웹 기반의 언어자원 객체화에 근거한 사전 개발 시스템

A Dictionary Constructing System based on a Web-based Object Model of Distributed Language Resources

황 도 삼*
(Dosam Hwang)

요 약 본 논문에서는 각기 다른 장소에 다양한 형태로 분산되어 있는 여러가지 언어자원들을 웹 기반에서 객체화시키는 모델을 제안한다. 웹 기반에서 객체화된 언어자원들은 다양한 응용 시스템 개발에 간단한 방법으로 이용되어 강력한 자연언어처리 응용 시스템을 구성할 수 있다. 또한, 초기 개발 이후에 이루어진 각 언어자원들의 개량은 별도의 처리과정 없이 자동으로 각 응용 시스템에 반영되므로 효과적인 유지보수가 가능하다는 장점이 있다. 제안한 모델의 적합성을 검증하기 위해 사전 개발 시스템 YDK2000를 설계하고 구현하였다. 개발한 YDK2000은 기존의 각종 사전의 여러가지 사전정보를 통합할 수 있을 뿐 아니라 여러 자연언어처리 시스템들과의 인터넷 접속을 통해 언어처리를 위한 사전정보를 손쉽게 통합할 수 있어 고품질의 사전을 개발할 수 있다.

키워드 전자사전, 자연언어처리, 한국어정보처리, 전자사전 개발 시스템

Abstract In this paper, we present a web-based object model of language resources that are distributed in different places in variable forms. Language resources organized as objects distributed over web sites can be easily utilized to produce application systems of natural language processing. So, it renders effective maintenance of overall language processing environment in that upgrading language resources can lead to the mechanical upgrading of application systems. We implemented a dictionary constructing system for Korean Language (YDK2000). This system can integrate various linguistic dictionaries, and also allow to construct high quality application specific dictionaries by connecting them to natural language systems on the Internet.

1. 서론

자연언어처리 시스템들은 대개 많은 양의 문법정보, 의미정보, 용례 등을 필요로 한다. 이러한 정보는 전자사전을 통해 제공되며, 제공되는 정보의 양과 질에 따라 프로그램의 성능도 영향을 많이 받는다[1].

전자사전을 구축하는 일은 노동집약적이면서도 전문지식을 필요로 하기 때문에 개발과정이 길고, 개발한 사전은 지속적으로 단어가 추가될 뿐만 아니라, 때로는 사전의 구조 자체도 변경되게 된다. 이러한 전자사전을 기반으로 하여 개발되는 자연언어처리 시스템들은 사전의 변화에 영향을 받게 되며, 필요로 하는 전자사전들을 시스템 내에 모두 설치해두어야 하는 문제점들을 갖고 있다[1-3]. 뿐만 아니라, 응용 시스템이 다른 자연언어처리 시스템들의 처리결과를 이용하고자 할 경우에는 별도의 복잡한 재개발과정을 거쳐 시스템 내에 그 시스템을 포함시켜 두어야만 한다. 이러한 과정을 거쳐 개발된다 하더라도 실제로 사용하기 위해서는 매우 큰 저장공간과 고성능의 시스템이 필요하므로 여러 면에서 비효율적이다. CORBA는 복잡한 재개발과정을 없애주는 특징이 있어 많이 이용되지만 별도의 미들웨어를 설치해야하며, 사용이 어렵다는 단점이 있다.

본 논문에서는 인터넷의 기본 서비스인 웹(World Wide Web) 환경에서 분산된 언어자원들을 객체화시키는 방법을 제안한다. 웹 기반에서 객체화된 언어자원

* 영남대학교 컴퓨터공학과
연구세부분야 : 한국어정보처리
Tel : 053-810-3515
FAX : 053-816-1976
본 연구는 KAIST KORTERM과 AITrc를 통하여 과학재단의 지원을 받았음.

들은 다양한 응용 시스템 개발에 간단한 방법으로 이용되어 강력한 자연언어처리 응용 시스템을 구성할 수 있다. 제안한 모델의 적합성을 검증하기 위해 한국어 통합정보사전 개발시스템(YDK2000이라 부름.)을 설계하고 구현하였다. YDK2000은 기존의 각종 사전의 다양한 사전정보를 통합할 수 있을 뿐만 아니라, 여러 자연언어처리 시스템들과 인터넷 접속을 통해 언어처리를 위한 사전정보를 손쉽게 통합할 수 있어 고품질의 전자사전을 개발할 수 있는 자연언어처리 응용 시스템이다.

2. 한국어 통합정보사전 개발 시스템

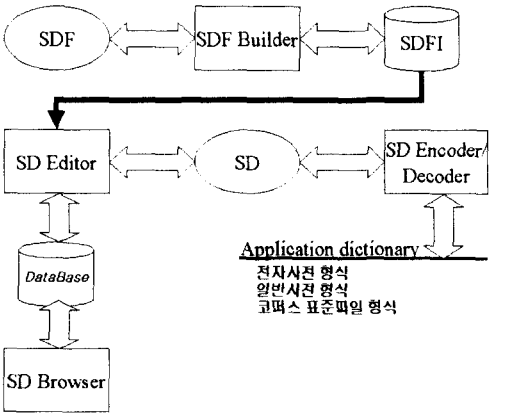
본 장에서는 대표적인 사전 개발 시스템인 KAIST의 TDMS(Text corpus and Dictionary Management System)에 대해 소개하고, 언어자원을 활용한 사전개발 시스템인 YDK98(1998년 개발, YDK1이라고도 함)에 대해 설명한다.

