한국어 통합정보사전 시스템

황 도 삼[†] · 최 기 선[††]

요 약

사전은 각종 자연언어처리 시스템에 있어서 고도의 언어처리 및 성능향상을 위한 필수 요소이며, 아무리 좋은 언어처리 도구와 알고리즘이라도 계산언어학에 근거한 양질의 체계적인 전자사전이 없는 한 이의 실용화는 불가능하다. 기존에 출판된 일반 사전은 자연언어처리 및 이해를 목적으로 하여 개발된 사전이 아니다. 또한, 자연언어처리 도구 및 응용시스템을 위해 개발된 사전은 각 시스템의 목적에 따라 각기 다른 체계에 의해 구축되어 있기 때문에 이용하는데 있어서 비효율적인 점이 있다. 따라서, 고도의 언어처리 및 이해를 목적으로 한 체계적이고 과학적인 방법론을 이용하여 형태소, 구문, 의미정보 등 각종 정보가 통합된 통합정보사전의 개발이 필요하다.

본 논문에서는 통합정보사전을 구축하기 위한 방법론을 제시하고, 이에 근거하여 개발한 통합정보사전 개발 시스템을 제시한다.

YDK : A Thesaurus Developing System for Korean Language

Do-Sam Hwang[†] · Key-Sun Choi[††]

ABSTRACT

Dictionaries are indispensable for NLP(natural language processing) systems. Sophisticated algorithms in the NLP systems can be fully appreciated only with matching dictionaries that are built systematically based on computational linguistics. Only few dictionaries are developed for natural language processing. Available dictionaries are far from complete specifications for practical uses. So, it is necessary to develop an integrated information dictionary that includes useful lexical information for processing and understanding natural languages such as morphology and syntactic and semantic information.

In this paper, we propose a method to build an integrated dictionary, and introduce a dictionary developing system.

1. 서 론

자연언어처리 도구 및 응용시스템을 위해 개발되어 있는 기존의 전자사전은 용도에 따라 각기 다른 체계에 의해 구축되어 있으므로 전자사전 간의 호환성이 부족하다. 또한, 이러한 전자사전은 대부분 기계가독형 체계로 만들어져 있으므로 사람이 사용할 수 있는 형태로 출판하기 어렵다. 또한, 어떤 단어에 대해 기술되어 있지 않은, 보다 더 상세한 정보를 얻기 위해서는 또다른 여러 사전들을 참조해야 한다는 불편함이 있을 뿐만 아니라, 이러한 사전들을 입수하는 것도 쉽지 않다. 따라서, 고도의 언어처리 및 이해를 목적으로 한 체계적이며 과학적인 새로운 방법론을 이용하여 형태소정보, 구문정보, 의미정보 등 각종 정보가 통합된 통합정보사전의 개발이 필요하다[6-11].

본 논문에서는 통합정보사전 구축을 위한 사전구축 방법론을 정립하고, 정립된 방법론을 바탕으로 한 통합정보사전 개발 시스템을 제시한다. 이 사전은 특정 분야에 맞춘 전문 사전이 아니며, 기존 사전의 정보를 통합하고 이미 개발되어 있는 여러 자연언어처리 도구

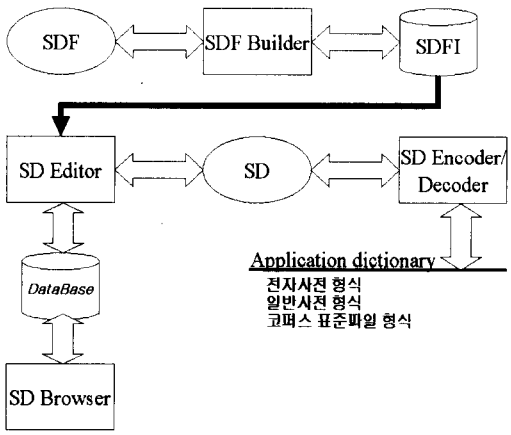
※ 본 연구는 KAIST KORTERM을 통하여 과기부의 지원과 첨단 정보기술 연구센터를 통하여 과학재단의 지원을 받았음.
† 정 회 원 : 영남대학교 전자정보공학부 교수
†† 종신회원 : 한국과학기술원 전문용어언어공학연구센터 소장
논문접수 : 1999년 5월 12일, 심사완료 : 2000년 8월 29일

를 이용하여 사전 정보를 얻을 수 있도록 하였으며, 각종 정보를 통합한 사전으로 국어정보처리 응용시스템 개발에 도움을 줄 수 있다. 또한, 통합정보사전 개발 시스템은 각종 전자사전의 정보와 여러 자연언어처리 시스템들로부터 추출한 정보를 손쉽게 통합할 수 있게 함으로써 고품질의 사전 개발을 가능하게 한다.

