한국어 명사의 시소러스 구축을 위한 시스템 설계 및 구현

이 종 인† · 한 광 록†† · 양 승 현††† · 김 영 섬†††

요 약

본 논문에서는 한국어 명사의 의미 개념의 계층을 생성하기 위한 시소러스 구성 방법과 시소러스를 구축하기 위한 개발 시스템을 구현하였다. 기존의 시소러스 구축에 있어서 나타나는 계층 설정의 비객관성 및 작업속도 문제, 비구조성, 비일관성 등의 문제를 해결하기 위하여 상향식과 하향식 방법을 혼합 적용하는 다단계 구축 방법을 사용한다. 온라인 전자 사전의 뜻풀이 문을 이용하여 객관성을 유지하고, 기존 시소러스의 기본 모델을 참조하여 비구조성과 비일관성의 문제를 해결한다. 또한 방대한 양의 표제어를 포함하는 시소러스를 빠른 시간 내에 구축하기 위하여 클라이언트/서버 환경의 개발 도구를 구현하여 여러 사람이 다중 입력 작업을 할 수 있도록 하였다.

Design and Implementation of a System for Constructing Thesaurus of Korean Nouns

Jong-In Lee† · Kwang-Rok Han†† · Seung-Hyun Yang††† · Young-Sum Kim†††

ABSTRACT

We present a method of thesaurus construction in order to produce semantic concept hierarchy of Korean nouns and implement a system for constructing the thesaurus in this paper. Multiple-step construction method is applied to this system which uses bottom-up and top-down method complementarily for solving the nonobjectivity of word hierarchy, working speed, structural contradiction, and incoherency of existing thesaurus. This system maintains objectivity using the meaning sentence of machine-readable dictionary and solves structural contradiction and incoherency with reference to existing thesaurus. We implement a developmental tool based on client/server system to construct thesaurus including massive entries as soon as possible and multiple clients can work simultaneously.

1. 서 론

한국어 단어의 의미 영역 정보는 자연어 처리 시스템에 있어서 형태소 분석이나 통사 분석으로는 처리되지 않는 중의성의 문제를 해결하기 위해 적용되는 의미 분석 분야와 인터넷의 등장으로 그 중요성이 증대되고 있는 정보 검색 분야에서 사용되는 정보로 전체 시스템의 성능을 결정 지을 수 있는 중요한 정보이기 때문에 이를 구축하기 위한 다양한 연구들이 이루어져 왔다.[1][4][9] 1990년 미국의 Princeton대학에서 제작된 WordNet이 대표적인 시스템으로 단어의미, 상위개념, 구성개념, 반대어, 연결어 정보를 포함하여 단어의 의미영역 정보를 제공하고 있다.[1] 그러나 WordNet을 위시한 초기 시스템에서는 전체 시소러스가 연구자들의 수작업에 의해 구축이 되었기 때문에 시소러스 구축에

† 준 회 원 : 호서대학교 컴퓨터학과
†† 종신회원 : 호서대학교 컴퓨터공학과 교수
††† 정 회 원 : 한국전자통신연구원 선임연구원
논문접수 : 1998년 8월 6일, 심사완료 : 1998년 11월 12일

많은 시간이 소모되고 연구자의 주관이 과도하게 개입된다는 문제를 안고 있었다. 그리고 WordNet등 기존에 구축된 시소러스를 한국어 분석 및 정보 검색에 적용하려는 시도가 있었지만 실제로 한국어 처리에 적용하기에는 한계가 있다.[3][11]

최근에 이러한 문제를 해결하기 위해 제시되고 있는 것이 전자 사전을 이용한 상위어 추출 방법이다. 전자 사전을 이용하는 방법은 사전의 뜻풀이문을 이용하여 상위어를 추출함으로 객관성을 유지할 수 있고 작업 속도를 높일 수 있다는 장점이 있으나 사전의 뜻풀이문이 시소러스 구축에 적합한 구조로 되어 있지 않기 때문에 전체 시소러스가 구조적이지 못하고 단어의 일관성을 잃어 버리기 쉽다는 문제점을 안고 있다.[10]

따라서 본 논문에서는 전자 사전의 뜻풀이문에서 상위어를 추출하는 방식을 택하되 시소러스의 구조성과 일관성을 유지하기 위해 다단계로 나누어 구축하는 방법을 제안한다.

2. 기존 방식에 대한 고찰

기존의 방식을 크게 두 가지로 분류하면 중심 단어들에 새로운 하위 단어들을 추가하는 하향식과 단어의 상위어를 추적함으로써 시소러스를 완성하는 상향식으로 나눌 수 있다.

하향식 방식의 대표적인 예로는 미국의 WordNet과 일본 EDR(Electronic Dictionary Research)에서 제작한 "개념사전"이 있다.[2][5] 개념사전의 제작 과정을 살펴보면 첫 단계에서 연구자의 연구에 의해 20개의 대분류노드를 설정 후 이 노드들과 연관된 단어들을 연쇄적으로 추가/확장 시키고 있다.[7] 이 방식은 초기 대분류 노드에서부터 문제를 노출시키고 있다. 대분류 노드로 선택된 20개의 노드가 연구자의 주관에 의해 선택되어 졌기 때문에 객관성을 가지지 못하고 있다. 게다가 모든 작업이 수작업이기 때문에 이 20개의 노드를 선택하기 위해 연구자는 많은 시간을 투자하고, 각 노드들에 관련된 노드들을 찾기 위해 더 많은 시간을 소비해야 한다는 문제이다.