2.1 TDMS (Text corpus and Dictionary Management System)

지금까지 전자사전은 대부분의 경우 응용 시스템마다 각기 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 만들 때마다 별도의 사전관리 시스템을 만들어야 한다[1-6].

이러한 문제점들을 해결하고 각종 사전들의 표준 형태를 정의하여 표준 사전을 개발하고 관리하기 위한 도구로 KAIST에서 개발한 TDMS(Text corpus and Dictionary Management System)가 있다. TDMS는 SGML을 기반으로 하여 여러 분야에서 필요로 하는 각종 사전들의 표준 형태(SDF:Standard Dictionary Format)를 정의하고, 표준 사전(SD:Standard Dictionary)을 구축하는데 사용하는 시스템이며, SDF의 정의 및 SD의 편집, 수정, 검색, 변환할 수 있는 사전 및 텍스트 관리 통합시스템이다[3]. 이를 (그림 1)에 나타낸다.

그러나, TDMS는 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. 또한 TDMS는 (그림 1)과 같이 SDF Builder를 이용하여 SDF를 먼저 작성하여야 하고, 그 이후에 SD Editor를 사용하여야 하며, 사전의 검색을 위해서는 SD Browser를 사용해야 하므로, 그 구성과 사용법이 복잡하여 사용법을 숙지하는데 장기간이 소요되는 단점이 있다.



(그림 1) TDMS의 시스템 구성

따라서, 기계번역 시스템과의 연동을 통해 국내외의 다양한 사전들의 자료를 직접 참조할 수 있으며, 여러 자연언어처리 시스템의 처리 결과를 사전 구축 자료로 활용할 수 있는 사전 시스템의 개발이 필요하다. 또한, 이러한 기능과 함께 사용하기 쉬운 사용자 인터페이스를 가져야 하며, 실제 사전 개발에 이용할 수 있어야 한다.

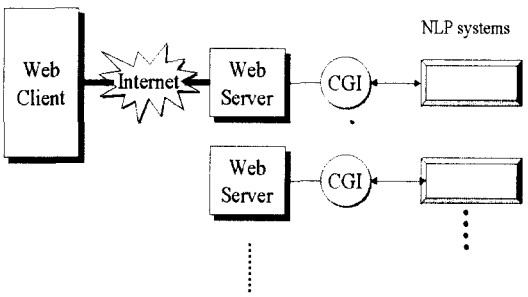
2.2 인터넷을 기반으로 한 분산 언어자원 통합 모델

지금까지, 자연언어처리 시스템을 비롯한 각종 사전 개발 시스템들은 각각의 목적에 맞추어 각기 다른 방법으로 개발되어 있으므로 특정 시스템의 처리 정보를 참조하기 위해서는 복잡한 절차를 거쳐야만 가능하다. 최악의 경우, 각 시스템의 운영체제나 운영환경의 차이점으로 인해 처리해야 할 정보 자체를 참조하지 못하게 될 수도 있다. 이는 동시에 여러 가지 시스템을 필요로 하는 다양한 자연언어처리 응용 시스템의 개발에 매우 부정적인 요소이며, 이를 해결하기 위해서는 자연언어처리 시스템의 개발과정과 처리 절차의 표준화를 통해 각 시스템들을 통합해야만 한다. 그러나, 이 방법은 현재까지 개발된 시스템들을 많은 시간과 노력을 들여 수정해야 하므로 현실성이 없다.

최근에 인터넷이 전세계적으로 급속히 보급되고 웹

이 인터넷의 기본 서비스로 자리잡음에 따라 웹을 기반으로 한 자연언어처리 시스템들이 개발되고 있지만, 이 시스템들 역시 호환성이 없어 다른 시스템에 재활용하는 것은 불가능하다.

YDK98에서는 인터넷을 기반으로 하고 웹 기술을 이용한 분산언어자원 통합모델을 제안하였다. 제안한 모델은 웹의 CGI(Common Gateway Interface)기술을 이용하여 기존 자연언어처리 시스템들 간의 통신기능만 부여한다. 이 방법은 현존하는 자연언어처리 시스템들의 대부분과 앞으로 개발되는 자연언어처리 응용 시스템들간의 정보교환 방법을 제공하므로 다양한 방면에 응용될 수 있다. (그림 2)에 분산 언어자원 통합을 위한 웹 기술인 CGI의 작동 개념을 보인다.