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[1-2, 4-5].

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)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 통합시스템이다[3]. 이를 (그림 1)에 나타낸다.



(그림 1) TDMS(Text corpus and Dictionary Management System)의 시스템 구성

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

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

3. 사전개발 시스템의 설계

본장에서는 사전을 효과적으로 개발하기 위한 사전 시스템의 설계에 관해 기술한다.

3.1 사전 구축 방법론

통합정보사전을 개발하기 위한 방법으로 수작업 방식과 반자동 방식이 있다. 수작업은 사람의 손으로 직접 모든 사전 자료를 수집하고 입력하여 구축하는 방법으로 많은 노력과 시간 및 비용이 든다. 또한 사전정보를 입력하는 사람의 능력에 의해 사전의 품질이 결정되므로 객관성이 떨어진다는 단점이 있다. 이에 비해 반자동 방식은 사전개발 시스템을 이용하여 구축하는 방법으로 많은 양의 데이터를 빠른 시간에 입력할 수 있을 뿐 아니라, 사전의 원시 정보를 가지고 있는 코퍼스나 전자사전과 같은 정보원으로 부터 사전구축에 필요한 정보들을 자동적으로 추출할 수 있으므로 실질적인 사전정보를 가진 고품질의 사전을 개발할 수 있다. 본 연구에서는 반자동 방식으로 통합정보사전을 개발하기 위해, 통합정보사전 개발 시스템인 YDK[1]를

1) YDK : 본 논문에서 개발한 사전 개발 시스템의 이름이다. YDK는 Yeungnam university Dictionary Kucuk(construction) system 또는 개발자의 이름인 Yongjun Dosam Keysun의 뜻을 가지고 있다.

개발하였다(그림 3). YDK 시스템은 각종 사전정보의 참조 및 자연언어처리 도구들과의 통신을 통한 정보추출기능을 가지고 있어 효과적으로 사전을 개발할 수 있는 도구이다.

3.2 통합정보사전 개발 시스템의 기능

통합정보사전 개발 시스템은 사전을 개발하고, 사전자료를 입력하는 모든 작업을 수행할 수 있어야 하므로, 1존의 다른 사전들의 정보를 직접·간접적으로 참조할 수 있어야 한다. 이러한 다른 사전 참조 기능은 이미 개발되어 있는 전자사전을 데이터베이스 차원에서 시스템에 연결시키면 비교적 간단히 구현할 수 있지만, 참조하는 사전이 한국어 사전이 아닐 경우에는 번역을 해야하는 문제가 발생한다. 해당 사전 전체를 미리 번역시켜 연결시킬 경우에는 번역 결과에 대해 검증을 해야하는 문제가 발생하며, 다양한 사전들을 모두 다 번역시켜야 하는 단점이 있다. 이를 해결하기 위해서는 사전 참조 기능과 기계번역 시스템을 연동시켜 필요시에만 번역 결과를 도출하는 방법의 개발이 필요하다. 또한 다른 사전에 포함된 자료가 부족하거나 새로운 자료를 추가하기 위해서는 다양한 자연언어처리 시스템들과의 연동을 통해 부가적인 정보를 얻을 수 있어야 한다.

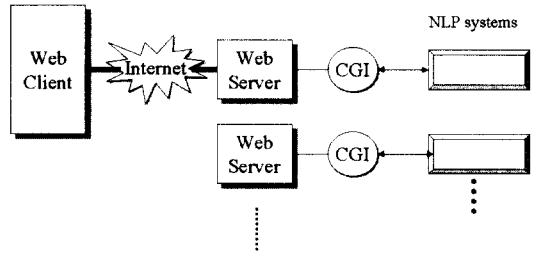
통합정보사전 개발 시스템은 이러한 특성으로 인해 사전을 설계하고 자료를 입력하는 기본 기능 외에 다른 사전을 참조하고, 경우에 따라서 자연언어처리 시스템과 연동할 수 있는 기능이 필요하다.

