기계가독사전을 이용한 한국어 시소러스 구축

이주호[0] 은광희 최기선
한국과학기술원 전자전산학과, 전문용어언어공학연구센터
{mywork, koaunghi, kschoi}@world.kaist.ac.kr

Construction of Korean Thesaurus Using Machine Readable Dictionary

Juho Lee[0] Koaunghi Un Key-Sun Choi
KOTERM, Dept. of EE CS, Korea Advanced Institute of Science and Technology

요 약

시소러스는 자연언어처리의 여러 분야에서 이용 가능한 아주 유용한 정보이다. 본 논문에서는 기존의 구축된 시소러스를 기반으로 우리말큰사전을 이용하여 한국어 명사 시소러스를 반자동으로 구축하는 과정을 소개한다. 우선 코퍼스의 고빈도어를 중심으로 사전에서 추출한 기본명사들의 각 의미에 1차로 의미번호 부착 후 그 결과를 이용하여 사전 정의문으로 각 의미별 클러스터를 구성했다. 그리고, 전단계에서 의미번호를 붙이지 못한 명사의 의미에 대하여 그 정의문과 클러스터들 간의 유사도를 계산하여 가장 유사한 의미번호를 후보로 제시하였다. 마지막으로 사전의 하이퍼링크를 사용하여 아직 의미번호가 붙지 않는 명사의 의미에 의미번호를 부여했다. 각 단계에서는 사람의 후처리를 통해서 시소러스의 정확도를 높였다.

1. 서론

시소러스는 자연언어처리의 여러 분야에서 이용 가능한 아주 유용한 정보이다. 특히 단어의미구분이나 정보검색, 기계번역 등에서 의미를 다루는 데 있어서 시소러스는 지식 베이스로써 큰 역할을 한다. 영어에 대해서는 기존의 Roget의 시소러스나 WordNet이 이러한 목적으로 널리 사용되고 있으며, 일본어에서는 NTT 시소러스가 실제 기계번역 시스템에 이용되었다 [8][9][10].

하지만 이러한 시소러스를 처음부터 수동으로 구축하는 일은 많은 시간과 노력을 필요로 한다. 본 논문에서는 기계가독사전을 이용하여 한국어 명사 시소러스를 비교적 쉽게 구축하는 방법론을 제시하고자 한다. 기계가독사전의 명사의 각 의미에 기존의 구축된 개념체계의 의미번호를 부여함으로써 이 두 지식을 하나로 합칠 수 있다. 의미번호 부여 작업은 일단 자동으로 부여하고, 사람이 후처리를 거쳐서 수정함으로써 비교적 적은 노력을 들이면서 정확한 시소러스를 구축할 수 있다. 여기에서는 NTT 시소러스를 기반으로 우리말큰사전을 이용하여 한국어 명사 시소러스를 반자동으로 구축하였다.

본 논문의 구성은 다음과 같다. 2절에서는 시소러스에 대한 관련연구를 살펴본다. 3절에서는 시소러스 구축의 기본이 되는 NTT 시소러스에 대해서 간단하게 알아보고, 4절에서 우리말큰사전을 이용하여 시소러스를 구축하는 과정을 설명한다. 그리고, 5절에서는 구축작업에 사용된 도구를 소개하고, 마지막으로 6절에서 결론을 맺는다.

2. 관련연구

한국어 시소러스 구축에 대한 연구 다음과 같은 것들이 있었다.

[2]에서는 초등학교 국어용 어휘집의 5,000단어를 대상으로 국어사전에서 자동으로 상의어 정보를 추출하고, 이를 자동 번역한 영어 WordNet과 합성하여 한국어 명사 WordNet 예비 리스트를 구축했다. 상의어 사전과 번역된 영어 WordNet을 합성할 때는 전산학자와 언어학자에 의해 수작업으로 pruning 작업이 이루어졌다. 또한 이렇게 만들어진 예비 리스트를 기반으로 분야가 명시된 20,000개 정도의 명사를 추가로 입력하여 한국어 명사 WordNet 리스트를 구축했다. 여기에서 pruning 작업은 일관성있고 신중하게 수행되어야 하며 무척 까다롭고 시간이 많이 걸리는 수작업 단계로서, 상식과 언어학자의 어휘개념을 참조하여

수행되었다.