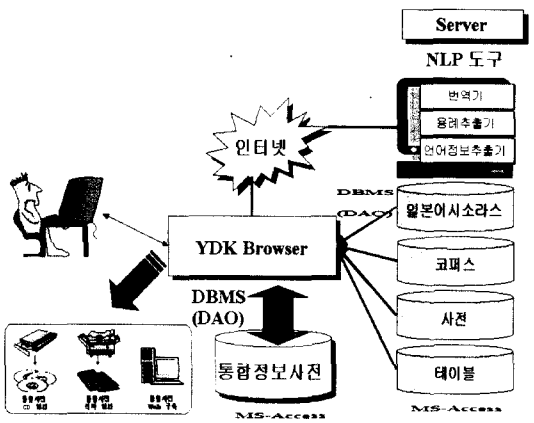


(그림 2) CGI기반 NLP시스템의 개념

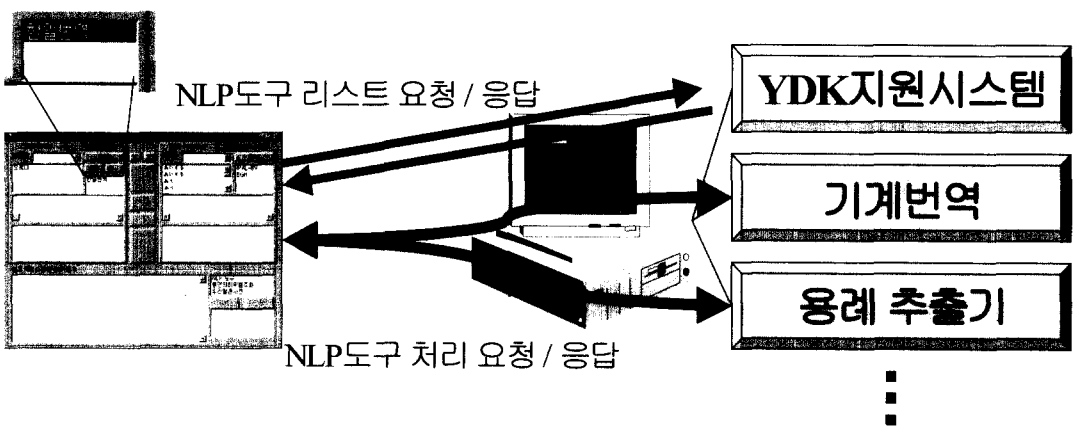
2.3 YDK1

YDK98(1998년 10월 version)은 단일 사용자를 위한 시스템으로 사전 정보는 사용자 시스템의 디스크에 저장되며, 자연언어처리 시스템들은 서버 시스템에 설치되어 있다. 사용자는 사전 정보를 검색하고자 할 때나 자연언어처리 시스템의 결과를 필요로 할 때는 간단한 버튼 조작만으로 결과를 얻을 수 있다. YDK98의 구성을 (그림 3)에 보이고, YDK98의 실행 예를 (그림 5)에 보인다[7].

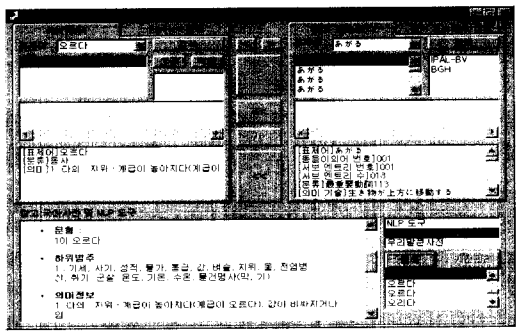
YDK98은 실행시에 YDK 서버시스템(nlp.yeungnam.ac.kr)에 접속하여, 등록된 자연언어처리 도구의 리스트를 얻어서 리스트 영역에 표시해 준다. 사용자의 자연언어처리 입력(버튼 클릭)을 받게되면 서버 시스템에 설치된 자연언어처리 도구에 접속하여 query를 넘겨주고, 그 처리 결과를 넘겨 받아 사용자에게 보여준다. 이를 (그림 4)에 나타내었다.



(그림 3) YDK98의 구성



(그림 4) YDK98의 시스템 처리 과정



(그림 5) YDK98의 실행 예

자연언어처리 시스템 등록 환경은 웹 사이트 형태로 개발되어 있으며, YDK98의 분산 언어자원 통합 모델에 맞도록 사전에 개량된 자연언어처리 시스템의 등록만 가능하다. 등록한 자연언어처리 시스템은 YDK98 지원 시스템의 데이터베이스에 그 정보가 입력되어 YDK98 실행시에 정보를 제공하게 된다.

3. 웹 기반의 언어자원 객체화 모델

2.3에서 설명한 YDK98은 사전에 입력할 자료를 NLP 시스템으로부터 직접 추출하기 위한 인터페이스를 가지고 있었으나, 입력된 사전 자료는 PC내에만 저장되어 사전구축 작업을 공동으로 수행할 수 없다는 점과 입력한 사전자료를 즉시 사용될 수 없다는 제한이 있어 사전자료 자체도 하나의 언어자원으로 처리할 수 있는 기법이 필요하다.

따라서, 본 장에서는 YDK98의 분산언어자원 통합모델을 전자사전을 포함한 NLP자원 전체에 걸쳐 확대한 웹 기반의 언어자원 객체화 모델을 제안하며, 인터넷 상에서 공동작업과 분산작업을 지원하기 위하여 구성한 Extranet형태의 사전개발 환경에 대해 설명한다.