3.3 인터넷을 기반으로 한 분산 언어자원 통합 모델

지금까지의 자연언어처리 시스템을 비롯한 각종 사전 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 처리정보 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. 이 방법은 현재까지 개발된 시스템들을 많은 시간과 노력을 들여 수정해야 하므로 현실성이 없다.

최근에는 인터넷이 전세계적으로 급속히 보급되고 웹이 인터넷의 기본 서비스로 자리잡음에 따라 웹을 기반으로 한 자연언어처리 시스템들이 개발되고 있지만, 이 시스템들 역시 호환성이 거의 없어 다른 시스템에 재활용하는 것은 거의 불가능하다.

본 논문에서는 인터넷을 기반으로 하고 웹기술을 이용한 분산언어자원 통합모델을 제안한다. 제안하는 모델은 웹의 CGI(Common Gateway Interface)기술을 이용하여 기존 자연언어처리 시스템들 간의 통신기능만 부여한다. 이 방법은 현존하는 대부분의 자연언어 처리 시스템들과 앞으로 개발되는 자연언어처리 응용 시스템들 간의 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

YDK 시스템은 다른 사전을 참조하기 위해 사전을 검색할 수 있어야 하며, 참조하는 사전에 외국어 사전도 있으므로, 이를 번역시스템과 연계하여 번역결과를 얻어낼 수 있어야 한다. 또한 자연언어처리 도구들로부터 다양한 정보를 얻기 위해서는 하나의 인터페이스로 통합하여야 하는데, 바로 여기에 본 논문에서 제안한 분산 언어자원 통합 모델을 적용하였다. 즉, 사전개발에 필요한 자연언어 자원들을 이용할 수 있도록 지원하기 위하여 전자사전과 자연언어처리 도구들을 하나의 인터페이스 내에 통합하는 것이 YDK 설계의 기본 개념이다.

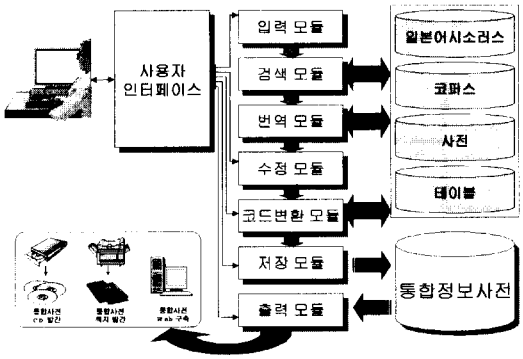
3.4 YDK 설계

3.4.1 YDK 개발 개념

통합정보사전을 구축하기 위해서는 가능한 많은 정보들을 참조해야 한다. 이를 위해서는 현재까지 개발된 여러 전자사전들의 정보를 참조할 수 있어야 하며, 여러 자연언어처리 시스템들의 처리 결과도 참조할 수

있어야 한다. 이와 같이 필요하다고 생각하는 모든 자원들을 하나의 플랫폼에 통합시키자는 것이 YDK의 개발개념이다.

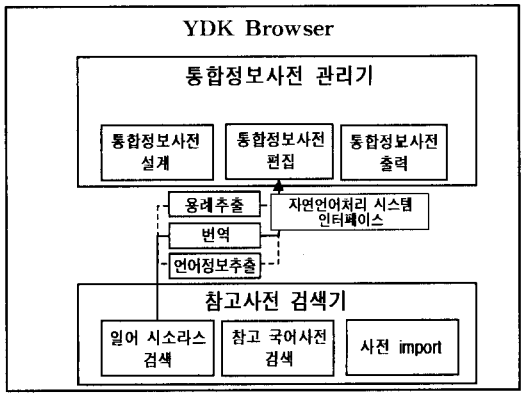
이를 구현하기 위해서는 전자사전, 기계번역 시스템 등 분산된 언어자원들을 하나의 플랫폼으로 통합하기 위한 통합모델이 필요하다. 제안하는 통합모델은 각 자원들의 고유한 작동 플랫폼을 그대로 둔 상태에서 모듈의 추가만으로 간단히 YDK에 연결되어 사전개발 환경을 구성하게 된다. 즉, 분산되어 있는 자연언어처리 자원들이 각각의 자원에 적합한 모듈로 개발되어 있으며, 각 모듈들의 인터페이스를 표준화시켜 통합하는 체계를 따른다. 이를 (그림 3)에 보인다.