[3],[4]에서는 영어 WordNet을 기반으로 하여 한영사전과 국어사전을 이용하여 한국어 명사의 개념체계를 자동으로 구축하였다. 여기에서는 한영사전과 국어사전으로부터 뽑아낸 한국어 일부의 의미를 다양한 WSD (Word Sense Disambiguation) 방법을 적용시켜 WordNet의 synset에 자동으로 연결시켰다.

본 논문에서는 우리말큰사전의 명사의 각 의미에 NTT 시소러스의 의미번호를 자동으로 부여함으로써 한국어 명사 시소러스를 구축한다. 사람의 후처리 과정을 거침으로써 좀 더 정확한 시소러스를 구축하도록 하고, 정확한 의미번호 후보 부여와 작업지원도구를 사용함으로써 후처리 과정의 번거로움을 최대한 줄인다.

3. NTT 시소러스

일본전신전화주식회사(NTT)가 10년 이상에 걸쳐 연구 개발해 온 일영 기계번역시스템 ALT-J/E (Automatic Language Translator-Japanese to English)의 번역사전 중 일본어 의미사전에 관한 부분을 정리하여 만든 것이 '일본어 어휘 대계'이다[8]. '일본어 어휘 대계'는 크게 '단어의미 속성체계', '단어의미사전', '구문의미사전'으로 구성되어 있고, 이중에서 '단어의미 속성체계'를 NTT 시소러스라 지칭한다.

NTT 시소러스는 그 하위로 세가지 의미 체계를 가지고 있다. 이는 일반명사 의미속성(12단계, 2,710노드), 고유명사 의미속성(9단계, 130노드), 용언 의미속성(36노드) 등 모두 3,000 여 노드로 의미가 분류되어 있으며 각 노드에 속해있는 모든 단어의 수는 약 40만 단어이다. 각 노드는 계층적인 구조를 가지며 상하의 단어의미속성이 is-a 관계 혹은 has-a 관계를 가지며 한 단어는 여러 의미 속성에 동시에 포함될 수 있다.

본 논문에서는 NTT 시소러스 중에서 일반명사 의미속성체계만을 사용한다. 이후 본 논문에서 NTT 시소러스는 일반명사 의미속성체계를 가리키는 말로 사용하였다. 그림 1에서 NTT 시소러스의 최상위 노드 부분을 대략적으로 나타냈다.

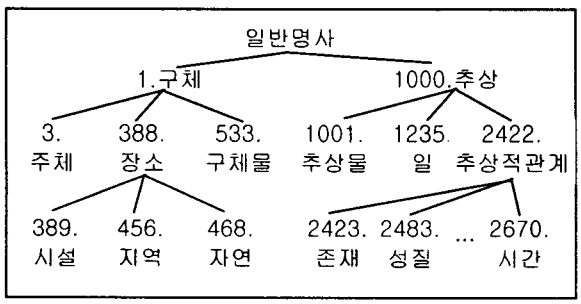

그림 1. NTT 시소러스의 최상위 노드 부분

4. 한국어 시소러스 구축

기계가독사전과 NTT 시소러스를 기반으로 한국어 명사 시소러스를 구축하는 전체 과정은 아래의 그림 2와 같다.