상향식은 전자 사전의 뜻풀이문에서 상위어를 추출하여 시소러스를 구축하는 방식으로 하위어로부터 시작하여 상위어를 추가하고 있다. 울산 대학교 조평옥씨의 SHKN(Semantic Hierarchy of Korean Nouns)이 대표적인 예로 일정 단어들을 선택하여 동시에 상위어들을 추적하고 있다. 이 방식은 전자 사전을 이용함으로 객관성을 유지할 수 있고 작업 속도를 높일 수 있지만 다음과 같은 구조적 문제와 비일관성의 문제를 안고 있다.

첫번째로 비일관성 문제를 들 수 있다. 두 단어가 같은 부류의 단어일 때 같은 부모 노드나 영역에 속해야 함에도 불구하고 뜻풀이문에 차이로 인해 전혀 다른 영역에 속해 버리는 문제점이다. 예를 들면 "중학교"와 "대학"에서 "중학교"는 "~하는 곳"이라는 뜻풀이문이 있어 장소를 나타내는 "곳"의 영역에 속하지만 "대학"이 "~하는 기관"이라는 뜻풀이문을 가짐으로써 "중학교"와는 다른 영역인 "기관"의 영역에 속해 버리는 문제가 나타나고 있다.

두 번째로 순환의 문제를 들 수 있다. 즉 두 단어의 뜻풀이문이 서로 맞물려서 더 이상 상위로 올라가지 못하는 경우로 "동물"과 "짐승"이 대표적이 예이다. "동물"은 "인간 이외의 짐승"이라는 뜻풀이문이 있고 "짐승"은 "~하는 동물"이라는 뜻풀이문이 있어 "동물"은 "짐승"을 상위로 삼고 "짐승"은 "동물"을 상위로 추출하여 더 이상 상위로 이동하지 못하고 두 단어 사이만을 순환하게 된다.

세 번째로 비적합성의 문제를 안고 있다. "요리"를 예로 들면 "요리"가 의미상 "음식물"과 "행위"의 의미를 모두 가지고 있기 때문에 "카레"라는 단어에 대해 "음식물"이라는 구체물과 "행위"라는 구체물과 반대되는 추상물의 의미를 동시에 도출해 내고 있다. 즉, 서로 상반되는 속성을 한 단어가 소유하게 되는 것이다. 이는 비단 상향식뿐만 아니라 하향식에서도 문제가 되고 있는 성질이다.[6][17]

마지막으로 상위어 추출이 불가능한 경우가 발생할 수 있다는 문제이다. 예를 들어 "축하"라는 단어의 뜻풀이문은 "잘했다고 하는 것"이라고 되어있다. 이 경우 어떤 것을 상위어로 추출해야 하는지 애매하게 된다.

이상의 두 방식을 살펴 보면 상호 보완적인 관계임을 알 수 있다. 즉 하향식이 구조성이나 일관성을 유지할 수 있는 반면 상향식은 객관성을 유지할 수 있고 작업속도를 높일 수 있다는 장점이 있다.

3. 다단계 시스템 제안

다단계 시스템은 기존의 두 방식을 절충하여 양 방식의 단점은 최소화하고 장점을 극대화하려는 시스

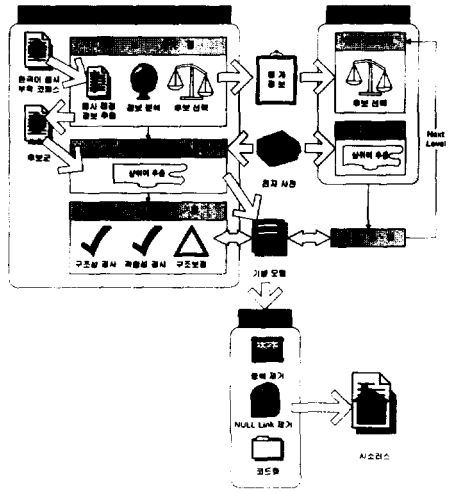
템이다. 본 시스템에서는 하향식을 기본으로 하여 사전의 뜻풀이문에서 상위어를 추출하고 시소러스의 구축하는 과정을 몇 가지 단계로 나누어 하향식과 구조조정을 반복 적용함으로써 양 방식의 장점을 흡수하고 있다. 초기에 전자 사전에서 특정한 선정 기준을 설정하여 후보 단어들을 추출한 후 이들에 대해 상향식 방법에 의해 작은 크기의 시소러스를 생성하고 이 시소러스를 기준 시소러스로 삼아 나머지 단어들에 대한 추가/확장에 이용하는 방식이다. 이때 기본 시소러스의 생성 단계에서 생성된 시소러스에 대해 상향식의 문제점이 되는 비구조화 연결이나 비일관성 연결을 찾아내어 재구성하는 구조 보정 처리를 하게 된다. 이렇게 구조화된 기준 시소러스에 대해 나머지 단어들을 확장시켜나가므로 상위 계층에 대한 구조적 문제가 해결되고 여러 단계를 거쳐 작업이 이루어지므로 각 단계를 거치면서 하부 시소러스에 대한 비구조성 문제나 비일관성 문제를 해결하면서도 객관성을 유지하고 작업 속도의 향상을 이룰 수 있다.

