 WordNet을 이용하여 그림 1처럼 검사한다. 다음의 식에서 "ew"는 한국어 단어의 영어 대역어이고, "hew"는 한국어 단어의 상의어의 영어 대역어이다.

$$S(ew,i) = \begin{cases} sim(sense(ew,i), sense(hew,j)) \\ \quad \text{if } sense(ew,i) \ ISA \ sense(hew,j) \\ 0 \qquad otherwise \end{cases}$$

$$where,\ i \in [1,\ n_{ew}],\ j \in [1,\ n_{hew}]$$

$$i_{max} = \arg\max_i S(ew,i)$$

$$where,\ i \in [1,\ n_{ew}]$$

**Step D : 임계값을 넘는 의미를 선택**

S(ew,$i_{max}$)가 임계값(threshold)을 넘으면, 한국어 단어의 의미를 WordNet의 synset인 sense(ew,$i_{max}$)에 연결한다.

$$\text{한국어 단어 의미} = \{x \mid x = sense(ew, i_{max}),\ S(ew, i_{max}) > \theta\}$$

상하위 관계를 이용한 방법은 유사도 행렬을 이용한 방법에서는 적용할 수 없었던 부분을 해결할 수 있다. 즉, 대역어의 수가 1이고 대역어의 의미가 2이상인 경우, 유사도 행렬을 이용한 방법에서는 대역어간의 유사도를 비교할 수 없으므로 적용할 수 없지만, 상하위 관계를 이용한 방법에서는 그 단어의 상의어의 대역어를 찾아 원래 단어의 대역어의 의미와 비교할 수 있으므로 적용이 가능하다. 그러나 이 방법은 한국어 단어가 다의어일 경우, 상의어가 여러 개 나오므로 나머지 상의어가 올바른 의미의 상의어에 잡음의 역할을 하게 되어 정확도가 떨어지게 된다.

## 3.3 실험 및 결과

실험에서 사용한 한영사전은 한국어 명사의 영어 대역어 사전이며, 총 23,478 개로 구성되어 있고, 한국어 단어의 각 의미별로 영어 대역어가 나와 있다. 본 실험에서는 한영사전에 나온 한국어 단어의 의미별로 적용율과 정확도를 측정하였다.

### 3.3.1 적용율

한영사전에 나와있는 한국어 단어의 의미를 기준으로 하여 얼마나 많은 한국어 단어의 의미가 WordNet의 synset에 연결되었는가를 측정하였다. 한영사전의 한국어 단어는 총 23,478 개이고, 한영사전의 한국어 단어의 의미는 총 30,656 개다. 이로부터 한국어 명사의 평균 중의성은 대략 1.31임을 알 수 있다.

30,656 개의 의미 중에서 유사도 행렬을 이용한 방법은 19470개의 의미를 WordNet의 synset에 연결시켰으며, 나머지 11186개의 의미는 연결하는데 실패하였다. 이 방법에서 임계값을 0.4로 했으며, 적용율은 63.5%이었다.

상하위 관계를 이용한 방법은 30,656 개의 의미 중에서 13,035 개의 의미를 WordNet의 synset에 연결시켰다. 이 방법에서 임계값은 0.4로 했으며, 적용율은 42.5%이었다.

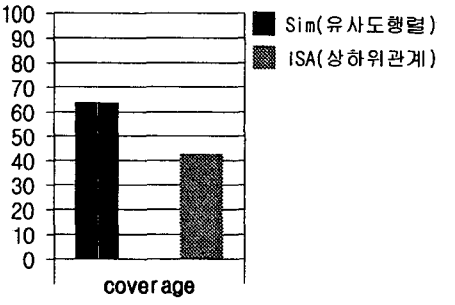


그림 2 : 적용율 (Sim, ISA)

### 3.3.2 정확도

정확도는 한영사전에 나온 한국어 단어의 영어 대역어의 여러 후보 중에서 얼마나 정확하게 올바른 의미를 선택했는지를 기준으로 하였다.

유사도 행렬을 이용한 방법은 임계값을 0.4로 했으며, 결과에서 무작위로 추출하여 정확도를 측정하였다. 상하위 관계를 이용한 방법도 임계값을 0.4로 했으며, 결과에서 무작위로 추출하여 정확도를 측정하였다.

유사도 행렬을 이용한 방법은 정확도가 90.3%이었고, 상하위 관계를 이용한 방법은 76.2%이었

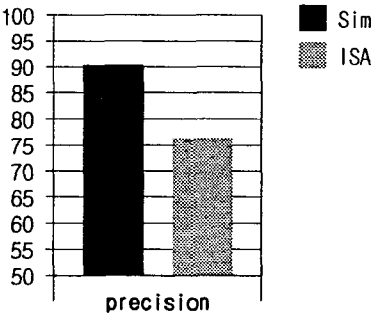


그림 3 : 정확도 (Sim, ISA)

다.

