2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

Graph Editor형식의 통합정보사전 개발 시스템

남동수*, 최용준**, 황도삼*
*영남대학교 컴퓨터공학과

e-mail:dshwang@yu.ac.kr

A thesaurus development system with an embedded graphic editor

DongSu Nam*, Yongjun Choi**, Dosam Hwang*

*Department of Computer Science Yeungnam university and
Advanced Information Technology Research Center(AITrc)
**Dept of Computer Engineering, Yeungnam University

요약

통합정보사전은 고도의 언어처리 및 이해를 목적으로 한 것이며, 체계적이고 과학적인 방법론을 이용하여 형태소, 구문, 의미정보 등 각종 정보가 통합된 전자사전으로, 이를 개발하는데는 막대한 개발 시간과 노력을 필요로 한다. 이러한 특성 때문에 통합정보사전을 구축하기 위해서는 정보를 통합하고 관리하는 사전개발 시스템의 개발이 선행되어야 한다. 현재까지의 사전개발 시스템은 사전 항목을 정의하고, 정의된 항목에 표제어별 정보를 입력하는 시스템으로, 단순한 정렬 및 검색에 의한 표제어 찾기 및 편집을 지원하고 있다. 본 논문에서는 사전의 계층화된 항목정보를 트리 형식으로 나타내어 사전의 개발 및 구축작업을 효율적으로 지원하기 위한 통합정보사전 개발 시스템인 YDK3를 설계하고 구현하였다. 구현한 YDK3는 기존의 각종 사전의 다양한 사전정보를 입력하는 기본적인 기능 외에, 항목정보를 기반으로 한 graph editor형식의 사용자 인터페이스가 제공되어, 사전의 개발, 자료입력 및 검색을 보다 쉽게 할 수 있다는 특징이 있다.

1. 서론

통합정보사전은 고도의 언어처리 및 이해를 목적으로 한 것이며, 체계적이고 과학적인 방법론을 이용하여 형태소, 구문, 의미정보 등 각종 정보가 통합된 전자사전으로, 이를 개발하는데는 막대한 개발 시간과 노력이 필요하다[1-4]. 전자사전을 구축하는 일은 노동집약적이면서도 전문지식을 필요로 하기 때문에 개발과정이 오래 걸리며, 개발한 사전은 지속적으로 단어가 추가될 뿐 아니라 사전의 구조 자체도 변경되게 된다[5-7]. 이러한 특성 때문에 통합정보사전을 개발하고 구축하기 위해서는 정보를 쉽게 통합하고 관리하는 사전개발 시스템의 개발이 선행되어야만 한다. 현재까지의 사전개발 시스템은 사전 항목을 정의하고, 정의된 항목에 표제어별 정보를 입력하는 시스템으로, 단순한 정렬 및 검색에 의한 표제어 찾기 및 편집기능을 지원하고 있다. 본 논문에서는 사전의 계층화된 항목정보를 트리 형식으로 나타내어, 사전의 개발 및 구축작업을 효율적으로 지원하기 위한 통합정보사전 개발 시스템인 YDK3를 제시한다. 구현한 YDK3는 기존의 각종 사전의 다양한 사전정보를 입력하는 기본적인 기능 외에 항목정보를 기반으로 한 graph editor형식의 사용자 인터페이스가 제공되어 있어, 사전의 개발, 자

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

료입력 및 검색을 보다 쉽게 할 수 있다는 특징이 있다.

2. 사전 시스템

지금까지 전자사전은 대부분의 경우 각 응용 시스템마다 다른 구조의 사전 시스템을 사용하고 있으며, 정보를 추출하는 방법이 명확하지 못하여 사전을 개발하고 자료를 입력하는데 많은 시간과 비용이 소요된다. 또한 특정 구조와 환경에 맞춰 개발되었기 때문에 새로운 사전이 만들어질 때마다 별도의 사전관리 시스템이 만들어져야만 한다[5].

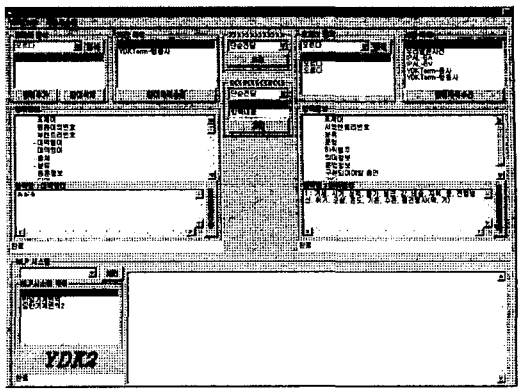
이러한 문제점들을 해결하고 각종 사전들의 표준 형태를 정의하여 표준 사전을 개발하고 관리하기 위한 도구로 KAIST에서 개발한 TDMS(Text Dictionary Management System)와 영남대학교에서 개발한 YDK(Yongjun-Dosam-Keysun Dictionary System)가 있다.