3.1 웹 기반의 언어자원 객체화

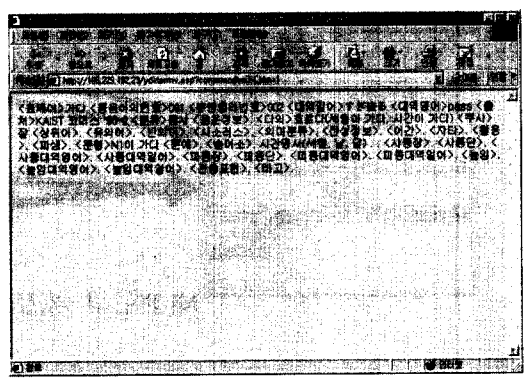
본 논문에서는 YDK98의 분산 언어자원 통합모델을 사전자료의 입력과 검색과정 전체에 확대한 웹 기반의 언어자원 객체화 모델을 제안한다. 즉, 웹의 특성을 YDK2000의 개발에 적용하기 위해서 현재까지 개발된 각종 전자사전들을 웹 환경에서 접근이 가능하도록 CGI 기술을 이용하여 약간의 수정을 적용하게 하며, 그 결과로 인터넷에서 각종 정보자원의 주소를 명시하는 표준화된 체계인 URL을 체계를 따르는 고유의 주소를 가지게 되므로, 용이하게 자원에 접근할 수 있게 된다. 실제로 전자사전을 이용하고자 하는 시스템들은 HTTP를 이용하여 자원의 주소에 접근하면, 별도의 시스템 설치과정을 거칠 필요 없이 바로 사전정보를 참조할 수 있다. 그리고, 각종 자연언어처리 도구들을 같은 방법으로 처리하면 분산되어 있는 언어자원들이 객체화되어 존재하게 된다. 자연언어처리 응용 시스템은 객체화된 언어자원을 하나의 부품으로 사용할 수 있게 되는 것이다. 이는 웹 브라우저에서 처리 결과를 바로 볼 수 있다는 점에 의의가 있는 것이 아니라, 다른 응용 프로그램에 간단히 통합될 수 있다는 점에 의의가 있다. 즉, 각종 응용프로그램에서 HTTP프로토콜만 지원하면 간단히 언어자원에 접근하여 정보를 얻을 수 있다.

객체화된 언어자원은 사용자의 query를 받아들여 해당하는 정보를 제공해야 하므로, 정보자원의 접근 method와 property가 필요하게 된다. 즉, 정보의 제공형식을 결정하는 method와 정보처리의 조건을 명시하는 property를 가짐으로써, 사용자가 원하는 정보를 즉각 제공할 수 있게 된다. 즉, 다음과 같은 형식의 URL로 표현된다.

http://호스트명/경로/aspfile?변수=value[&n=value2]
value : method + property, value2 : 단어식별번호

여기에서 n은 단어가 여러 개일 경우에 각 단어들을 식별하기 위한 번호이다. 이 형식에 따라 실제로 단어 "가다"에 대한 정보를 한국어 통합정보 사전(YDK-Term)에서 검색하기 위한 형식은 다음과 같이 표현되며, 웹 브라우저를 이용하여 해당 URL에 실제 접근한 결과를 (그림 6)에 보인다.

http://165.229.192.23/ydktermv.asp?command=a가다&n=1

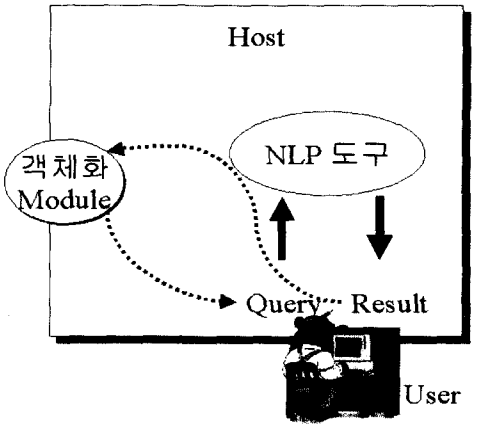


(그림 6) 객체화된 언어자원의 사용

객체화된 모델은 다른 시스템에서 사용할 수 있도록 하기 위해 시스템에 대한 접근 방법을 특정 호스트에 명시해 둘 필요가 있으며, 각 시스템들을 이용하게 되는 응용 시스템들은 이 호스트에 접근하여 해당 정보를 얻은 다음, 원하는 시스템으로 직접 접속하여 정보를 검색할 수 있다. 실제로, 본 논문에서는 '용언의 하위범주화사전', '우리말큰사전', 일본의 정보처리진흥사업협회기술센터의 계산기용 일본어 기본동사 사전(IPAL-BV)[9]과 기본 형용사 사전 IPAL-BA[10]등을 객체화시켜, YDK2 지원 시스템에 명시해두었으며, 이 자원들을 적절히 사용하여 통합정보사전 개발 시스템을 개발하였다. 본 연구에서 직접 객체화시킨 시스템들뿐 아니라, 국내외에서 개발된 다양한 자연언어처리 응용 시스템들도 이 형식에 맞추면 비교적 간단한 방법으로 시스템간의 정보교환이 가능하다.

3.2 언어자원 객체화 모듈

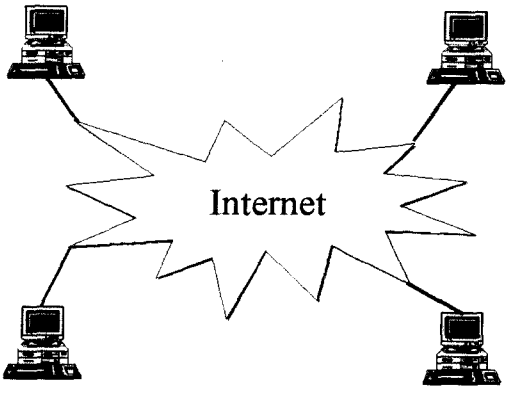
언어자원을 객체화시키기 위해서는 별도의 개발과정이 필요한 것이 아니라, 실제 시스템과 웹을 연결해주는 모듈만 작성하면 된다. 이 모듈은 각 개발자들이 CGI, ASP 등 적절한 서버측 응용 프로그램 실행 기술들을 이용하면 되며, 시스템 재개발은 필요하지 않다. 대부분의 자연언어처리 시스템은 몇 시간 정도의 모듈 작성 시간만으로 충분히 객체화될 수 있다. 객체화 모듈의 작동원리를 (그림 7)에 보인다. 대부분의 언어자원 즉, 자연언어처리 도구들은 사용자의 조작에 의해 실행되는 시스템인데, 객체화 모듈은 사용자의 조작을 대신 처리해주는 일종의 중계기 역할을 수행한다.