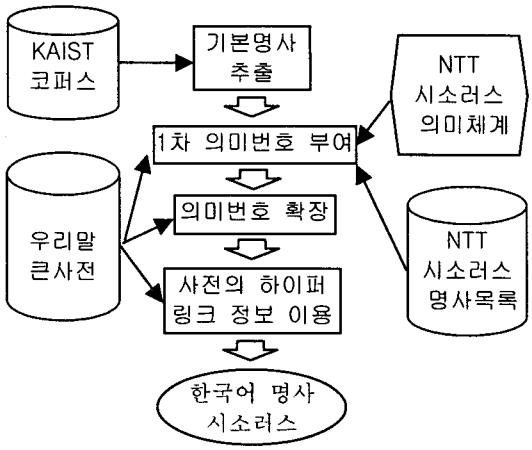

그림 2. 전체 구축 과정

먼저 시소러스의 각 의미노드에 들어갈 기본명사 목록을 사전과 코퍼스로부터 추출하도록 한다. 첫번째 단계에서는 기본명사목록을 NTT 시소러스의 명사 목록과 비교하여 각 기본명사에 해당하는 시소러스 의미번호 후보를 부여하고, 후처리를 통해서 이들 후보들 중에서 각 명사 의미에 대한 정확한 의미번호를 선택한다. 두번째 의미번호 확장 단계에서는 첫번째 방법으로 의미노드를 붙일 수 없는 것들에 대해서 국어사전의 정의문과 전단계에서 구축한 정보를 사용하여 자동으로 의미번호를 후보를 제시하고, 수작업을 통해서 올바른 후보를 선택하도록 한다. 마지막 단계에서는 아직 의미번호가 붙어있지 않는 기본명사에 대해서 국어사전의 동의어, 준말 같은 하이퍼 링크 정보와 전단계의 결과를 이용하여 의미번호가 부착 가능한 경우에 이를 수행하도록 한다. 이렇게 하여 기본명사의 각 의미에 시소러스의 의미번호를 부여함으로써 한국어 명사 시소러스를 구축할 수 있다. 그럼 각 중간 과정에 대해서 좀 더 자세히 알아보자.

4.1 기본명사 추출

시소러스를 추출하기 위한 기본명사목록은 국어사전을 기반으로 하여 추출하도록 한다. 시소러스가 여러 자연언어처리 응용시스템에서 유용하게 사용되기 위해서는 시소러스를 구성하는 어휘들이 실제로 자주 사용되는 어휘들로 되어 있는 것이 좋을 것이다. 여기에서는 코퍼스의 빈도수를 기본으로 삼아 시소러스 구축에 사용될 기본명사목록을 추출했다. 코퍼스를 통해 조사한 고빈도어를 대상으로 하여 국어사전에서 명사와 그 명사에 해당하는 의미를 추출한다.

본 논문에서는 우리말큰사전을 기본으로 하여 기본명사목록을 추출했다. 또한 빈도수 계산을 위해서 약 천만 어절 수준의 KAIST 품사부착 코퍼스를 이용하였다. 이 코퍼스에서 형태소 빈도를 조사하여 빈도가 7이상인 "형태소/품사" 쌍을 모두 추출하였다. 이는 코퍼스에서 전체 "형태소/품사" 쌍의 약 97.9% 에 해당하는 수준이다. 여기에서 명사만을 따로 추출하고, 우리말 큰사전과 비교하여 사전에 속한 단어로 이루어진 최종 기본명사목록을 만들었다. 기본명사목록에 대한 통계값은 아래 표와 같다.

항목	값
전체 유일한 명사수	25,368개
전체 의미수	69,242개
한 명사 당 평균 의미수	2.73개

표 1. 기본명사목록에 대한 통계값

시소러스를 구축하는 작업은 위에서 추출한 모든 기본명사의 의미에 시소러스의 의미번호를 부착하는 작업으로 볼 수 있다. 즉, 앞으로의 작업은 69,242개의 각 의미에 NTT 시소러스의 전체 의미분류 2,710개의 의미번호 중 하나를 부착하는 것이고, 바꾸어 말하면 2,710개의 의미노드에 69,242개의 의미를 부여하여 시소러스를 구축하는 것이다.

4.2 1차 의미번호 부착

이 단계에서 우선 기본명사목록과 NTT 시소러스의 명사목록을 비교하여 각 명사에 대응하는 의미번호 리스트를 얻는다. NTT 시소러스의 명사목록은 대략적으로 한국어로 번역되어 있는 상태이고, 이미 NTT 시소러스의 명사목록에는 의미번호가 부착되어 있기 때문에 이것과 매칭시키면 우리가 추출한 기본명사에 의미번호를 붙일 수 있다.