유사도 행렬을 이용한 방법은 높은 정확도를 보였으나, 상하위 관계를 이용한 방법은 80%에 못미치는 낮은 정확도를 보였다.

의미 수에 따른 각각의 정확도를 측정하기 위해서, 두 방법에 대해서 한국어의 의미 수가 1인 경우(monosemous)와 한국어의 의미 수가 2 이상인 경우(polysemous)를 나누어 시험을 해보았다. 실험 결과 한국어 단어의 의미 수가 1인 경우에는 정확도가 유사도 행렬을 이용한 방법에서는 92.8% 이고, 상하위 관계를 이용한 방법에서는 81.4% 이었다. 의미 수가 2 이상인 경우에는 유사도 행렬을 이용한 방법에서는 85.1% 이고, 상하위 관계를 이용한 방법에서는 65.4% 이었다.

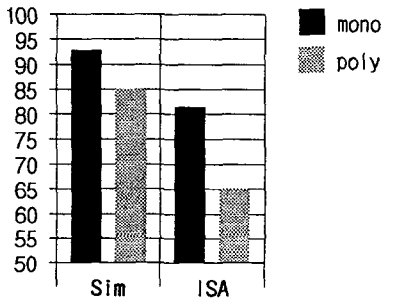


그림 4 : 의미 수에 따른 정확도 (Sim, ISA)

실험 결과를 보면 상하위 관계를 이용한 방법인 경우, 정확도가 매우 낮으므로 한국어 단어의 의미 수가 2 이상인 경우에는 적용하지 않는 것이 좋음을 알 수 있다.

이번에는 앞의 실험에 나온 결과를 이용하여 상하위 관계를 이용한 방법에서 한국어 단어의 의미 수가 1인 경우(monosemous)에만 적용하고, 임계값을 0.4, 0.7, 0.85로 변화시켜 실험을 해보았다. 임계값이 0.4인 경우, 정확도가 81.4% 이었고, 임계값이 0.7인 경우에는 86.1%, 임계값이 0.85인 경우에는 86.7%의 정확도가 나왔다.

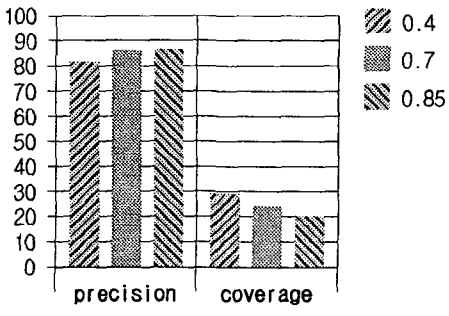


그림 5 : 임계값에 따른 정확도 (ISA)

결과를 보면 임계값이 0.85일 때 가장 높은 정확도를 보임을 알 수 있다. 그러나 임계값이 높을수록 정확도는 높아지지만 적용율이 낮아지기 때문에 정확도와 적용율을 고려하여 임계값을 정해야 한다.

지금까지의 결과를 보면 유사도 행렬을 이용한 방법이 상하위 관계를 이용한 방법보다 높은 정확도와 적용율을 보인다. 마지막 실험에서는 두 방법을 조합하여 사용하기 위해서 먼저 유사도 행렬을 이용한 방법을 적용시키고, 실패할 경우에 상하위 관계를 이용한 방법을 적용하도록 하였다. 그리고 상하위 관계를 이용한 방법에서는 한국어 단어의 의미 수가 1인 경우에만 적용시키며, 임계값은 0.7을 사용하였다.

실험 결과, 총 30,656 개의 한국어 단어의 의미 중에서 21,157 개의 의미를 WordNet의 synset에 연결시켜 적용율이 69.0%가 되었다. 이 결과에서 무작위로 추출하여 정확도를 측정한 결과 90.2%의 정확도가 나왔다.

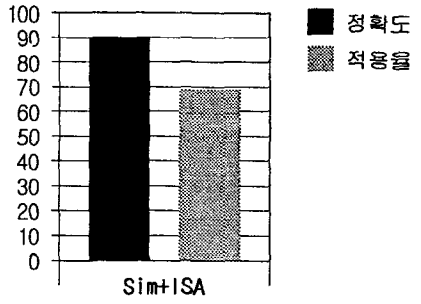


그림 6 : 최종 결과 (Sim+ISA)

마지막 실험을 통해서 18,362개의 단어와 21,390개의 의미로 이루어지고 정확도가 90.2%인 한국어 명사 개념체계를 완전 자동으로 구축하였다.