(그림 3) YDK 시스템의 작동 개념

3.4.2 YDK의 구성

YDK는 [통합정보사전 관리기]와 [참고사전 검색기] 및 [자연언어처리 시스템 인터페이스]로 구성된다, 이를 (그림 4)에 나타낸다.



(그림 4) YDK 시스템 구성도

(1) 통합정보사전 관리기

통합정보사전을 설계하고 자료를 입력하는 기능을 가진다. 사전 항목을 정의하거나 변경할 수 있으며, 항목 자료를 입력할 수 있어야 한다. 사전은 내부적으로는 데이터베이스를 기반으로 하여 작동한다. 사전을 구성하는 데이터베이스는 사전항목의 정보를 저장하는 데이터베이스와 사전 자료 자체를 저장하는 데이터베이스로 구성되며, 필요할 경우에는 사전 자료 입력 중이라도 사전 구조 자체를 수정할 수 있어야 한다. 사전의 구조를 작성하고 변경하는 '통합정보사전 설계' 기능과 사전 자료를 입력하고 편집하는 '통합정보사전 편집' 기능으로 구성된다.

입력이 완료된 사전은 인쇄 매체를 통해 출판할 수 있어야 하며, 타 사전시스템과의 자료 호환을 위해 SGML형식의 파일의 형태로도 출력할 수 있어야 한다. 이를 위해 '통합정보사전 출력' 기능을 가지고 있다.

(2) 참고사전 검색기

현존하고 있는 여러 전자사전들을 참조할 수 있는 기능으로서 YDK에 적합한 형태의 데이터베이스 형태로 변환된 전자사전들만을 연결할 수 있는데, 이미 변환되어 있는 사전을 시스템에 등록하는 기능으로는 '사전 import'가 있다. '사전 import'를 통해 사전을 등록하고, 검색하는 방법에 대해 정의한다. 확장성을 가지려면 임의의 사전을 가지고 와서 변환할 수 있는 기능이 필요하지만 현재의 시스템에서는 고려하지 않았다.

한국어 사전들을 참조할 때 적용되는 기능은 '참고 국어사전 검색' 기능인데, 검색한 표제어에 대한 상세한 정보를 출력한다. 사용자는 이 정보를 참고하여 사전편집 작업을 할 수 있다.

일어 시소러스들을 참고하기 위한 '일어 시소러스 검색 기능은 KS 코드로 변환된 사전들을 검색할 수 있는 시스템이다. 일어 사전들의 각 항목은 '자연언어처리 시스템 인터페이스'를 통해 기계번역 시스템에 연결되어 온라인으로 번역시킬 수 있다. 사전 개발자는 번역된 내용을 참고할 수 있으며, 필요에 따라 번역하지 않은 내용 자체를 참고할 수도 있다.

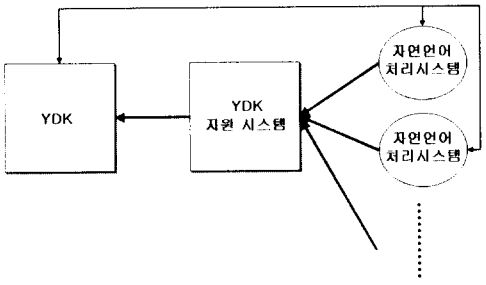
(3) 자연언어처리 시스템 인터페이스

자연언어처리 시스템들로부터의 정보 획득을 위한 처리부로서 서버 시스템에 접속하여 query를 넘겨주고 처리 결과를 돌려 받는 역할을 한다. 통합정보사전의 특성상 여러 자연언어처리 시스템들의 처리 결과를 수