2.1 TDMS

KAIST에서 개발한 대표적인 사전시스템인 TDMS는 SGML을 기반으로 하여 여러 분야에서 필요로 하는 각종 사전들의 표준 형태(SDF: Standard Dictionary Format)을 정의하고, 표준 사전(SD : Standard Dictionary)을 구축하는데 사용하는 시스템으로 SDF의 정의 및 SD의 편집, 수정, 검색, 변환할 수 있는 사전 및 텍스트 관리 통합시스템이이다[7]. TDMS와 같은 일반적인 사전 시스템들은 국내 사전들의 자료를 참조하여 하나의 사전을 개발하는 데는 효과적이지만 외국 사전들의 참조는 번역 문제로 인해 어려우며, 자연언어처리 도구의 처리 결과를 사전자료로 활용하는 것이 불가능하므로, 통합정보사전과 같은 대규모의 사전을 개발하는 데는 부적합하다.

2.2 통합정보사전 개발 시스템:YDK

영남대학교에서 개발한 YDK는 전자사전, 기계번역 시스템 등 분산된 언어자원들을 하나의 플랫폼으로 통합하기 위한 통합모델을 기반으로 한 통합정보사전 개발 시스템이다. 이 시스템은 각 자원들의 고유한 작동 플랫폼을 그대로 둔 상태에서 모듈의 추가만으로 간단히 YDK에 연결되어 사전개발환경을 구성하게 된다. YDK 시스템의 사용자 인터페이스를 [그림 2]에 보인다.

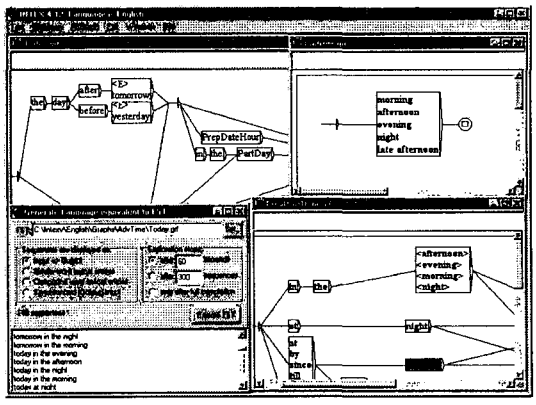


[그림 1] YDK Browser

YDK는 다양한 사전정보의 통합, 자연언어처리(NLP)도구와의 연계를 지원하여 정보를 통합하는 것이 특징인 통합정보사전 개발 시스템이다.

2.3 graph editor

graph editor는 graph 형식으로 사전을 개발하는 것을 지원하는 시스템으로 LADL(Laboratoire d'Automatique Documentaire et Linguistique)의 INTEX가 있다. INTEX는 각각의 정보들을 노드화시키고 이를 연결하여 그 연계성을 정의하는 방식으로 작동하며, 사전을 직접적으로 개발하는 도구가 아니라, 언어적인 자원 개발에 사용되는 범용도구다[8]. INTEX의 사용자 인터페이스를 [그림2]에 보인다.



[그림 2] INTEX

TDMS와 YDK는 사전개발 및 구축작업을 손쉽게 하도록 지원해주지만, 항목별 단순한 정보입력이라는 체계를 벗어나지 못하고 있다. 사전의 개발과정에 항목에 입력될 정보형식에 대한 정의가 변경될

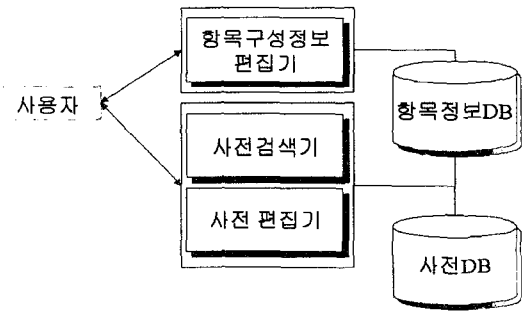
경우 이를 일괄적으로 처리하기가 매우 어려우며, 상하계층관계가 존재하는 항목정보일 경우 그 연계성을 검색하기가 어렵다. 따라서 전자사전들의 자료를 직접 참조할 수 있으며, 항목의 정보체계를 관리하고 사전개발작업을 효과적으로 수행할 수 있는 graph editor형식의 통합정보사전 개발이 필요하다. 개발하는 사전 시스템은 강력한 기능과 함께 사용하기 쉬운 사용자 인터페이스를 가져야 하며, 실제 사전 개발에 이용할 수 있어야 한다.