(그림 7) 객체화 모듈

3.3 익스트라넷(Extranet)

익스트라넷은 인터넷 기술을 이용한 기업간 정보 인프라이며, 기업간에만 적용되는 것이 아니라 업무상 파트너인 거래처, 제휴기업, 관련 회사, 판매점, 단골 고객등을 모두 포함한다. 즉, 인터넷을 정보를 주고받기 위한 하나의 통신매체로 사용하는 것이다. 이것은 인터넷의 기본 기술을 이용하여 원격지간 통신을 가능하게 하는 것으로 공동작업 및 분산작업이 가능하다는 장점이 있다. 본 논문에서 제안한 언어자원 객체화 기법은 익스트라넷 형태의 사전개발환경을 가능하게 하는 기술이며, 실제 구현된 시스템 YDK2000은 익스트라넷 환경에서 작동하는 응용 시스템이다. 인터넷을 인프라로 사용하는 전형적인 익스트라넷 시스템을 (그림 8)에 보인다.



(그림 8) Extranet

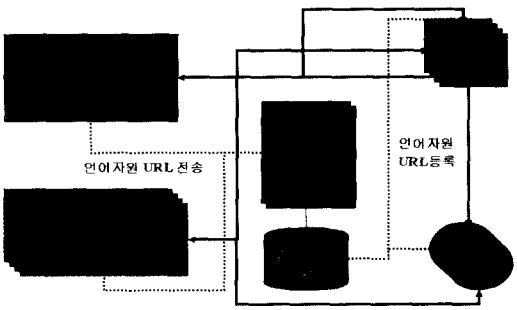
4. YDK2000의 설계 및 구현

YDK2000은 인터넷 환경에서의 언어자원 객체화 및 객체화된 언어자원을 이용한 통합정보사전 개발 시스템이다. 본 연구에서는 부산대의 '용언의 하위범주화사전', '우리말큰사전', 일본의 정보처리진흥사업협회기술센터의 IPAL-BV, IPAL-BV등을 객체화시켰으며, YDK2000으로 개발한 한국어통합사전(YDK-Term)도 객체화시켰다.

4.1 YDK2000의 설계

YDK2000(이하 YDK라 함.)는 YDK 서버 시스템과 YDK 브라우저로 구성된다. YDK 서버 시스템은 객체화된 언어자원들의 목록을 유지하여 YDK 브라우저에게 접근할 수 있는 방법을 제공하여 주며, YDK 브라

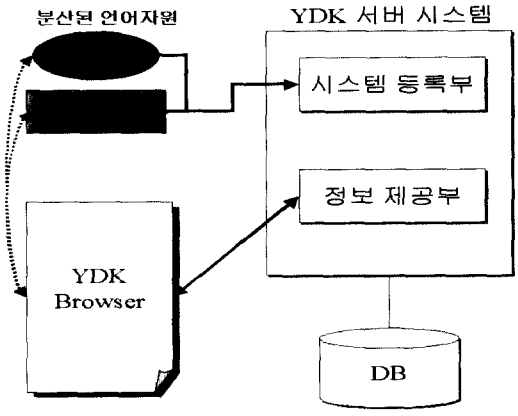
우저는 각종 전자사전 및 자연언어처리 도구들을 참조하여 사전을 개발할 수 있게 한다. 이 과정에서 개발된 YDK 서버 시스템은 YDK 브라우저뿐만 아니라 다양한 자연언어처리(이후 NLP라 함) 응용시스템의 개발에도 사용할 수 있는 다목적 시스템이다. 전체적인 시스템 구성도를 (그림 9)에 나타낸다.



(그림 9) YDK2 시스템 구성도

(1) YDK 서버 시스템

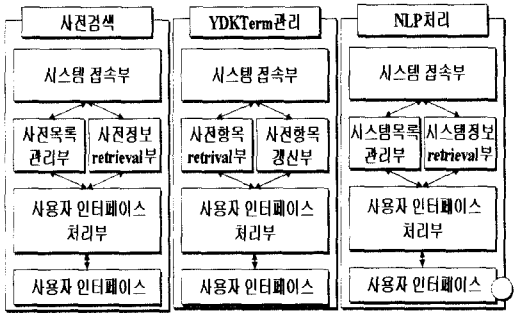
YDK 서버 시스템은 아래 그림과 같이 시스템 등록부와 등록 정보 제공부로 구성되어 있다. 분산된 언어자원들은 시스템 등록부에 접속하여 자신의 URL과 정보를 검색하기 위한 method와 property를 등록한다. 이 작업만으로 YDK 브라우저가 이 언어자원의 주소를 획득하게 되며, 이후에는 YDK 브라우저가 해당 언어자원에 직접 접속하게 된다. 이 화면을 아래 (그림 10)에 나타내었다.