집하게 되는데, 자연언어처리 시스템 인터페이스는 query로 구성하여 서버에 넘겨준 다음, 그 처리 결과를 출력 영역을 이용하여 사용자에게 보여준다. 자연언어처리 시스템들로부터 처리결과를 얻기 위해서는 사전에 각 자연언어처리 시스템들이 YDK에 연결할 수 있는 형태로 구성되어 있어야 하는데, 이를 위해서는 YDK 분산 언어자원 통합 모델을 따르도록 특별한 모듈을 개발하여 적용시켜야 한다. YDK 분산 언어자원 통합 모델을 따르는 시스템들은 YDK 지원 서버에 설치된 YDK 지원 시스템에 등록하여야 하며, 이 과정은 웹을 통해 이루어진다. YDK 지원 시스템은 자연언어처리 시스템들의 목록을 보관하고 있으며, YDK가 접속해 오면 그 목록을 넘겨준다. YDK는 실행시에 YDK 지원 시스템과 통신하여 사용할 수 있는 자연언어처리 시스템들의 목록을 넘겨받으며, 이 시스템들을 YDK환경에서 실제 사용할 수 있도록 구성한다.

(4) YDK 지원 시스템

YDK 지원 시스템은 자연언어처리 시스템들의 목록을 보관하고 있으며, YDK가 접속해 오면 그 목록을 넘겨주는 역할을 수행한다. 각 자연언어처리 시스템들은 웹을 통해 자신의 정보를 등록하며, URL(Uniform Resource Locators)형식으로 자신의 시스템을 등록하면 된다. 이를 (그림 5)에 나타낸다.



(그림 5) YDK 지원 시스템의 작동원리

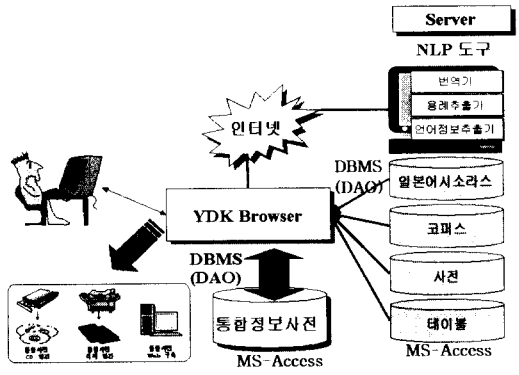
4. YDK의 구현

4.1 YDK의 구현 환경

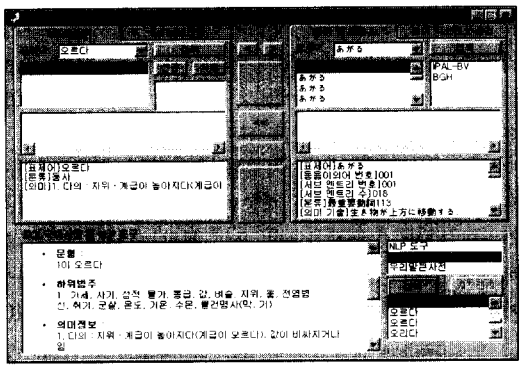
YDK는 Intel Pentium PC, MS-Windows95환경에서 MS-Visual BASIC 5.0을 이용하여 개발하였으며, 서버에 설치된 YDK지원 시스템은 SUN ULTRASparc, Solaris 2.5환경에서 gcc 2.7.2.3과 gdbm 1.7.3 및 CERN httpd 3.0A를 이용하여 개발하였다.

4.2 YDK Version 1.0

YDK Version 1.0(1998.10월 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. YDK Version 1.0의 구성을 (그림 6)에 보이고, 실제로 구현한 YDK의 실행 예를 (그림 7)에 보인다.

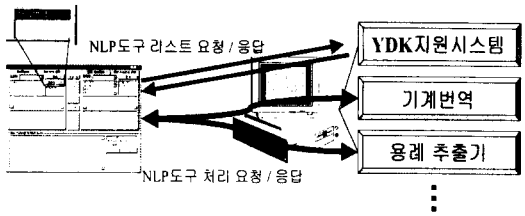


(그림 6) YDK Version 1.0의 구성

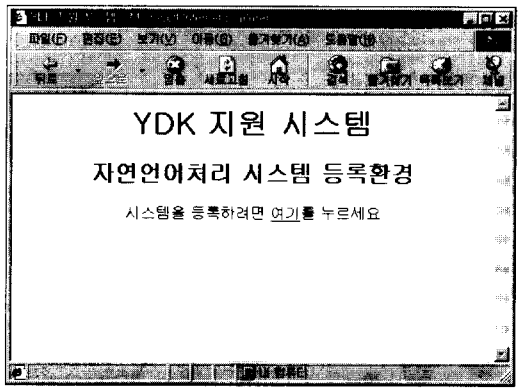


(그림 7) YDK의 실행

YDK는 실행시에 YDK 서버 시스템(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 8)에 나타내고 YDK 지원 시스템의 자연언어처리 등록환경을 (그림 9)에 나타낸다.