3. graph editor형식의 YDK3

YDK3는 기존의 각종 사전의 다양한 사전정보를 입력하는 기본적인 기능 외에 항목정보를 기반으로 한 graph editor형식의 사용자 인터페이스를 가지고 있어 사전의 개발, 자료입력 및 검색을 보다 쉽게 할 수 있어야 한다.

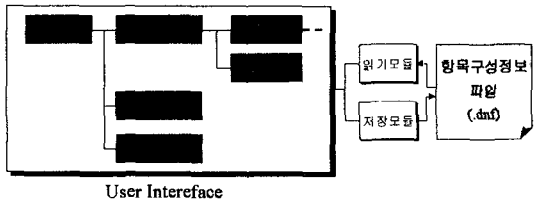
3.1 시스템 설계

통합정보사전에서 graph형식으로 나타나야 할 부분은 항목정보를 기준으로 나타내는 경우이다. 즉, [그림 3]과 같이 계층적으로 구성된 항목정보를 바탕으로 하여 사용자 인터페이스에 나타내고 이를 저장하게된다. 사전의 자체 DB에는 변화를 주지 않고 별도의 항목정보DB를 구성하여 사용한다.



[그림 3] YDK3의 시스템 구성도

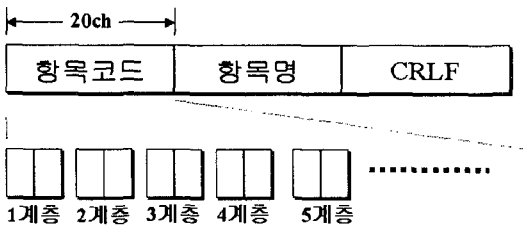
항목정보 DB는 시스템의 운영과정에서 작동을 하기 위한 DB로 정보의 편집이 완료되어 '사전 저장'을 선택하게 되면, [그림 4]와 같이 항목구성정보 파일로 디스크에 저장하게 된다. 해당사전을 선택하면, 다시 이 파일에서 불러와서 항목정보DB에 넣어 사전편집에 사용할 수 있게된다. 이렇게 하면 특정 사전의 개발과정뿐 아니라, 다양한 사전의 개발에 사용할 수 있게 되는 장점이 있다.



[그림 4] YDK3의 항목구성정보 관리구조

3.2 항목구성정보 저장구조

항목정보는 디스크에 파일로 저장되는데, 특정한 Markup Language를 사용하지 않고 각 항목별로 텍스트 파일에 한 줄씩 저장한다. 항목이름은 길이가 유동적이므로 길이를 지정하지 않고 항목정보의 계층정보를 나타내기 위해 20ch크기를 할당한다. 하나의 계층을 나타내는데 2ch를 사용하도록 하였으므로, 총 10계층까지 나타낼 수 있다.
이 구조를 [그림 5]에 나타낸다.



[그림 5] 항목구성정보 저장구조

[그림 4]에 일부 나타낸 정보를 저장할 경우 [그림 6]과 같은 구조로 저장이 된다. [그림 6]에서 '자동사'와 '타동사'같은 경우는 '동사'의 하위계층에 들어있음을 알 수 있다.

```
01 00 00 00 00 00 00 00 00 00 00 00 품사CRLF
01 01 00 00 00 00 00 00 00 00 00 00 동사CRLF
01 01 01 00 00 00 00 00 00 00 00 00 자동사CRLF
01 01 02 00 00 00 00 00 00 00 00 00 타동사CRLF
01 02 00 00 00 00 00 00 00 00 00 00 형용사CRLF
01 02 00 00 00 00 00 00 00 00 00 00 부사CRLF
```