또한 비적합성의 문제를 해결하기 위해 의미 영역 정보를 표제어가 아닌 뜻풀이문을 한 단위로 하였다. 즉 "카레"라는 단어의 상위어가 "음식"이라는 모든 표제어가 되는 것이 아니라 "먹고 마시는 물건"이라는 뜻풀이문을 가지는 "음식"만을 상위어로 삼는 것이다. 이렇게 함으로서 "카레"라는 단어가 "음식"을 통해 "구체물"도 되고 "추상물"도 되는 비적합의 문제를 해결한다.

3.1 작업 모델

전체 작업은 크게 기본 모델 생성 단계와 추가/확장 단계, 인코딩 단계로 이루어진다. 기본 모델 생성단계는 선출된 후보 단어에서 상위어를 추출하여 구조적 문제가 없는 기본 모델을 만드는 단계로 이 기본 모델을 바탕으로 새로운 단어들을 추가/확장하기 때문에 구조적 문제를 해결한다. 또한 기본 모델의 구조가 전체 시소러스의 구조에 골격이 됨으로 별도의 단계로 나누어 이루어진다. 추가/확장 단계는 구조적 문제가 없는 기본 모델에 대해 새로운 단어들을 추가/확장 시키는 단계로 통계 정보를 이용하여 단어들을 몇 단계로 분류하고 해당 단계에 해당하는 단어들에 대하여 처리를 하고 있다. 마지막으로 인코딩 단계는 완성된 시소러스에서 사용되지 않는 Null Link나 중복 단어를 제거하고 코드화 하는 단계이다.

(그림 1)은 작업 모델을 도식화 한 것이다.



(그림 1) 작업 모델
(Fig. 1) Working model

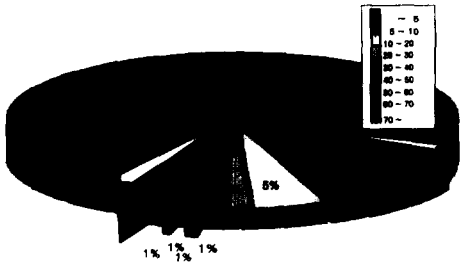
3.1.1 기본 모델 생성 단계

기본 모델 생성 단계는 시소러스의 상위 구조를 결정짓는 단계로 기본 모델 생성에 이용될 단어들을 선택하는 분석 후보 선정 단계와 상위어 추출을 통한 기본 모델 생성 단계, 생성된 기본 모델에 대하여 구조적 문제나 비일관성의 문제를 해결하는 구조 보정 단계로 이루어진다.

전체 시소러스의 구조에 근간이 되는 기본 모델은 전체 시소러스의 구조를 결정짓기 때문에 전체 작업 중 가장 중요한 단계라고 할 수 있다. 또한 기본 모델의 구조는 모델 생성에 이용되는 단어들에 따라 영향을 받기 때문에 분석 후보단어의 선정에는 주의를 기울여야만 한다. 본 시스템에서는 한국 과학 기술원에서 제작한 "한국어 품사 부착 코퍼스"를 이용하여 단어의 출현 빈도를 조사하고 이 통계 정보를 이용하여 분석 후보 단어를 선정하였다.[12][13]

(그림 2)는 단어의 출현 빈도수를 차트화 한 것이다. 차트에 나타난 바와 같이 전체 25,650개의 단어 중 74%인 18,896개의 단어가 5회 이하의 출현 빈도를 나타내고 있다. 나머지 26%의 단어가 5회 이상의 출현 빈도를 나타내고 있다. 본 논문에서는 26%인 6,754개의 단어를 그 후보로 선택하였다. 이 수치는 표제문을 기준으로 한 것으로 평균 한 표제어당 1.5개 정도의 뜻풀

이문을 가지고 있어 9,632개의 뜻풀이문이 선택되어 졌다.



(그림 2) 단어 출현 빈도
(Fig. 2) Appearance frequency of word

그런데 이 선택된 16,233의 뜻풀이문을 살펴보면 <표 1>과 같이 다빈도수의 표제어를 가진 모든 단어를 선택하였기 때문에 현재는 사용되지 않는 고문이나 잘 사용되지 않는 뜻풀이문이 상당수 포함되어 있는 것을 알 수 있다. 예를 들어 提供이 많은 빈도수를 나타내는 표제어이지만 祭供이나 諸公은 현대에 잘 사용되지 않는 뜻풀이문이다.[6] 따라서 본 시스템에서는 이 사어들을 후보에서 제외하였다.

〈표 1〉 사어의 예
〈Table 1〉 Example of unused word

제공[祭供] 【명】 제사에 이바지하는 일 제공[提供] 【명】 바치어 이바지함,쓰라고 줌 제공[諸公] 【명】 점잖은 여러분

3.1.2 추가/확장 단계

기본 모델 생성 단계를 전체 단어에 대한 처리로 확장하는 단계이다. 이것은 기본 모델 생성 단계에서 사용된 단어를 제외한 나머지 단어들에 대해 통계 정보를 적용하여 단계별로 기본 모델에 단어를 추가/확장하는 작업이 이루어진다. 본 시스템에서는 기본 모델 생성 단계와 동일하게 후보 단어의 출현 빈도를 계산하여 5단계로 나누어 처리하게 되었다. 이와 같이 단계를 나누어 처리함으로써 상위구조가 먼저 생성되고 하위가 상위 구조에 추가되는 형태를 취하게 됨으로 구조적 문제를 해결하고 있다. 또한 각 단계의 마지막에 구조 보정 단계를 두어 문제가 되는 노드들에 대한 조정을 하여 비구조성이나 비일관성의 문제를 해결하고 있다.