우리가 추출한 기본명사목록을 NTT 시소러스의 명사목록과 매칭시킬 때 검색이 안되는 명사에 대해서는 검색될 때까지 그 상의어를 따라 올라가면서 검색을 한다.

명사의 상의어는 사전의 정의문에서 자동으로 추출할 수 있다[1][5]. 이 논문에서는 다음의 규칙에 따라 자동으로 상의어를 추출하였다.

- 사전 정의문의 첫번째 문장에서 가장 뒷부분에 오는 명사가 상의어
- 'A의 하나', 'A의 일부', 'A의 한 갈래' 등의 패턴일 때는 A가 상의어
- '것', '일', '함' 등 너무 범위가 넓거나 의미없는 명사는 상의어에서 제외

아래의 그림 3은 '국립묘지'에 대한 자동 의미부착의 예이다. 여기에서 보면 '국립묘지'가 NTT 시소러스 명사목록에 존재하지 않기 때문에 그것의 상의어 '공동묘지'로 다시 검색을 하였고, 이에 실패하여 그것의 상의어인 '묘지'로 검색하여 결과를 얻어냈음을 알 수 있다.

```
상의어 검색: 국립묘지->공동묘지->묘지
  국립묘지: 나라에서, 순국한 사람들의 유해나
    영령을 모시는 공동묘지.
  공동묘지: 일정한 곳에 공동으로 쓰게 된
    묘지.

검색 결과: 456[묘지]460[지역(인간활동)]
```

그림 3. '국립묘지'에 대한 자동의미부착의 예

하지만 이러한 과정들에는 의미구분이 포함되지 않는다. 따라서 기본명사의 의미 하나하나에 올바른 의미번호를 부착하기 위해서는 의미구분이 필요하다. 여기에서는 위의 과정을 거쳐 자동으로 제시한 의미번호를 대상으로 다음과 같은 원칙을 세워서 수동으로 의미구분을 하고, 올바른 의미에 해당 의미번호를 부착하였다.

- 우리말 큰사전의 정의문에 기초하여 의미번호를 부착
- 상하위 관계가 있는 의미번호가 동시에 관계있는 경우에는 하위 의미를 선택
- 한 의미에 하나 이상의 의미번호 부착 가능

위의 작업은 정확도를 높이기 위해서 두 사람이 각각 수행한 후에 그 두 결과를 비교하여 일치하지 않는 부분은 제3자가 다시 검토하도록 하였다.

이 과정을 통해서 전부 19,663개의 명사에 대한 29,637개의 의미에 의미번호를 부착하였다.

4.3 사전의 정의문을 이용한 의미번호 확장

전단계에서는 기본단어와 상의어 정보만을 가지고 NTT 시소러스 명사목록에서 찾아 의미번호를 부착하였다. 하지만 NTT 시소러스 명사목록에 속하지 않는 명사가 기본 명사에 있을 수 있고, NTT 시소러스 명사목록을 번역하는 과정에서 오류가 있을 수 있다. 따라서 전단계를 거친 결과에는 듬성듬성 의미번호가 부착되지 않은 의미가 존재한다. 이번 단계에서는 전단계의 결과와 사전의 정의문을 이용하여 나머지 의미번호가 부착되지 않은 부분에 대해서 확장한다.

4.3.1 접근 방법

'사전에서 같은 의미분류에 속하는 의미들은 비슷한 어휘들로 기술된다'고 가정하고 Chen[7]이 사용한 방법을 기초로 하여 접근했다.

이미 의미번호가 붙어있는 의미에 대하여 같은 의미번호가 붙은 의미의 사전 정의부끼리 클러스터링하

여 각 의미번호당 하나의 문서로 취급하면 의미번호를 붙이는 일을 일반적인 정보검색과 비슷하게 취급할 수 있다. 즉, 아직 의미번호가 붙어있지 않은 새로운 의미가 들어오면 그것의 정의부를 검색어로 하여 앞서 같은 의미번호끼리 클러스터링한 문서를 대상으로 검색해서 가장 유사한 문서가 얻어지면, 그 문서에 해당하는 의미번호를 새 의미에 부착하면 된다.