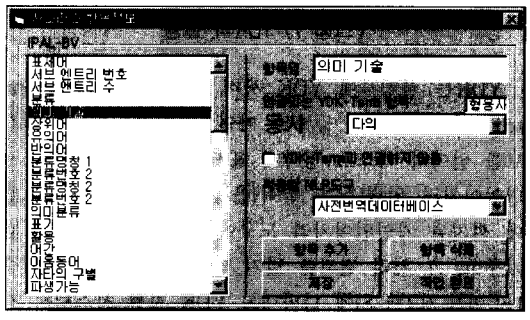
(그림 8) YDK 지원 시스템의 작동



(그림 9) 자연언어처리 시스템 등록환경

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

YDK는 대량의 사전 자료를 참조하기 위해 사전자료간의 mapping 기능을 제공한다. 사용자가 일어 시소러스와 YDK-Term(한국어 용언 통합정보사전)간의 항목정보 mapping 테이블을 작성하면 사전 자료 입력을 자동으로 진행할 수 있게 된다. 설정의 예를 (그림 10)



(그림 10) 항목정보 mapping 테이블

에 보인다. 연결할 YDK-Term의 항목을 선택한 후에 적용할 자연언어처리 도구를 선택해 두면, 이후에 자동변환 기능 실행시에 해당하는 자연언어처리 도구를 실행시켜 그 결과를 YDK-Term의 항목 내용 입력창에 넘겨주는 동작 구조를 가지고 있다.

YDK는 사전간의 자료 변환에 두가지 원칙을 가지고 있다. 일어 시소러스는 버튼 클릭만으로 사전 자료의 한국어 변환 및 YDK-Term으로의 입력까지 자동으로 실행해 주며, 국어사전의 경우에는 Windows95의 클립보드를 이용한 복사 기능을 이용하여 하나하나 입력해야 한다. 이렇게 한 것은 특별한 이유가 있는 것은 아니며, 아직까지 국어사전의 사전정보 분류가 완전하지 않기 때문이다. 기본적으로는 일어 시소러스 사전에 국어사전을 등록해 놓고 사용할 수 있도록 되어 있다. 일어 시소러스 전체를 자동으로 변환하는 기능은 일부가 개발되어 있는데, 개발이 완료되면 일어 시소러스를 단 몇 시간만에 한국어 사전으로 변환할 수 있다.

4.3 YDK-Term

본 연구에서 개발한 사전 시스템으로 한국어 용언의 통합정보사전인 YDK-Term을 개발하였다. 개발한 사전은 동사 사전과 형용사 사전으로 구성되어 있으며, 사전의 보완 작업 및 사전 자료 입력 작업에 YDK를 사용하고 있다. 개발한 통합정보사전의 항목 정보를 <표 1>과 <표 2>와 같으며, 실제의 사전 내용예는 부록에 실었다. 현재, 동사 300단어, 형용사 100단어를 YDK를 이용하여 구축하였다.

〈표 1〉 YDK-Term 동사 정보체계

항 목	부 항 목	
1. 표제어정보	<1> 표제어	
	<2> 동음이의번호	
	<3> 부엔트리 번호	
	<4> 대역정보	일어 ★
		영어 ★
	<5> 출처	
	<6> 분류	
	<7> 음운정보	
2. 의미정보	<8> 의미 기술	다의
		부사 (BC)
	<9> 관련어	상위어
		유의어
		반의어
	<10> 시소러스	
	<11> 의미분류	

항 목	부 항 목			
3. 형태정보	<12> 전성형			
	<13> 어간			
	<14> 자타			
	<15> 활용			
	<16> 파생			
4. 통어정보	<17> 문형			
	<18> 문례			
	<19> 술어소			
	<20> 의미소	N1		
		N2		
		N3		
5. 문법정보	<21> 사동	장 (BC)		
		단 (BC)		
		대역	영어	
			일어	
	<22> 피동	장 (BC)		
		단 (BC)		
		대역	영어	
			일어	
6. 높임정보	<23> 높임형			
6. 기타정보	<24> 관용표현			
	<25> 비고			