3.1.3 상위어 추출

사전의 뜻풀이문을 분석하여 상위어를 추출하는 단계로 뜻풀이문의 형태에 따라 유형을 분류해 보면 상위어 출현 위치에 따라 크게 후위형과 중위형의 2가지의 상위어 추출 유형이 나타난다.[14]

1) 후위형

전체 유형의 85.7%정도를 차지하는 형태로 뜻풀이문에 마지막 단어가 상위어인 형태이다. 이를 다시 분류하면 다음과 같다

- "수식언+명사" : 가장 많은 비유을 가지는 형태로 전체에서 82%가 해당된다
  "ㄱ자집 : ㄱ자 모양으로 지은 집"과 같이 뜻풀이문의 전체 문장이 마지막 한 단어를 수식하는 형태로 마지막 단어 "집"을 상위어로 삼는다
- 동의어 : 전체에서 3.7%를 차지한다
  뜻풀이문 자체가 동의어이고 이 동의어를 상위어로 정하여 둘 사이에 동의 관계를 설정해 준다

2) 중위형

나머지 뜻풀이문 안에 상위어가 포함된 형태로 "~의" 형태, "~를" 형태와 용언화 형태가 있으면 전체에서 12.9%가 이에 해당한다

- "~의" 형태
  중위형의 대부분을 차지하는 형태로 "산토끼 : 산에 사는 토끼의 일종"과 같이 "'상위어'의 XXX"에 구조로 되어 있다. 이 XXX의 종류를 살펴보면 "이름, 말, 명칭, 원말, 이칭, 동칭, 동의어, 유의어, 일종, 조각, 품종, 종류, 일종, 준말, 뜻, 용어, 구용어, 속칭, 높임말, 낮춤말, 경어, 존어"등이 있다.
- "~를"
  "스님 : 중을 이르는 말"과 같이 목적어가 상위어가 되는 형태
- 용언화 접미사형
  "가결 : 옳다고 결정함"과 같이 상위어가 용언화 접미사와 결합되어 있는 형태로 용언화 접미사에서 상위어를 분리하여 준다

3) 다의형

중위형이나 후위형과 결합되어 나타나는 형으로 뜻

풀이문안에 상위어 후보가 여러 개 나타나는 형태로 "와"형과 "또는"형이 전체에서 1.4%를 차지한다.

- "와"형
  표제어가 뜻풀이문에 나타나는 상위어들이 가지는 속성을 모두 가지고 있는 형태로 상위어들의 공통 속성을 나타내는 단어가 존재하면 상위어를 추정하고 상위어간의 공통 속성이 존재하지 않는 경우는 모두 상위어로 삼는다
- "또는"형
  하나의 표제어에 많은 뜻풀이문이 중복된 형태로 각각을 별도로 분리하여 새로운 표제어로 만들어 준다

4) 상위어 추정 불가능형

사전의 뜻풀이문만으로 상위어 추출이 불가능한 형태로 전체에서 1.4%를 차지한다. 단어가 행위나 동작, 상태 등 추상물을 나타나는 말일 경우, 또는 "~것"이나 "~일", "~짓"과 같이 우리말에서 포괄적 의미로 많이 쓰이는 단어가 상위일 경우가 이 형태에 해당한다. 이 경우 기존에 형성되어 있는 시소러스를 참조하여 3.1.4에 제시된 방법으로 강제로 상위어를 할당한다

3.1.4 구조 보정 단계

후보 선정 단계에서 선택된 단어들을 중심으로 상위어 추출로 얻은 시소러스에 대해 전체 구조를 헤치는 노드나 비일관성 문제를 발생시키는 노드를 제거하는 단계로 전체 시스템에서 연구자의 주관이 가장 많이 개입되는 단계이지만 뜻풀이문을 기반으로 하여 최대한 객관성을 유지하도록 한다. 그리고 전체적인 형태의 조정뿐만 아니라 노드들 사이의 강제 할당이 필요한 경우 상위어 강제 할당을 한다

1) 구조적 문제 해결

구조 보정 단계에서 처리하여야 하는 구조적 문제는 영역의 일관성 유지와 레벨 유지이다. 영역의 일관성 유지는 같은 부류의 단어임에도 뜻풀이문 차이로 다른 영역에 속해 버리는 문제로 "기관"으로 연결되는 "대학"과 "곳"으로 연결되는 "중학교"와의 관계등에서 발생하는 문제를 해결하여 준다. 본 논문에서는 상기 문제에 대해 "대학"과 "중학교"를 "학교"라는 단어의 하위로 옮기고 "학교"가 "곳"과 "기관"의 속성을 모두 같도록 하여 줌으로서 문제를 해결하고 있다. 또한 단어간의 유사도가 높음에도 불구하고 뜻풀이문의 차이로 위치한 레벨이 틀려지는 문제가 있는 경우 그 높이를 맞추어준다

2) 상위어 강제 할당

뜻풀이문에서 상위어를 추출할 수 없거나 추출된 상위어가 적절하지 못한 경우 사용자에 의한 상위어 강제 할당을 하여 준다. 다음과 같은 경우 상위어 강제 할당을 하여준다.