전단계에서 일부 의미에 대하여 의미번호를 부여했으므로 이를 기반으로 이 방법을 적용시키면 나머지 의미번호를 붙이지 않은 의미에 대하여 의미번호를 부여할 수 있을 것이다.

4.3.2 알고리즘 설명

새로운 정의문에 대하여 기존의 사전 정의문 클러스터중에 가장 유사한 것을 찾기 위해서 이 둘 간에 유사도를 구해야 한다. 여기에서는 유사도를 계산할 때 간단하게 tf (term frequency), idf (inverted document frequency)를 사용하고, 추가로 새로운 정의문에서 상의어에 대하여 따로 가중치를 주었다.

새로운 정의문 Q에 대한 클러스터 C_i의 유사도는 아래의 식으로 계산한다.

$$\sum_{t_j \in Q} g(t_j) \times tf_{ij} \times idf_j$$

$$idf_j = \log(\frac{N}{df_j})$$

t_j: 새로운 정의문 Q에 속하는 내용어
$g(x)$: 상의어에 대해 가중치를 주는 함수
　　x가 상의어이면 w, 그렇지 않으면 1
tf_{ij}: t_j가 C_i에 나타난 빈도
N: 클러스터의 수, 의미분류 수와 일치
df_j: t_j가 나타난 클러스터의 수

정의문에서 조사, 어미 등의 기능어는 제외하고 내용어만 고려하기 때문에 정의문을 품사태거를 사용하여 품사를 부착하는 과정이 필요하다.

위의 식을 이용하여 유사도를 계산한 후 가장 유사도가 높은 클러스터를 찾으면 그 클러스터에 해당하는 의미분류에 새 의미가 속한다고 볼 수 있다.

4.3.3 실험 및 평가

전단계에서 의미번호를 부착한 29,637개의 의미에 해당하는 우리말큰사전의 정의문을 2,710개의 의미별로 각각 모아서 클러스터를 만들었다. 또한 상의어에 대한 가중치 w를 2로 하여 실험했고, 4.2에서 제시한 상의어 추출 방법을 사용하여 가중치를 계산하는 단어가 상의어인지 검사하였다. 실험데이터로 아직까지 의미번호가 붙어있지 않은 150개의 의미를 임의로 추출하여 사용했다.

우선 유사도가 높은 상위 n개 안에 정답이 포함되는 의미개수의 변화를 나타낸 그래프는 그림 4와 같다. 그래프에서 보면 n이 증가하면서 처음에는 정답이 포함된 의미개수가 급격하게 증가하다가 차츰차츰 완만한 곡선을 그리는 것을 알 수 있다. 150개 중에서 25개의 의미에 대해서는 의미번호 후보를 전혀 찾을 수 없었고, 정답을 찾을 것 중에서 정답의 순위가 가장 낮은 경우에는 278번째 후보가 정답인 경우였다.

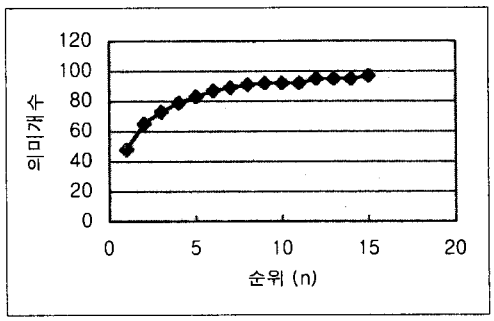

그림 4. 상위 n개까지 정답이 포함된 의미개수의 변화

의미번호 후보를 낮은 순위까지 많이 제시할수록 재현율을 높힐 수 있지만 그만큼 후처리 과정의 작업속도가 떨어질 수 있다. 여기에서는 이를 고려하여 10개까지의 의미번호 후보를 제시하도록 했다. 10개까지 후보를 제시할 때 재현율은 61.3%이다.

이 단계의 방법을 적용하여 전단계에서 의미번호를 붙이지 못한 각 의미에 대하여 10개씩 의미번호 후보를 유사도 순으로 제시한 후 사람이 후처리로 올바른 결과를 선택하여 모두 9,006개의 명사에 대한 19,263개의 의미에 추가로 의미번호를 붙일 수 있었다. 이는 첫번째 결과와 비교해 보면 한 명사당 평균 의미수가 더 높다. 즉, 두번째 단계의 방법이 더 적용성이 높음을 알 수 있다.