(그림 10) YDK 서버 시스템 구성도

(2) YDK 브라우저

YDK 브라우저는 컴포넌트 형태로 개발된 인터페이스와 각 컴포넌트를 연결하는 모듈로 구성되어 있다. YDK 브라우저에서 처리해야 하는 것은 사전을 검색하는 기능과 NLP 시스템과의 접속 기능 및 YDK-Term사전을 개발하는 기능인데, 이들이 각각 컴포넌트로 개발되어 있다. 따라서 (그림 11)과 같이 YDK 브라우저에는 사전검색 컴포넌트와 NLP처리 컴포넌트 및 YDK-Term 관리 컴포넌트로 구성되어 있다.



(그림 11) YDK 브라우저의 컴포넌트

사전검색부는 YDK 서버 시스템에 접속하여 사전의 목록을 수신하여 사용자에게 제공해 주고, 사용자의 검색요구를 서버에 전달하여 처리한 후에, 그 결과를 수신하여 사용자에게 보여 준다. 이 컴포넌트는 특별한 조건을 제시하면 YDK-Term에 대해서만 정보를 보여주고 사전자료 편집기능을 제공하는 형태로 변환되도록 되어 있다. 따라서 YDK-Term 편집시에는 사전정보 검색부가 편집기능을 겸한다. YDK-Term 관리부는 사전항목을 편집하게 하는 것으로 YDK-Term에 대해서만 작동하는 기능으로 Client/Server 모델이다. NLP처리부는 2가지의 작동모드를 갖는다. 한가지는 NLP 처리 결과를 사전편집에 사용할 수 있도록 외부에서의 처리요구를 수용해 주는 것으로, 기계번역 시스템과의 접속을 이용한 자동번역을 예로 들 수 있다. 또 한가지는 아직까지 자동으로 처리할 수 없는 정보가 많아 사용자가 직접 수작업으로 정보를 참조하고자 할 때, 웹 페이지 형식의 정보를 사용자에게 제공하는 기능이다. YDK 브라우저는 이 3가지의 컴포넌트를 상호 연동시켜 작동한다.

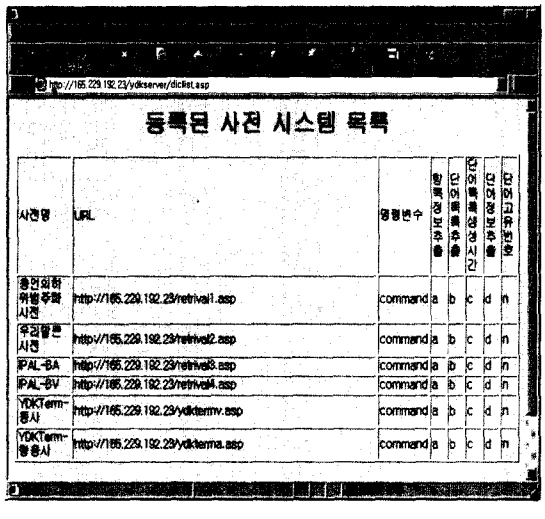
4.2 YDK2000의 구현

(1) 구현환경

YDK2000은 Intel Pentium PC, MS-Windows98 환경에서 MS-Visual BASIC 6.0을 이용하여 개발하였으며, 서버에 설치되는 YDK 서버 시스템은 MS-WindowsNT, IIS 4.0환경에서 ASP(Active Server Page)를 이용하여 개발하였다.

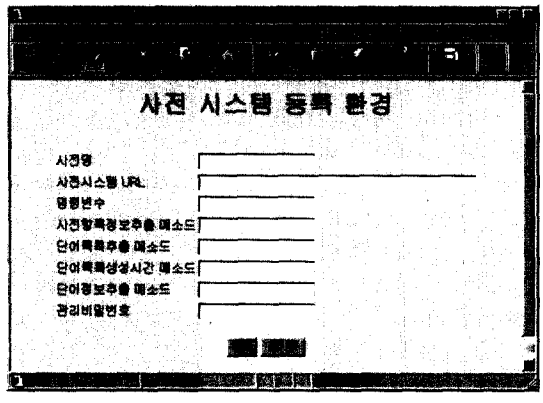
(2) YDK2000

YDK2000은 다수의 사용자가 동시에 사용할 수 있는 사전개발시스템으로 모든 정보는 서버시스템에 보관되고 브라우저 자체에는 아무런 정보가 없다. 사용자는 사전정보를 검색하고자 할 때나 자연언어처리 시스템의 결과를 필요로 할 때, 간단한 입력과 버튼 조작만으로 결과를 얻을 수 있다. 실제로 구현한 YDK 서버 시스템을 (그림 12)와 (그림 13)에 보이고, YDK2000 브라우저의 실행 예를 (그림 14)에 보인다.