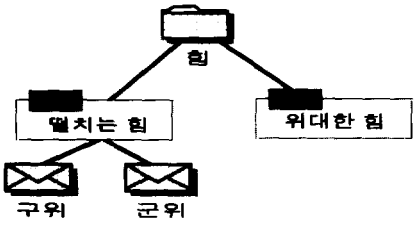
- 상위어 추적을 일정 높이만큼 하여 주었으나 기존의 시소러스와 만나지 못한 경우로 다음 3가지 방법 중 한가지를 이용한다
  - 전체 경로를 다시 추적하면서 상위어 설정이 제대로 설정되어있는가를 살피고 오류를 수정한다
  - 확장된 부분을 포함해서 기본 모델에서 의미상 가장 유사한 노드를 찾아 연결하여 준다
  - 유의어가 없을 때는 새로운 노드로 인식하여 추가하여 준다
- 뜻풀이문에 상위어가 존재하지 않는 경우
  "각이 : 각각 다름"과 같이 뜻풀이문만으로는 상위어를 찾을 수 없는 경우로 이미 만들어진 상의 노드들을 이용하여 추정을 통해 설정해 주게 된다. 또한, "기도 : 공기가 허파로 들어가는 통로"의 경우처럼 뜻풀이문안에 상위어가 존재하지만 그것이 올바르지 못한 경로를 설정할 때 역시 강제 설정이 요구된다.
- 계층의 깊이가 길어질 경우
  의미적으로나 형태적인 문제는 아니지만 계층의 깊이가 길어지면 계층화의 궁극적인 목표인 하위범주화의 특성상 처리의 문제가 발생함으로 깊이를 제어하여 주어야만 한다. 대부분 깊이가 깊어지는 경우는 유사한 의미의 단어들이 상위어로 선택되면서 유사 단어들 사이를 거쳐 진정한 상위로 올라가기 때문이다. "흑석"이라는 단어를 예로 들어보자. 상위어를 이용하여 경로를 추적하면 "흑석"->"흑요석"->"화산암"->"화성암"->"암석"->"바위"->"자연물"->"물체"->"무생물"->"구체물"의 단계를 거치게 된다. 이 것을 자세히 보면 의미상 "흑석"에서 "바위"나 "암석"으로 직접 연결되어도 되나 단어에 대해 불필요할 정도로 상세한 분류에 의해 그 깊이가 깊어지고 있다. 따라서 필요 이상 깊어질 경우 "흑석"을 "암석"에 연결 시켜주는 것과 같이 강제 할당을 하여주어야 한다. 물론 이런 경우의 강제 할당은 신중을 기해야만 한다. 즉, 현재 목표로 하는 구조 체계가 모든 것에 대하여 상세 분류가 요

구되는지 그렇지 않은지를 결정하여 계층의 깊이에 제약 조건을 가한다.

- 포괄적 의미의 상위어가 선택된 경우
  "일"이나 "것", "짓"이 가장 대표적인 예로 "~하는 것"이나 "~하는 일"과 같이 끝나는 경우가 빈번하게 출현하는데 그 의미상 적합하게 연결될 수 있는 노드가 존재함에도 상위어가 "것"이나 "일"이기 때문에 적합한 속성이 부여되지 못하게 된다. 이 경우는 특별한 방법이 없기 때문에 추정에 의한 방법을 사용한다. 즉 "것"이나 "일"이 가지는 속성을 가질 수 있는 단어의 부류를 한정 시켜 놓고 나머지는 전체 기본 모델을 참조 가장 의미가 유사한 노드를 찾아 연결 시킨다

3.1.5 인코딩

인코딩은 구조적으로 존재의 의미가 없는 노드나 동일한 형태를 갖는 중복 노드에 대한 제거 작업을 하는 단계이다. 존재 의미가 없는 NULL link는 단어의 하위어가 존재하지 않고 하위어를 가진 동일한 표제어 노드가 존재할 경우 해당 노드의 존재 가치가 없어지는 것이다. 다음 (그림 3)은 이 NULL Link의 상황을 나타낸 것이다.



(그림 3) NULL Link의 예
(Fig. 3) Example of NULL link

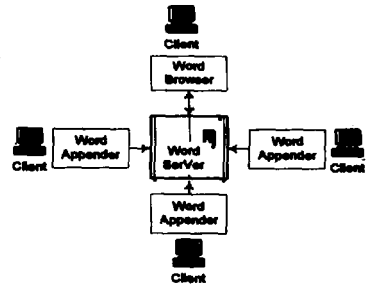
그림에서 "위력"에 대한 두 가지 뜻풀이문중 "위대한 힘"은 하위어를 가지고 있지 않고 동일한 상위어 구조를 갖는 같은 형태의 표제어 "떨치는 힘"이 존재하기 때문에 큰 존재 의미를 가지지 못하기 때문에 "위대한 힘"은 제거하여도 시소러스에 영향을 주지 않아 제거하여 준다.