4.4 사전의 하이퍼 링크 정보 이용

이 단계에서는 이전단계까지 의미번호를 부착하지 못한 의미에 대하여 국어사전의 하이퍼 링크 정보가 이용가능한 것들은, 하이퍼 링크 정보와 이전 단계의 결과를 이용하여 의미번호를 부착한다. 여기에서 이용 가능한 국어 사전의 하이퍼 링크 정보는 동의어, 높임말, 낮춤말, 준말 등의 정보이다. 이런 관계에 있는 두 의미는 서로 같은 의미분류에 속한다고 볼 수 있으므로 어느 한쪽의 의미번호만을 아는 경우 나머지 한쪽에 의미번호를 부착할 수 있을 것이다.

이 단계에서는 모두 4,658개의 명사에 대한 7,623개의 의미에 이 방법을 적용할 수 있었다.

5. 작업지원 도구

4.2 ~ 4.4의 작업내용을 보면 사람이 작업하는 후처리 과정들도 중요한 역할을 하고 있는 것을 알 수 있다. 우리는 이런 사람들의 작업에 도움을 주고자 국어사전과 구축한 시소러스를 손쉽게 볼 수 있는 작업지원 도구를 만들었다.

그림 5는 작업지원 도구의 화면이다. 이 도구는 크게 네 부분으로 구성된다. 화면 좌측상단은 검색어를 입력하는 부분, 우측상단은 구축한 시소러스를 탐색하는 부분, 좌측하단은 우리말큰사전을 탐색하는 부분, 우측하단은 시소러스의 의미체계를 탐색하는 부분이다. 각각에 대하여 간단하게 설명을 하도록 하겠다.

좌측상단은 검색어를 입력하는 부분으로서, 의미분류명, 의미번호, 명사로 검색이 가능하다. 이 부분에서는 또한 우리말큰사전의 각 의미에 시소러스 의미번호를 부착한 표 형태의 검색결과를 표시한다.

우측상단은 구축한 시소러스를 직접 탐색하는 부분이다. 가운데 중심 노드는 현재 탐색중인 의미분류를 나타내고, 그 위쪽은 상의분류, 아래쪽은 하의분류, 왼쪽은 형제분류, 오른쪽은 현재 의미분류에 속한 의미들을 나열하고 있다. 각 항목을 클릭함으로써 시소러스를 계속 탐색하는 것이 가능하다.

좌측하단은 우리말큰사전 탐색기로서, 사전의 내용을 표시해준다. 사전의 풀이와 구축한 시소러스를 비교하면서 시소러스가 잘 구축되었는지 쉽게 확인할 수 있다.

우측하단은 시소러스의 의미체계를 계층적으로 보여줌으로써 의미체계의 전체적인 모습을 볼 수 있게 한다. 각 항목을 클릭함으로써 그 하위 분류를 보이게 또는 숨기게 할 수 있다.

이 네 가지 부분은 독립적인 것이 아니라 모두 연동이 된다. 즉, 한쪽에서 탐색을 하면 나머지 부분에서 그것과 관련된 항목을 보여준다.

이 작업 도구는 사용자가 쉽게 웹 브라우저를 통해서 사용할 수 있고, 실제로 이 작업도구를 사용함으로써 여러 작업 부분에 많은 도움을 얻을 수 있었다.

6. 결론

본 논문에서는 기계가독사전과 기존의 의미체계를 이용하여 비교적 쉽게 한국어 명사 시소러스를 구축하는 방법론을 제시했다. 코퍼스의 고빈도어를 중심으로 사전에서 추출한 기본명사들의 각 의미에 1차 의미번호 부착 후 그 결과를 가지고 사전 정의문으로 클러스터를 구성하여, 전단계에서 의미번호를 붙이지 못한 명사의 의미에 대하여 그 정의문과 클러스터들 간의 유사도를 계산하여 가장 유사한 의미번호를 후보로 제시하였다. 마지막으로 사전의 하이퍼링크를 사용하여 아직 의미번호가 붙지 않는 명사의 의미에 의미번호를 부여했다. 각 단계에서 사람의 후처리를 통해서 시소러스의 정확도를 높였다.