〈표 2〉 YDK-Term 형용사 정보체계

항 목	부 항 목		
1. 표제어정보	<1> 표제어		
	<2> 동음이의번호		
	<3> 다의번호		
	<4> 대역정보	일어	
		영어	
	<5> 출처		
	<6> 분류		
	<7> 음운정보		
2. 의미정보	<8> 의미 기술	다의	
		부사	
	<9> 관련어		상위어
			유의어
			반의어
	<10> 시소러스		
	<11> 의미분류		
3. 형태정보	<12> 전성형		
	<13> 어간		
	<14> 활용		
	<15> 파생		
4. 통사정보	<16> 문형		
	<17> 문예		
	<18> 술어소		
	<19> 의미소		N1
			N2
			N3
5. 기능정보	<20> 일반	성질	
		상태	
5. 기능정보	<21> 특수	심리	
		비교	
		지시	
		의문	
6. 문법정보	<22> 기능성 동사화		
	<23> 행동성 동사화		
	<23> 사동 동사화		
7. 기타정보	<24> 관용표현		
	<25> 비고		

4.4 평가 및 고찰

본 연구에서 개발한 사전 시스템인 YDK는 사전 개발 과정에서의 작업 절차와 작업 시간을 줄여준다. 이 시스템을 사용한 사전의 구축은 평균적으로 한시간에 6단어의 사전정보 입력이 가능하며, 일괄처리 기능을 이용하면 시간당 100단어에 대한 사전자료 구축이 가

〈표 3〉 사전자료 입력 절차 비교

		단어별 직접입력	일괄입력기능 활용
일반적인 사전자료 입력	코퍼스로부터 추출한 단어입력	① 표제어 입력 ② 표제어 번역 ③ 일어 시소러스 검색 ④ 항목선택 ⑤ 항목정보 번역 ⑥ 항목정보 입력 ⑦ 자료 저장 ▶ 7 단계	없음
	일어 시소러스 번역	① 표제어 선택 ② 표제어 번역 ③ 일어항목 선택 ④ 한국어항목 선택 ⑤ 일어항목의 번역 ⑥ 항목정보 입력 ⑦ 자료 저장 ▶ 7 단계	없음
YDK에 의한 사전자료 입력	코퍼스로부터 추출한 단어입력	① 표제어 입력 ② 일어 시소러스 검색 ③ 일어항목선택 ④ 한국어항목 선택 ⑤ 항목정보의 import ⑥ 자료 저장 ▶ 6 단계	① 표제어 입력 ② 일어시소러스 검색 ③ 항목정보의 import ④ 자료 저장 ▶ 4 단계
	일어 시소러스 번역	① 일어표제어 선택 ② 표제어 import ③ 일어항목 선택 ④ 한국어항목 선택 ⑤ 항목정보의 import ⑥ 자료 저장 ▶ 6 단계	① 입력 시작 → 입력 완료 ▶ 1 단계

능하여 매우 빠른 시간에 사전자료의 구축이 가능하다. 사전자료 입력절차를 최소 40% , 최대 85%를 줄임으로써 사용하기 편리하다는 장점이 있다. 그 절차 비교를 <표 3>에 보인다. 이 결과는 KAIST에서 개발되어 사용중인 대표적인 프로그램인 TDMS의 사전자료입력 절차와 YDK의 사전자료 입력절차를 단적으로 비교한 것이다.

5. 결 론

본 논문에서는 다국어 통합정보사전 구축을 위한 사전구축 방법론을 정립하고, 정립된 방법론을 바탕으로 하여 통합정보사전 개발 시스템을 구현하였다. 개발한 사전은 특정 분야에 맞춰 전문화된 사전이 아니라 기존 사전의 정보를 통합하고 여러 자연언어처리 시스템을 이용하여 얻을 수 있는 정보를 통합한 사전으로 국어정보처리 응용시스템 개발에 도움을 줄 수 있다. 또한, 통합정보사전 개발 시스템은 각종 전자사전 및 자연언어처리 시스템들의 정보를 손쉽게 통합할 수 있으므로 고품질의 사전을 개발할 수 있다. 향후 연구과제로는 YDK-Term 사전 체계의 보완을 통해 고품질의 통합정보사전으로 발전시켜나가는 것과 하나의 시스템에서만 사용이 가능한 현재의 YDK를 인터넷 환경을 기반으로 한 다중 사용자 환경으로 버전업하는 것 등이 있으며, 지속적인 사전자료 입력을 통해 현재의 문제점들을 분석하여 대규모의 통합정보사전 개발에 적합한 사전 시스템으로 발전시키는 과정이 있다.

