1998 年度 大韓電子工學會 夏季綜合學術大會 論文集 第 21 卷 第 1 號 98/6

# 한국어 단어 시소러스 구축 시스템의 설계

° 이 종 인*, 한 광 록*
* 호서대학교 컴퓨터 공학과

# Design of Construction system for Thesaurus of Korean Word

° Jong-In Lee*, Kwang-Rok Han*
* Dept. of Computer Engineering, Hoseo University

요약 본 논문은 한국어 단어의 의미 영역 정보를 저장하기 위한 시소러스를 설계하고 이를 구축하기 위한 시스템을 설계하였다. 기존에 Top_Down 이나 Bottom_Up 방식을 이용하는 경우 각각 비 객관성과 작업 속도의 문제와 비구조성,비일관성의 문제를 안고 있어 이를 혼합하여 어의문을 이용하여 객관성을 유지하면서도 기본모델을 이용하여 비구조성과 비일관성의 문제를 해결하고 있다. 또한 그 동안 필요성이 증가되었으나 작업을 하지 못했던 가장 큰 이유인 단어의 방대성으로 기인한 작업 속도의 문제해결하기 위하여 C/S 모델을 적용하여 다수의 입력자들에 의해 동시 입력을 가능케 함으로써 작업 속도의 향상을 이루었다

## 1. 서론

한국어 단어의 의미 영역 정보는 자연어 처리에 있어서 형태소 분석이나 통사 분석 같은 문법 요소의 분석만으로는 처리할 수 없는 언어적 모호성의 문제를 해결하기 위해 전체 문장이 가지는 의미를 분석해 올바를 결과를 산출해 내는 의미 분석 단계에서 하위 범주화 정보와 함께 이용되는 정보이다.[1] 한국어 단어의 시소러스를 구축하는데 있어서 핵심이 되는 것은 일관성의 유지, 구조화, 객관성 유지, 작업 속도의 향상이다. 이를 위한 방법론으로서 일본 EDR 등에서 사용된 Top_Down 방식과 이의 문제를 해결하기 위해 나온 Bottom_Up 방식이 있다.[3][4] 그러나 Top_Down 의 경우에는 객관성의 문제와 작업 속도의 문제가 있고 Top_Down 의 경우는 일관성의 문제 및 구조성에 문제를 안고 있다.[4] 본 논문에서는 이를 절충하여 Top_Down 의 구조성과 일관성, Bottom_Up 의 객관성 및 작업 속도를 얻을 수 있는 다단계 방법을 제시한다.

## 2. 다단계 방법

다단계 방법은 시소러스의 구축에 몇 가지 단계를 두고 Bottom_Up 과 구조조정을 반복 적용하여 양방법의 장점을 흡수하려는 시도이다. 즉, 초기에 후보 단어들을 추출하고 이들에 대해 Bottom_Up 을 적용하여 기본적인 시소러스의 형태를 생성한 후 나머지 단어들에 대해 추가/확장을 이루는 방식이다. 이때 기본 시소러스의 생성 단계에서 생성된 시소러스에 대해 Bottom_Up 의 문제점이 되는 비구조화 연결이

나 비일관성 연결을 찾아내어 재 구성하는 구조조정 처리를 하게 된다. 이렇게 함으로서 Bottom_Up 에서 문제가 되는 비구조성문제나 비일관성 문제를 해결하면서도 객관성을 유지하고 작업 속도의 향상을 이룰 수 있다.

## 3. 시스템 설계

### 3.1 기본 모델

본 논문에서 제시된 시스템은 크게 기본 모델 생성 단계와 추가 확장 단계로 이루어 져있다.

다음 그림 1 이 그 전체 시스템 흐름도이다.