중복 노드의 제거는 NULL Link제거의 확장 형태로 두 뜻풀이문이 모두 하위어를 가지고 있어도 만일 상위 구조가 동일하다면 (상위어, 다중 상위어, 동의 관계가 일치할 경우) 두 뜻풀이문을 하나로 합치고, 합쳐진 노드가 두 뜻풀이문의 모든 노드들을 하위어로 가지게 한다.

3.2 시스템 설계

본 논문에서 제시하는 시스템에서는 시소러스 구축에 가장 큰 장애중에 하나인 작업 속도의 향상을 위해 전자사전을 이용할 뿐만 아니라 클라이언트/서버 모델을 적용하여 작업자들의 다중 입력이 가능하도록 설계하였다.

(그림 4)는 전체 시스템의 구성도를 나타낸다

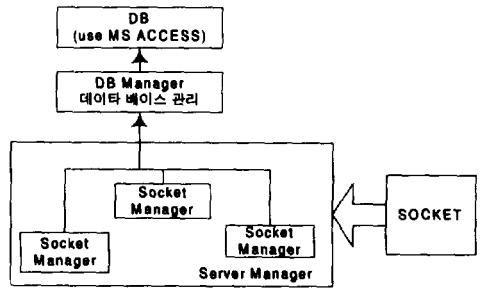


(그림 4) 시스템 구성도
(Fig. 4) System diagram

3.2.1 Word Server

본 시스템에서 가장 중심이 되는 부분인 WordServer는 실제로 시소러스를 관리하는 부분으로 클라이언트들에게 정보를 전달하고 입력을 받아들인다. 전체 구성은 크게 클라이언트와의 연결을 처리하는 Server Manager모듈과 시소러스 데이터를 관리하는 DB manager로 구성된다.

(그림 5)는 WordServer의 전체 구성도를 나타내고 있다.

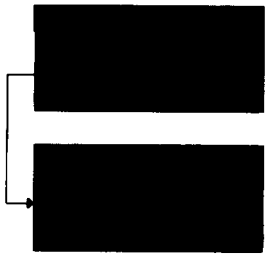


(그림 5) WordServer 구성도
(Fig. 5) Structure diagram of WorsdServer

본 시스템에서 사용되는 데이터베이스는 단어의 표제어와 뜻풀이문을 가지고 전자 사전의 역할을 하는 사전 테이블과 상위어 정보를 보관하는 ID 테이블의 2개 테이블로 구성된다.

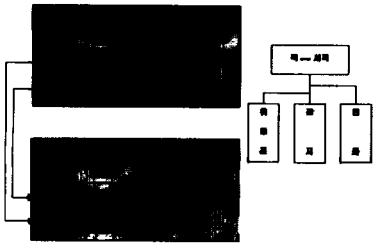
사전 테이블은 전자 사전으로 클라이언트에게 단어의 표제어와 뜻풀이문을 기반으로 하여 후보단어를 제공하는 역할을 담당한다. 상위어 테이블은 해당하는 단어의 ID와 이 단어의 상위어로 구성되어 진다. 따라서 전자 사전에서 원하는 표제어에 대한 상위어를 요청하면 해당 단어의 ID를 가지고 상위어 테이블을 참조한다. 상위어 테이블에서 해당 Id에 상위어ID필드를 참조하여 상위어를 획득하게 된다. 이때 해당하는 ID가 상위어 테이블에 다중으로 존재할 수 있기 때문에 하나의 ID에 대해 여러 개의 상위어를 설정할 수 있도록 한다.[7][8]

(그림 6)은 이들의 구조를 나타낸 것이다.



(그림 6) 테이블 구성도
(Fig. 6) Structure diagram of table

또한 Synonym필드를 두어 동의어 관계를 설정한다. 본 시스템에서는 동의 관계에 있는 단어들 중 하나를 대표로 두고 나머지는 이 단어의 하위로 둔 후 Synonym을 이용하여 동위 관계임을 나타내고 있다, (그림 7)은 "서적", "책", "위인전"의 관계에 대한 테이블과 결과를 트리 구조로 표현한 것이다.[15]

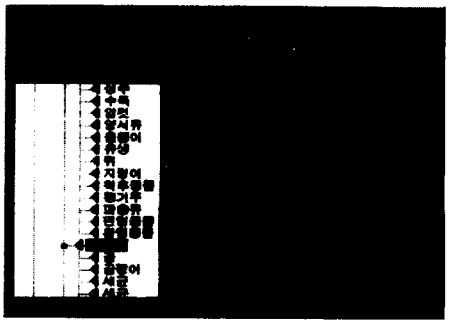


(그림 7) 동의어 관계의 예
(Fig. 7) Relational example of synnonym

Server Manager는 클라이언트와의 통신을 담당하는 부분으로 Socket을 이용하여 이루어진다. Server Manager는 크게 4개 부분으로 나뉘는데 우선 메시지를 번역하는massage interpreter부, 메시지의 내용을 분석하여 해당 작업을 실행 시키는 Message Dispatcher부, DB와 연결하는 DB부, 해당 결과를 클라이언트에 넘겨주는 Message Sender로 나뉜다.