부 록

YDK-Term의 실례

가다
[표제어] 가다
[동음이의번호] 001
[부엔트리번호] 001
[대역일어] ゆく
[대역영어] go
[출처] KAIST 코퍼스 '96
[분류] 동사
[다의] 어떠한 곳을 향하여 움직이다.
[부사] 일찍, 자주, 멀리, 늘, 갑자기
[상위어] 움직이다
[반의어] 오다
[시소러스] IPAL, BGH(2. 1 5 2 7)
[의미분류] 동작-변화-이동-출발,귀착
[어간] 가
[자타] 자
[활용] 거라 불규칙
[문형] N1이 N2(에, ~에게, ~로, ~에게로) 가다
[문예] 철수가 학교에 간다
[N1] 기동성이 있는 명사(사람, 동물, 이동수단)
[N2] HUM, 장소명사
[사동장] 가게 하다
[피동장] 가지다
[관] "가는날이 장날" 우연히 갔다가 의외로 공교로운 일을 만났을 때 쓰는 말.

참 고 문 헌

[1] 황도삼 외 4, , "심층 국어정보처리 품질관리체계", 한국과학기술원, 위탁과제 최종보고서, 1998.
[2] 이재성 외 3, "텍스트 및 전자사전 관리시스템의 설계", 한국정보과학회&한국인지과학회, 제8회 한국어 정보처리 학술대회 논문집, pp.408-414, 1996.
[3] 최병진 외 3, "표준화를 위한 일반사전의 논리 구조", 한국정보과학회&한국인지과학회, 제8회 한국어 정보처리 학술대회 논문집, pp.415-423, 1996.
[4] 한국과학기술원, "텍스트코퍼스 및 전자사전 관리시스템(TDMS)", 과학기술처, 통합 국어정보베이스 최종보고서, pp.17-150, 1996.
[5] 오름테크, "국어정보처리기술 개발-사전및텍스트관리 통합시스템 개발", 한국과학기술원 위탁과제, 제2차년도 최종보고서, pp3-4, 1996
[6] 황도삼 외2, "자연언어처리", 홍릉과학출판사, 1998.
[7] 技術センター, "計算機用日本語動詞辭典IPAL(Basic Verbs)", 일본 정보처리 진흥사업협회, 1987년 3월.
[8] 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.
[9] 부산대학교, "한국어 문장 분석을 위한 용언의 하위범주화에 관한 연구", 시스템공학연구소 최종보고서, 1997.
[10] 大野 晋, 竝書淨人, "角川 類語新辭典", 角川書店, 1980.
[11] 岩波書店, "日本語語彙大系", 日本電信電話株式會社, 1997.

황 도 삼

email : dshwang@yu.ac.kr

1980년 홍익대학교 전사계산학과 졸업(학사)

1983년 연세대학교 전자계산전공 (공학석사)

1995년 Kyoto University 전자정보통신전공(공학박사)

1980년~1995년 한국과학기술연구원 시스템공학연구소 책임연구원

1995년~현재 영남대학교 전자정보공학부 부교수

관심분야 : 자연언어처리, 전자사전, 기계번역, 정보검색 등

최 기 선

e-mail : kschoi@cs.kaist.ac.kr

1978년 서울대학교 수학과 졸업 (학사)

1980년 한국과학기술원 전산학과 졸업(공학석사)

1986년 한국과학기술원 전산학과 졸업(공학박사)

1985년~1986년 한국외국어대학교 전산학과 조교수

1987년~1988년 일본 NEC C&C 정보연구소 초빙연구원

1988년~현재 한국과학기술원 전산학과 교수

1998년~현재 한국과학기술원 전문용어언어공학연구센터 소장

관심분야 : 자연언어처리, 기계번역, 정보검색, 전문용어 등