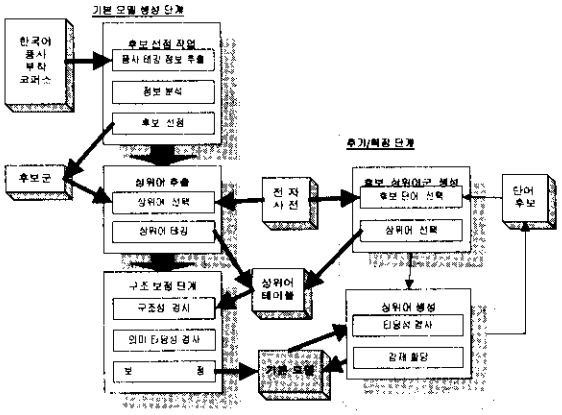



그림 1. 전체 시스템 흐름도

기본 모델 생성 단계에서는 한국 과학 기술원 제작의 "한국어 품사 부착 코퍼스"를 분석하여 단어의 출현 빈도수를 얻어내고 그 중 전체의 26%를 차지한 5 회 이상의 출현 빈도를 갖는 단어들을 이용하여 Bottom_Up 방식을 적용하여 기본 모델을 생성하였고 구조 보정 단계에서 구조적 문제점을 조사/보정하여 구조적 문제가 없는 기본 모델을 생성한다.[8][9]

이렇게 생성된 기본 모델에 대하여 본 논문에서 제시하는 시스템에서는 C/S 모델을 적용하여 다수의 클라이언트들이 어의문을 통한 상위어 추출을 하도록 허가하고 하나의 Administrator 를 통해 구조적 문제를 감시하도록 구성하고 있다.

### 3.2. 추가/확장 시스템

본 논문에서 제시한 시스템에서는 전체 작업의 속도를 높이기 위하여 C/S 모델로 개발이 되었다.

다음 그림 2 는 이를 도식화한 것이다

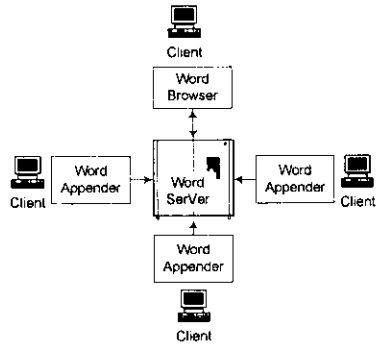



그림 2 전체 시스템도

.WordServer 는 크게 두 가지 기능을 수행한다. 가장 기본적인 기능으로 시소러스를 구성하는 데이터 베이스를 관리하는 데이터 베이스 관리 기능이 있고 데이터 베이스와 외부 클라이언트를 연결/관리해주는 서버 기능이 있다.

WordAppender 는 시소러스에 새로운 단어를 추가 확장하는 기능을 수행한다. 서버로부터 처리할 후보 단어와 단어의 어의문을 넘겨 받아 사용자로 하여금 상위어를 선택하도록 한다.

WordBrowser 는 WordAppender 를 통해 확장되는 시소러스가 구조적인 문제를 일으키지 않는지 감시하고 문제가 발생했을 경우 구조를 조정하는 역할을 담당한다.

### 3.3 WordServer

본 시스템에서 가장 중심이 되는 부분으로 전체적인 구성은 크게 두 부분으로 나뉘어 지는데 클라이언트와의 연결을 처리하는 Manager 부분과 데이터 베이스를 관리하는 DB Manager 부분이 그것이다.

다음 그림 3 이 전체 구성도를 보이고 있다.

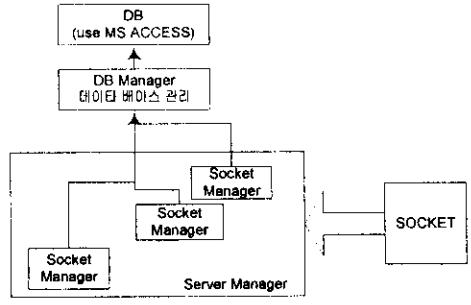



그림 3. 구성도

1) 데이터 베이스 관리

데이터 베이스는 3 개의 테이블로 구성되어진다. 다음 그림 4)가 그 구조이다.