[그림 6] 저장된 항목구성 정보의 예

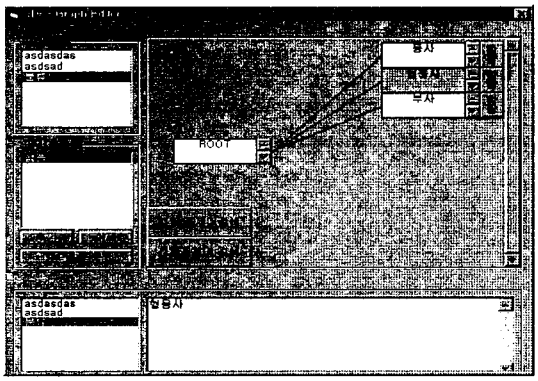
3.3 구현환경

YDK3는 MS-Windows 98을 운영체제로 사용하는 Intel Pentium-III 500Mhz을 사용하는 IBM-PC

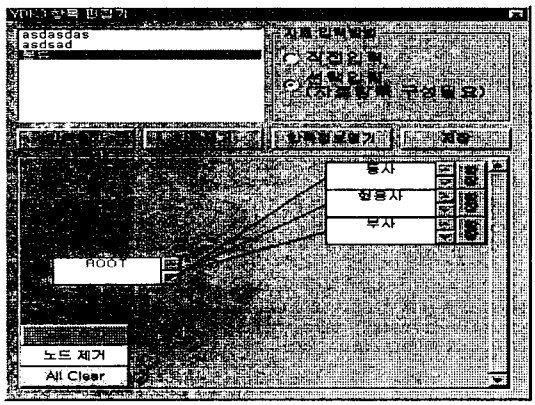
호환 컴퓨터에 개발하였다. 개발도구는 MS-Visual C++ 6.0과 MS-Visual Basic 6.0을 이용하였으며, MS-Access를 데이터베이스로 사용하였다.

3.4 결과

실제 구현한 YDK3에서 항목구성정보는 [그림 7]에 나타낸 것과 같이 가로방향의 트리 형태로 나타나며 마우스 오른쪽 버튼으로 메뉴를 호출하여 항목구성정보를 편집하거나 트리 구조 자체를 바꿀 수도 있도록 되어있다. [그림 7]에 사전자료 구축을 위한 사용자 인터페이스를 나타내었고, [그림 8]에 항목정보를 정의하는 사용자 인터페이스를 나타내었다. 특정한 분류에서 선택하는 경우와 단순히 입력하는 경우(예:용례 입력등)를 구분하여 입력하도록 구현하였다.



[그림 7] YDK3 사용자 인터페이스



[그림 8] YDK3 항목정보 입력 사용자 인터페이스

4. 결론

본 논문에서는 사전의 계층화된 항목정보를 트리 형식으로 나타내고 관리할 수 있는 통합정보사전 개발 시스템인 YDK3를 설계하고 구현하였다. 구현한 YDK3는 Graph Editor형식의 사용자 인터페이스를 가지고 있어 사전의 개발, 자료입력 및 검색을 보다 쉽게 할 수 있다는 특징이 있다.

향후연구과제로는 다양한 NLP도구의 처리결과를 자동으로 YDK3에 나타내고 입력할 수 있는 체계와 다국어 통합정보사전의 개발을 지원하기 위해 다양한 코드체계를 하나의 시스템에서 수용하는 방법을 연구하는 것이 있다.

참고문헌

[1] 황도삼, 최용준, 조성래, 최기선 "웹 기반 언어자원 객체화에 근거한 통합정보사전 개발 시스템", 한국정보처리학회, 한국정보처리학회 추계 학술발표회 논문집, 1999년 10월.

[2] 최용준, 황도삼, 최기선, "YDK : 한국어 통합정보사전 개발시스템의 설계 및 구현", 한국정보과학회, 한국정보과학회 추계학술발표회논문집, 1998년 10월.

[3] 최용준, 황도삼, 최기선, "YDK-Term : 한국어용언의 다국어 통합정보사전", 한국인지과학회&한국정보과학회, 제10회 한글및 한국어 정보처리 학회 논문집, 1998년 10월.

[4] 황도삼외, "심층 국어정보처리 품질관리 체계", 영남대학교, 대용량 국어정보 심층처리 및 품질관리기술 개발 최종보고서, pp.56-68, 1998.

[5] 이재성 외3, "텍스트 및 전자사전 관리시스템의 설계", 한국정보과학회&한국인지과학회, 제8회 한국어 정보처리 학술대회 논문집, pp.408-414, 1996.

[6] 최병진 외3, "표준화를 위한 일반사전의 논리 구조", 한국정보과학회&한국인지과학회, 제8회 한국어 정보처리 학술대회 논문집, pp.415-423, 1996.

[7] 한국과학기술원, "텍스트코퍼스 및 전자사전 관리시스템(TDMS)", 과학기술처, 통합 국어정보베이스 최종보고서, pp.17-150, 1996.

[8] LADL, "INTEX", http://www.ladl.jussieu.fr/INTEX/index.html