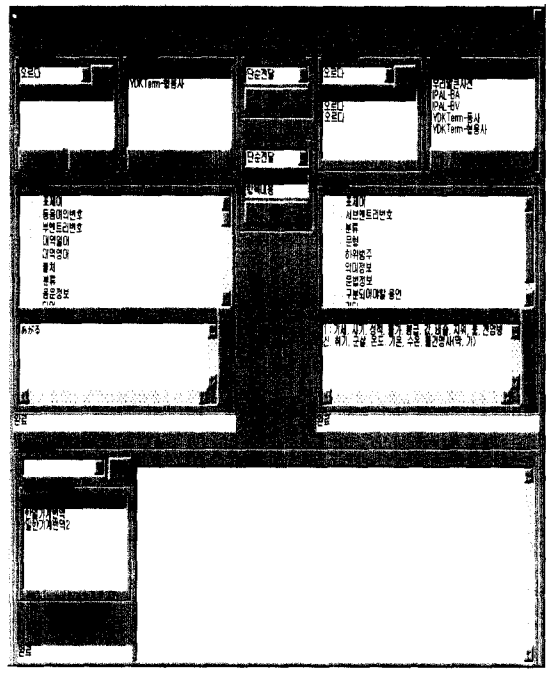


(그림 12) YDK 서버 시스템

YDK2000은 실행전에 반드시 YDK 서버시스템에 접속하여 등록된 언어자원의 리스트를 수신하여야 하며, 수신한 결과는 목록창에 표시해준다. 사용자의 각종 입력조작은 각 컴포넌트에 포함되어 있는 접속부에 의해 자동으로 각 시스템들과 통신하게 되며 처리결과를 넘겨받아 사용자에게 보여준다.



(그림 13) 사전시스템 등록 환경



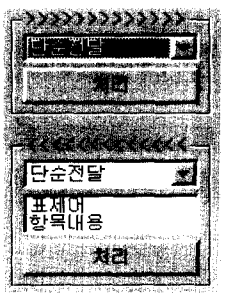
(그림 14) YDK2000 브라우저

4.3 사전간의 자료교환 지원 환경

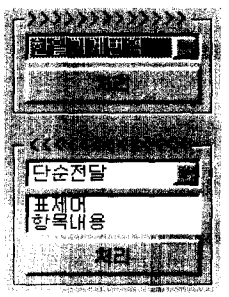
YDK2000은 사전의 설계, 편집 등 사전개발 시스템이 가져야 할 기본 기능 외에, 여러 다양한 사전들의 정보를 참조할 수 있으며, 사전의 정보만으로 부족할 경우에는 자연언어처리 시스템과도 연결될 수 있다. 기계번역 시스템과 용례 추출기 및 각종 언어정보 추출 시스템들을 연결하여 다양한 정보를 직접 참고할

수 있으며, 자동화된 작업 수행을 위해 사전간의 자료 교환시에 기계번역 시스템에 접근하여 그 처리 결과를 수신하여 상대편 편집창에 직접 전송할 수도 있다.

YDK2000은 사전을 개발할 때, 다른 사전의 자료를 최대한 참조할 수 있다. (그림 15)와 (그림 16)은 YDK 브라우저의 중앙에 위치한 사전자료 교환지원 기능의 예이다. (그림 15)는 YDK-Term의 표제어를 사전검색 영역의 표제어로 전송한다. 만약 대상 사전이 외국어 사전이라면 기계번역 시스템이 필요하게 되는데, 이 경우는 (그림 14)의 좌하 부분과 같이 처리할 NLP도구를 지정해 주면 된다. 다른 사전의 자료를 YDK-Term에 반영하는 경우는 2가지로 볼 수 있다. 하나는 타 사전의 표제어를 YDK-Term에서 찾고자 할 때이며, 다른 하나는 타 사전의 항목정보를 YDK-Term으로 가지고 올 경우이다. 이를 지원하고자 하는 것이 (그림 15)와 (그림 16)의 하단에 있는 영역이다. 표제어와 항목내용으로 나누어 처리할 NLP 도구를 선택할 수 있게 되어 있는데, 여기에 나타나는 도구 목록들은 NLP 도구 접속처리 영역에 나타나는 것과 동일하며, 실제로 NLP 도구 접속처리 영역의 기능을 이용하여 NLP 도구에 접속하게 된다.



(그림 15) 사전간의 자료교환 I



(그림 16) 사전간의 자료교환 II

4.4 결과 및 고찰

본 논문에서 제안한 웹 기반의 언어자원 객체화는 범용화된 인터넷을 기반으로 하여 각종 언어자원들의 활용성을 높일 수 있다. 특히 YDK2000은 한국어 용언의 다국어 통합정보사전인 YDK-Term개발에 활용되고 있다. YDK2000은 사전의 설계, 편집 등 사전 개발 시스템이 가져야 할 기본기능을 가지고 있으며, 여러 다양한 사전들의 정보를 참조하고, 자연언어처리 시스템과의 연결을 통한 자료 수집기능을 가지고 있어, 이미 개발되어 있는 대부분의 언어자원들을 간단한 방법으로 활용할 수 있다는 특징이 있다. 특히, YDK2000의 자연언어처리 시스템과의 연계부분은 보다 쉽게 자연언어처리 시스템들을 이용하게 함으로써 사전정보의 정확성을 높이는 데 결정적 역할을 하고 있다. 본 시스템은 사전자료를 모으고 입력하는 작업을 편리하게 처리하므로 사전의 개발 작업을 용이하게 하는 장점이 있다.

본 시스템 개발과 함께 연구된 한국어 통합정보 사전(YDK-Term)의 사전항목을 〈표 1〉에 보인다. 본 논문에서는 사전개발 시스템 YDK2000에 대해서만 기술하며, YDK-Term에 대한 설명은 참고문헌 [1]과 [8]을 참조하기 바란다.