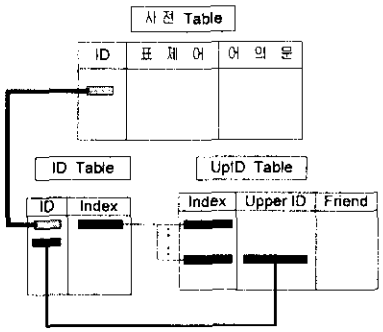



그림4. 데이터 베이스 구성도

사전 테이블은 단어의 표제어와 어의문을 보관하는 테이블로 클라이언트에게 표제어를 보여주거나 후보단어를 제공하는 역할을 담당한다. 다음의 ID 테이블은 해당하는 단어의 아이디와 이의 상위어를 보관한 UpID 테이블을 연결해 주는 인덱스로 구성되어진다.[10][11] 상위어를 두 단계로 구성한 이유는 단어의 다의성에 기인한 것으로 같은 의미의 단어가 두개 이상의 상위어를 가질 수 있기 때문이다.[6] 또한 UpID 테이블에 Friend 필드를 두어 동의어 관계를 설정하고 있다.[12] 본 시스템에서는 동의 관계에 있는 단어들 중 하나를 대표로 두고 나머지는 이 단어의 하위어로 둔 후 Friend 를 이용하여 동위 관계임을 나타내고 있다. 다음 그림 5 는 "서적","책","위인전"의 관계를 예로 든 테이블 모습이다.[7][5]

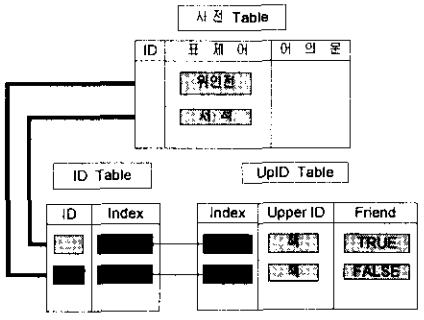



그림5. 서적, 책, 위인전의 예

그림 6 은 이를 트리 구조로 나타냈을 경우에 모습이다 .

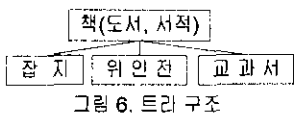



그림 6. 트리 구조

2) Socket Manager

클라이언트와의 통신을 담당하는 부분으로 Socket 을 이용하여 이루어진다. 다음 그림 7 이 그 구성도 이다.

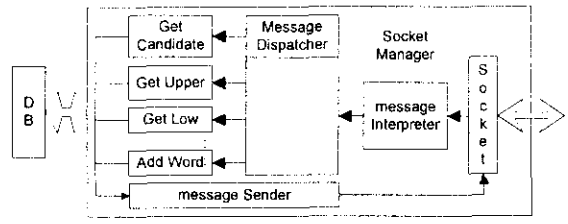



그림 7 . Socket Manager의 구성

이 부분은 크게 4 개 부분으로 나뉘는데 우선 메시지를 번역하는 Message Interpreter 부, 메시지의 Operation 부를 분석하여 해당 작업을 실행 시키는 Message Dispatcher 부, DB 와 연결하는 DB 부, 해당 결과를 클라이언트에 넘겨주는 Message Sender 로 나뉜다.

4. 작업 결과 및 고찰

다음 그림 8 은 새로운 단어를 추가/확장하는 WordAppender 의 작업 화면으로 사용자들은 서버로부터 전달되어 지는 단어들에 대하여 어의문을 통해 상위어를 선택하도록 하고 선택한 상위어가 기존 기본 모델에 존재할 경우 다음 후보로 넘어 갈수 있도록 하였다.

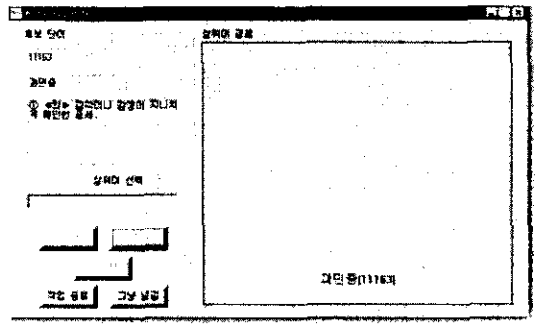
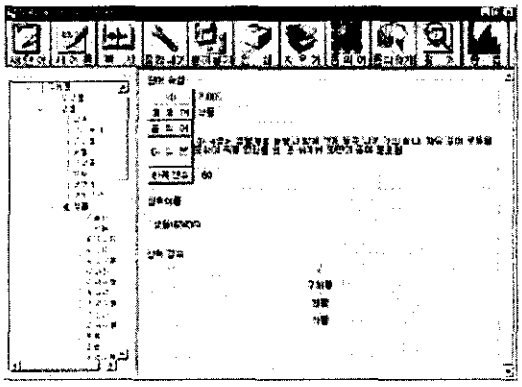


그림 8. WordAppender

다음 그림 9 는 시소러스를 관리하는 WordBrowser 로 트리 구조로 나타나는 결과에 대해 원하는 형태로의 변형이 가능하다.