## 4 결론

본 논문에서는 Princeton 대학의 WordNet과 한영사전, 그리고 국어사전을 이용하여 한국어 개념체계를 자동으로 구축함으로써, 이미 구축되어진 다른 언어의 개념체계를 이용하여 새로운 언어의 개념체계를 자동으로 구축할 수 있음을 보

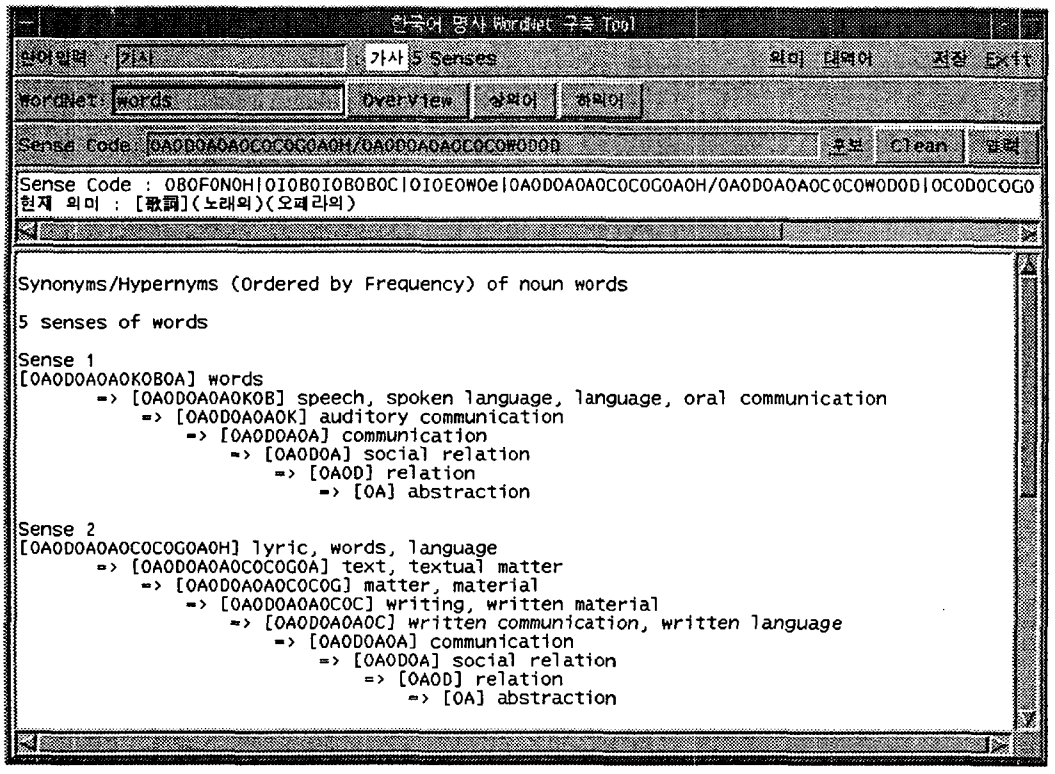


**그림 7 : 한국어 개념체계 반자동 구축 tool**

였다. 실험 결과 자동으로 구축되어진 한국어 개념체계는 총 21,390개의 한국어 단어 의미와 18,362개의 단어로 이루어져 있으며, 90.2%의 정확도를 보인다. 본 논문에서 사용한 방법은 특정 언어에 제한되지 않는 매우 일반적인 방법이다. 따라서 이 방법을 다른 언어에도 쉽게 적용할 수 있으며, 이미 이와 비슷한 방법으로 EuroWordNet의 일부 언어에서 자국의 WordNet을 구축하고 있다[4][5].

본 논문에서 사용한 방법은 적용율이 69%이므로, 나머지 31%는 자동으로 구축할 수 없다. 이를 위해서 다른 자동화 방법을 적용하거나, 다양한 방법을 조합하는 것도 가능할 것이다. 그러나 자동화 기법으로는 해결하지 못하는 부분이 존재하므로, 자동으로 구축하지 못하는 단어의 의미에 대해서 사람의 개입을 최소화하여 영어 대역어의 올바른 의미를 선택할 수 있는 도구를 개발하였다(그림 7). 이 도구로 자동으로 구축되어진 한국어 개념체계를 검증 및 수정을 할 수 있고, 자동으로 구축하지 못하는 의미들을 구축할 수 있게 된다.

## 참고문헌

[1] Dekang Lin. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1997

[2] Rigau G. and Atserias J. and Agirre E. Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1997

[3] Rigau G. and Rodriguez H. and Agirre E. Building Accurate Semantic Taxonomies from Monolingual MRDs. In Proceedings of the 36th Conference of the Association for Computational Linguistics (COLING-ACL'98), 1998

[4] Atserias J., Climent S., Farreras J., Rigau G. and Rodriguez H. Combining Multiple

Methods for the Automatic Construction of Multilingual WordNets. In proceeding of the Conference on Recent Advances on NLP, 1997

[5] Benitez L., Cervell S., Escudero G., Lopez M., Rigau G. and Taule M. Methods and Tools for building the Catalan WordNet. In workshop Language Resources for European Minority Languages at LREC'98, 1998

[6] 조평옥. 한국어 명사의 의미 계층 구조 구축. 울산대학교 교육대학원 석사학위논문, 1996

[7] 문유진. 한국어 명사를 위한 WordNet의 설계와 구현. 정보과학회논문지(c) 제2권 제4호, 1996

[8] 이승우. 포항공과대학교 대학원 석사학위논문, 1998