〈표 1〉 YDK-Term 사전항목

항목	부항목			항목	부항목		
1. 표제어 정보	표제어			3. 형태 정보	어원		
	항목작성자				전성형		
	표제어번호				이형태		
	동음이의번호				약어		
	부엔트리번호				어간		
	항목수정일				자/타		
	작성일자				활용		
	빈도정보				파생		
	외래어표기				높임		
	대역정보	일어			방언형	경상	
		영어				전라	
		중국어				:	
		독일어		4. 통사 정보	문형		
		불어			문예		
	출처				술어소		
	분류				의미소	N1	
	음운정보					N2	
	전문어표시					N3	
	메모			5. 문법 정보	사동	장	
2. 의미 정보	의미기술	다의				단	
		부사				대역	영어
	관련어	상의어					일어
		유의어			피동	장	
		반의어				단	
		전체어				대역	영어
		부분어					일어
		연상어		6. 기타 정보			
	시소러스						
	의미분류						
	숙어정보	숙어					
		풀이					

5. 결론

대부분의 자연언어처리 시스템들이 필요로 하는 전자사전은 많은 양의 문법정보, 의미정보, 용례 등의 정보를 제공하며, 제공하는 정보의 양과 질에 따라 프로그램의 성능도 영향을 많이 받게 된다. 이러한 전자사전을 구축하는 일은 매우 노동집약적이면서도 전문지식을 필요로 하기 때문에 개발과정이 오래 걸리며, 개발한 사전의 유지보수 자체가 매우 어렵다는 단점이 있다. 또한, 이러한 전자사전을 기반으로 하여 개발되는 자연언어처리 시스템들은 사전의 변화에 매우 민감하게 되며, 필요로 하는 전자사전들을 시스템 내에 모두 설치해두어야 하기 때문에 새로운 응용 시스템 개발에 많은 제약을 받고 있다는 것 외에, 응용 시스템이 다른 자연언어처리 시스템들의 처리결과를 이용하기가 매우 어렵다는 단점이 있다.

본 연구에서는 전자사전을 보다 효과적으로 구축하기 위해 각종 자연언어처리 시스템의 처리결과를 이용할 수 있는 언어자원 객체화 모델을 제안하였으며, 실제 모델을 구체화시킨 통합정보사전 개발 시스템을 설계하고 구현하였다. 본 모델은 복잡한 미들웨어를 사용하지 않고 인터넷의 기본 서비스인 웹(World Wide Web) 환경을 이용하여 분산된 언어자원들을 객체화시키는 것으로 웹 기반에서 객체화된 언어자원들은 다양한 응용 시스템 개발에 간단한 방법으로 이용되어 강력한 자연언어처리 응용 시스템을 구성할 수 있게 된다. 또한, 이 과정에 있어 별도의 추가 개발이나 구축과정이 거의 없다는 장점이 있으며, 구축한 사전도 하나의 언어자원으로 타 시스템에 정보를 제공할 수 있게 된다는 장점이 있다. 본 연구에서는 실제로 이미 개발되어 분산되어 있는 언어자원들을 웹기반에서 객체화시켰으며, 이를 이용하고자 하는 자연언어처리 응용 시스템들은 HTTP를 이용하여 자원의 주소에 접근하도록 하면, 별도의 시스템 설치과정을 거칠 필요 없이 바로 사전정보를 참조할 수 있게 된다. 이는 언어자원이 하나의 부품으로 사용되므로 효과적으로 응용프로그램을 개발할 수 있다는 장점이 있다. 또한, 본 연구에서 개발한 통합정보사전 개발 시스템인 YDK2000은 분산되어 있는 각종 사전의 정보를 통합할 수 있을 뿐 아니라 여러 자연언어처리 시스템들과의 접속을 통해 다양한 정보를 손쉽게 통합할 수 있어 고품질의 전자사전을 개발할 수 있다는 장점이 있으며, 동시에 여러명이 사전편집 작업에 참여할 수 있어 사전개발 시간을 단축시킬 수 있다. 추후 연구과제로는 각 시스템을 사용하는 권한을 지정하도록 하여 비 인가자가 자원을 사용할 수 없게 하는 방법의 연구를 비롯하여, 보다 다양한 자연언어처리 시스템들을 객체화시키는 것 등이 남아 있다.

참 고 문 헌

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

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

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

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

[5] 황도삼 외2, "자연언어처리", 홍릉과학출판사, 1998.

[6] 황도삼 외2, "자연언어이해", 홍릉과학출판사, 1999.

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

[8] 최용준, 황도삼, 최기선, "YDK-Term : 한국어용언의 다국어 통합정보사전", 한국인지과학회&한국정보과학회, 제10회 한글 및 한국어 정보처리학회 논문집, 1998년 10월.

[9] 技術センター, "算機用日本語動詞辭典IPAL (Basic Verbs)", 일본 정보처리 진흥사업협회, 1987년 3월.

[10] 技術センター, "計算機用日本語動詞辭典IPAL(Basic Adjectives)", 일본 정보처리 진흥사업협회, 1990년 7월.

[11] 부산대학교, "한국어 문장 분석을 위한 용언의 하위범주화에 관한 연구", 시스템공학연구소 최종보고서, 1997.

[12] 大野平, 竝書淨人, "角川 類語新辭典", 角川書店, 1980.

[13] 岩波書店, "日本語語彙大系", 日本電信電話株式會社, 1997.