3.2.2 WordBrowser

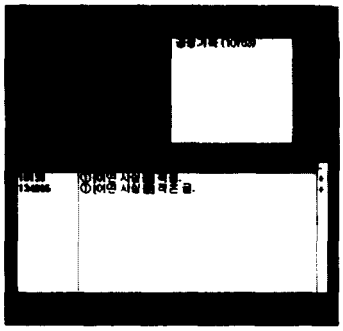
WordBrowser는 시소러스 관리 도구로 전체 시소러스의 구조를 한 눈에 확인해 볼 수 있도록 트리 구조를 제공한다. 이 WordBrowser를 통해 각 단계의 마지막에 구조 보정을 할 수 있을 뿐만 아니라 시소러스 구축 중간에도 보정이 가능하도록 추가/확장기와 동시에 서버와 연결되어 구조를 보정하여 준다. 또한 트리 구조에 대해 삭제, 이동, 복사를 가능하도록 하여 구조를 자유롭게 조정하여 줄 수 있도록 하고 있다. 다음 (그림 8)은 WordBrowser의 작업 화면이다.



(그림 8) WordBrowser 실행 예
(Fig. 8) Executing example of WordServer

3.3.3 WordAppender

새로운 단어를 추가/확장 시키는 프로그램으로 서버로부터 후보 단어를 받아 상위어를 선택하여 다시 서버에 넘겨주는 역할을 담당한다. 또한 뜻풀이문을 분석하고 기본 모델을 참조하여 상위어를 추측하여 사용자에게 제시하여 줌으로서 작업 속도를 높이고 있다. 다음은 "공정기록"의 상위어를 추적하는 과정을 예로 든 것이다. 우선 서버로부터 "공정기록"의 정보를 획득하고 형태소 분석기와 통계 정보를 이용하여 상위어 후보를 사용자에게 보여 준다.[14] (그림 9)는 "공정기록"의 상위어로 "기록"을 선택한 후 사용자의 선택을 기다리는 모습이다.



(그림 9) 상위어 선택을 위한 사용자 화면
(Fig. 9) User Interface for selecting upper word

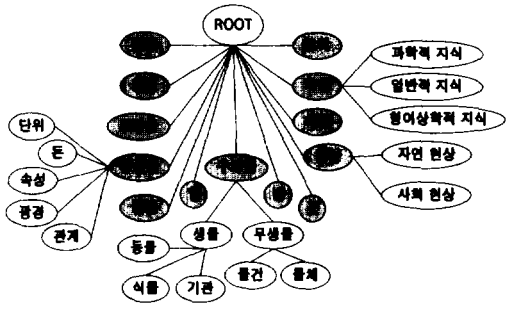
다음 (그림 10)은 상위어를 선택한 결과 화면이다

(그림 10) 상위어 선택 결과
(Fig. 10) The result of selecting the upper word

4. 결과 및 고찰

전체 작업은 총 5단계를 거쳐 이루어 졌다. 최초 단계는 기본 모델을 생성하는 단계로 코퍼스에서 선정된 6,754개의 후보 단어를 이용하여 상위어 추정을 하였다. 기본 모델 생성시 후보에 존재하지않는 단어가 나타나면 새로이 등록하여 주고 쓰이지 않는 단어가 나타나면 제거하여 9,754개의 뜻풀이문으로 이루어진 기본 모델이 생성되었다. 그 결과를 살펴보면 최상위 노드는 "행위, 상태, 표상물, 추상물, 방법, 힘, 구체물, 병, 곳, 현상, 사건, 지식, 조직"의 13개 단어로 구성되어 진다. (그림 11)은 이들 상위 노드를 나타낸다.

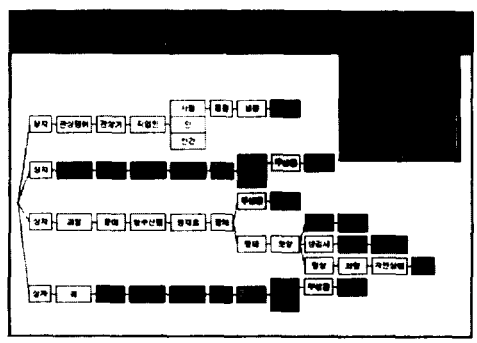
기본 모델이후 코퍼스에 출현 빈도에 따라 3단계를 거쳐 새로운 단어를 추가확장 시킨다. 첫번째 단계에서는 "한국어 품사 부착 코퍼스"에서 얻은 통계 정보 중 2~5개의 빈도를 가지는 5,386개의 표제문으로 시



(그림 11) 상위 노드 구조
(Fig. 11) Structure of super nodes

작하여 9,105개의 뜻풀이문을 추가하였다. 다음으로 2번째 단계에서는 1번의 출현 빈도를 가지는 53%의 단어인 13,510개의 표제문을 처리 대상으로 하였고 결과로 17,325개의 뜻풀이문이 추가되었다. 마지막 단계는 코퍼스에 나타나지 않은 나머지 11,588개의 어의문에 대하여 어의문을 처리하였다

결과적으로 전체 시소러스는 47,650개 뜻풀이문으로 이루어지고 최대 9레벨, 평균 5레벨의 높이를 가지고 있다. 다음 (그림 12)는 "상자"의 상위어 정보 결과를 보여 주고 있다.