기본 모델 생성을 위해 작업의 후보로 6754 개의 단어로 시작했으나 작업 중 후보에 존재하지 않는 중간 노드의 단어들에 의해 추가되고 일반적으로 잘 사용되지 않는 단어를 제외시키는 등의 처리 후에 12702 개의 어의문으로 이루어진 기본 작업 모델이 만들어 졌다. 이의 구조를 보면 가장 최상위 레벨의 노드로 다음의 13 개가 선택되어졌다.

"구체물, 추상물, 곳, 관계, 단위, 방법, 활동, 현상, 표상물, 조직, 속성, 상태, 사건"

## 5. 결론

본 논문에서는 객관성을 유지하면서도 일관성과 구조성을 유지하는 의미 체계의 구축 방법을 제시하였다. 구축에 객관성을 유지하기 위하여 기본 접근법으로 기존의 Bottom_Up 방식을 사용하였고 일관성과 구조성을 유지하기 위하여 기본 모델을 생성한 후 전체 단어로 확장 시켜 나가는 방식을 취하였다.

이 방법은 두 방식을 혼합한 형태를 취함으로 해서 Bottom_Up 방식에 비해 조금 주관적이라는 문제를 안고 있다. 특히 기본 모델 생성시 구조 조정단계에는 특히 주관성이 개입되고 있다. 그러나 기본 모델의 생성은 Bottom_Up 을 통해 만들어진 결과를 가지고 잘못된 노드를 수정하는 것이기 때문에 객관성을 해칠 만큼 주관성이 개입되지는 않는다. 또한 이 방식을 사용 시 Bottom_Up 의 객관성을 해치는 것보다는 Top_Down 방식에 비해 객관적인 것이 더 크므로 큰 문제가 되지는 않는다.

또한 C/S 모델을 이용하여 동시에 작업을 할 수 있도록 하여 기존의 모델에 가장 문제점이었던 작업 속도의 문제를 해결할 수 있었다.

그러나 본 논문에서 제시한 방법으로 처리되지 않는 몇 가지 문제가 존재한다.

우선 사전들이 단어의 상/하위 관계를 고려하지 않고 만들어 졌기 때문에 어의문에 있는 상의어가 진정한 그 단어의 상의어가 아닐 경우가 발생한다는 것이다. 본 논문의 경우 기본 모델에서 유사 단어를 찾는 방법을 통해 해결을 시도하였지만 전혀 알지 못하는 단어가 나타날 경우는 단지 추측에 의한 방법만이 가능할 뿐이다. 또한 체계 구축 시 어느 수준까지의 세부 분류가 필요한지에 대한 연구도 역시 필요하다. 세분 분류 수준은 전체의 깊이 한계를 결정하는데 영향을 주기 때문이다

## 참고 문헌


[1] 김영택, " 자연 언어 처리",교학사, 1994

[2] Thomas A Sebeok, " Approaches To Semiottics", Mouton Publishers,1975

[3] 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

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

[5] 김상형, "금성판 국어대사전",금성출판사, 1992

[6] 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

[7] 한글학회, "한글 우리말 큰사전 96",한글과 컴퓨터, 1996

[8] 최기선, "한국어 품사부착 코퍼스",한국 과학기술원, 1997

[9] 이공주, 김재훈 et al., "한국어 구문 트리 태깅 코퍼스 작성을 위한 한국어 구문 태그", 한국 과학기술원

[10] 田中穗積, 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

[11] 田中穗積, 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 64-5, 1987

[12] 문유진 , "한국어 명사를 위한 WordNet 의 설계와 구현", 정보과학회논문지 437-444, 1996