본 논문에서는 우리말 큰사전과 NTT 시소러스를 이용하여 23,823개의 명사와 이에 대한 56,523개의 의미를 포함하는 2,710개의 계층적 의미분류로 이루어진 한국어 명사 시소러스를 구축했다.

향후과제로써 기본명사의 의미중 아직 의미번호를 붙이지 못한 의미에 대해서 의미번호를 부여함으로써 보다 완전한 시소러스를 구축해야 할 것이다. 또한 만들어진 시소러스를 단어의미구분이나 구문분석, 정보검색 등 실제 응용에 적용해 보고, 계속 수정하여 다듬는 과정이 필요하다.

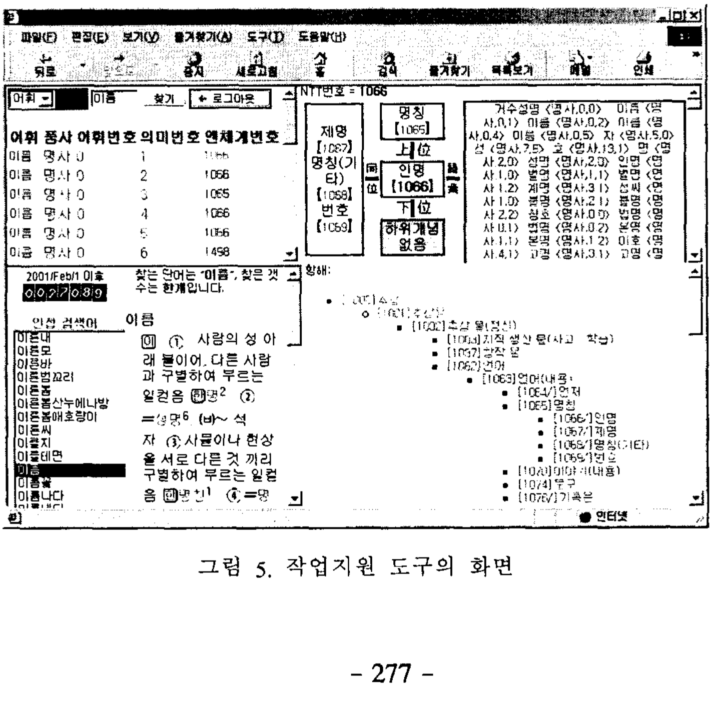

그림 5. 작업지원 도구의 화면

7. 참고 문헌

[1] 문유진, 김영택. 한국어 명사의 hypernym 자동 추출 방법. 한국정보과학회 '94 가을 학술발표논문집 A. 제 21권 2호, 613-616, 1994.

[2] 문유진. 한국어 명사를 위한 WordNet의 설계와 구현. 정보과학회논문지(c) 제2권 제4호, 437-445, 1996.

[3] 이창기, 이근배. WordNet을 이용한 한국어 시소러스 자동 구축. 제11회 한글 및 한국어 정보처리 학술대회 논문집, 156-163, 1999.

[4] 이창기, 이근배. 의미 애매성 해소를 이용한 WordNet 자동 매핑. 제12회 한글 및 한국어 정보처리 학술대회 논문집, 262-268, 2000.

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

[6] 우리말큰사전. 어문각, 1997.

[7] Chen, Jen Nan and Jason S. Chang. Topical Clustering of MRD Senses Based on Information Retrieval Techniques. Computational Linguistics Volume 24, Number 1.

[8] Ikehara, S. *et al.* The Semantic System, volume 1 of Goi-Taikei -- A Japanese Lexicon. Iwanami Shoten. 1997.

[9] Miller G. Wordnet: An on-line lexical database. International Journal of Lexicography, 1990

[10] Roget's Thesaurus of English Words and Phrases. 1987. Longman Group UK Limited, London.