(그림 12) "상자"의 상위어 정보
(Fig. 12) Hierachy of a word

5. 결 론

본 논문에서는 한국어 명사의 의미 개념의 계층을 생성하기 위한 시소러스 구성 방법과 시소러스를 구축하기 위한 개발 시스템을 구현하였다. 기존의 하향식의 단점이었던 비객관성과 작업 속도의 문제를 해결하기 위해 상향식을 적용하였고 상향식의 문제점인 비일

관성과 비구조성의 문제를 해결하기 위해 전체의 작업을 여러 단계로 나누어 처리하였다. 또한 비적합성의 문제를 해결하기 위해 작업의 기본 단위를 표제어가 아닌 뜻풀이문을 이용하여 같은 표제어를 가지고 있더라도 어의문이 다를 경우 다른 노드를 생성하도록 하여 주었다.

본 논문에서 제시한 시스템은 뜻풀이문을 이용하여 상위어 선정에 걸리는 시간을 단축하였을 뿐만 아니라 클라이언트/서버 모델을 적용하여 다수의 사용자에 의한 동시 작업이 가능하여졌기 때문에 전체 작업 속도를 높일 수 있었다. 또한 시소러스의 기본단위를 표제어가 아닌 뜻풀이문으로 하여 같은 표제어라도 뜻풀이문이 다르면 별개의 단어로 인식하여 줌으로서 비적합성의 문제를 제거하였다.

그러나 뜻풀이문을 이용하는 상향식을 적용함으로써 상위 구조가 정리되지 못하고 산만해 지는 현상이 발생하여 초기 기본 모델의 보정 단계에서 단순화 작업이나 추정 불가능한 단어에 대한 강제할당 작업시 주관이 많이 개입된다는 문제를 가지고 있다.

참 고 문 헌

[1] George A. Miller et al. "Introduction to WordNet : An on-line Lexical Database," in Five Papers on WordNet, CSL report. Cognitive science Lab., pp.1-9, Princeton University, 1993.

[2] George A. Miller, "Nouns in WordNet:A Lexical Inheritance System," in Five Papers on WordNet, CSL report. Cognitive science Lab., pp.10-25, Princeton University, 1993.

[3] Richard Beckwith, et al., "Design and Implementation of the WordNett Lexical Database and Searching Software," in Five Papers on WordNet,CSL report. Cognitive science Lab., pp. 62-77, Princeton University, 1993.

[4] William B. Frakes, Ricardo Baeza-Yates, "Information Retrieval," Prentice Hall, 1992.

[5] 仲尾由雄 et al., "日本電子化辭書研究所における概念體系", 自然言語處理 93-1, 1993.

[6] 田中慧積, 仁科喜久子, "上位/下位關係シソーラスISAMAPの作成[I]", 自然言語處理 64-4, 1987.

[7] 竹下克典, 伊丹克企 et al, 國語辭典情報いたシソーラスの作成について, 自然言語處理 83-16, 1991.

[8] 田中慧積, 仁科喜久子, "上位/下位關係シソーラスISAMAPの作成[II]", 自然言語處理 64-4, 1987.

[9] 김영택, "자연 언어 처리", 교학사, 1994.

[10] 박영자, "사전에서 추출한 의미 속성에 기반한 명사 의미 클러스터링", 정보과학회논문지, 25권 3호, pp.585-595, 1998.

[11] 문유진, "한국어 명사를 위한 WordNet의 설계와 구현", 정보과학회논문지, 2권, 4호, pp.437-444, 1996.

[12] "통합 국어 정보 베이스를 위한 한국어 형태통사 태그 설정", 한국 과학 기술원, 1996.

[13] 최기선, "한국어 품사 부착 코퍼스", 한국과학기술원, 1997.

[14] 조평옥, "한국어 명사의 의미 계층 구조 구축", 울산대학교 석사 학위 논문, 1996.

[15] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 졸업 논문, 1993.

[16] 김상형," 금성판 국어대사전", 금성출판사, 1992.

[17] 한글학회, "한글 우리말 큰사전", 한글과 컴퓨터, 1996.

이 종 인

email : nyquist@shinbiro.com

1997년 2월 호서대학교 컴퓨터공학과 졸업

1997년 3월~1999년 현재 호서대학교 컴퓨터공학과 석사과정

관심분야 : 자연어처리, 정보검색, 멀티미디어 저작시스템

한 광 록

email : krhan@dogsuri.hoseo.ac.kr

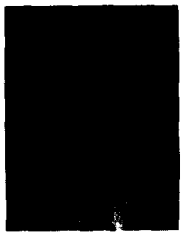
1984년 2월 인하대학교 전자공학과 졸업

1986년 2월 인하대학교 대학원 전자공학과(공학석사)

1989년 8월 인하대학교 대학원 전자공학과(공학박사)

1991년 3월~1999년 현재 호서대학교 컴퓨터공학부 교수

관심분야 : 자연어처리, 정보검색, HCI, 멀티미디어 저작시스템 등



양 승 현

e-mail : shyang@etri.re.kr

1989년 서울대 공대 컴퓨터공학과 졸업(학사)

1992년 서울대 대학원 컴퓨터공학과 졸업(석사)

1997년 서울대 대학원 컴퓨터공학과 졸업(박사)

1997~현재 한국전자통신연구원 자연어처리연구부 선임연구원

관심분야 : 자연언어처리, 기계번역, 정보검색

김 영 섬

e-mail : yskim@etri.re.kr

1985년 한양대학교 공대 대학원 졸업(석사)

1990년 한양대학교 공대 대학원 졸업(박사)

1989년~현재 한국전자통신연구원 자연언어처리연구부 선임연구원

관심분야 : 정보검색, 자연언어처리, 음성언어처리