한국어 전자사전 원형의 설계 및 구현

(하이퍼택스트 기법 사용)

양단희, 최윤철, 송만석

연세대 전산과학과
한국어 전자 사전 개발실
서울 서대문구 신촌동 134(우편번호 120-140)

요 약

언어 정보 처리가 정보과학의 가장 중요한 분야의 하나로 부상하면서 언어 정보 가치가 높은 실용성있는 우리말 사전의 필요성이 더욱 더 절실해지고 있으며, 또 우리 언어 사회에 외래어가 지나치게 범람함으로써 고유 언어 문화를 위협할 정도가 되어 외래 신생어에 대한 대책이 긴요하다 하겠다. 그러므로 본 논문은 최신 전산 기술을 이용하여 우리말 어휘 뭉치를 대량 정보 처리함으로써 우리말 언어 세계를 신속하고 정확하게 반영하며, 실용성있고 사용하기 편리하며, 우리말 어휘 확장에 보조 역할을 해줄 하이퍼텍스트화된 우리말 전자사전을 제안하며 그 원형을 설계 및 구현하였다.

1. 서론

국어 사전 하나가 편찬되기까지는 많은 인력과 오랜 기간의 수작업을 요한다. 예컨데 한글 학회의 우리말 '큰 사전'(1929~1957년)은 28년의 편찬 기간에, 총 20여 만의 어휘 수집에 16만여의 표제어를 실었고, 영국의 O.E.D.(Oxford English Dictionary, 1858~1928)는 70년의 세월에 총 5백만의 보기 인용문을 수집하여 40여 만의 표제어를 실었었고 증보판이 나오기까지도 30년이 걸렸다.[13,14]

때문에 그 장대한 기간은 시시각각 변화하는 우리 언어 세계를 신속하게 포착하여 언어 정보 가치가 높은 실용성있는 최신 우리말 사전을 편찬하는데 장해물이었다. 이에 종합적 최신 언어 자료가 부재하여 신외래 어휘(전문 용어와 과학 발달에 따른 신조어)의 우리말화에 대한 노력을 기하기가 몹시 어려워 그대로 외래어로 수용함으로써 실제 우리말은 세월의 흐름에 따른 사멸어가 증가하는 반면 생성어(신조어)는 거의 없는 실정이 되어 버렸다. 이는 자칫 선진국에 언어적 예속화가 야기되어 우리나라 고유 언어 문화가 흔들리게 될 지도 모른다.

그러므로 신외래 어휘가 토착화되어 외래어로 굳어지기 전에 우리말화에 대한 언어 과

학적인 대책과 올바른 모국어 사용 유도가 필요하겠다. 그래서 본 논문은 자연어 정보 처리 기법을 이용하여 사전 편찬 기간을 크게 단축 시켜주고(영국의 COBUILD 과제에서는 불과 6~7년으로 단축되었다.[14]) 현 언어 세계를 바로 바로 반영해주기 위해, 장차 온라인 상의 대규모 국어 사전 서비스를 내다보며 그 첫걸음으로 하이퍼텍스트화된 우리말 전자 사전 원형을 퍼스널 컴퓨터상에서 설계 및 구현하였다.

2. 관련 연구

2-1. 외국 현황

세계 각국은 대규모 언어 정보 데이타베이스 구축과 전자 사전 개발 및 그로인한 과학적 원칙하의 대규모 사전을 편찬했거나 상당 기간 연구중이다.[14,17,18,22,23]

1) 미국

◆Xerox PARC의 언어 정보 처리, 전자 사전 연구.

◆스탠퍼드대학의 'Center for the Study of Language and Information(SCLI)'의 연구.

2) 영국

◆버밍엄대학의 COBUILD 과제의 언어 정보 데이타베이스 개발 및 사전 편찬(연세대 한국어 사전 편찬 연구진이 1차 방문하여 기술 지원 약속을 받음).

◆옥스퍼드대학의 O.E.D.의 데이타베이스 전환 및 대규모 사전의 CD-ROM화

3) 일본 ◆전자 사전 연구 협회(EDR)의 언어 정보 데이타베이스 개발 및 이용 기술

4) 프랑스

◆파리 7 대학의 'Laboratoire d'Automatique Documentaire et Linguistique(LADL)'의 전산학과 언어학의 연합적 연구에 의한 각종 전자 사전의 개발.

◆'Center National de la Recherche Scientifique' 소속 기관인 'Institutee National de la Langue Francaise(INALF)'의 대규모 사전인 'Tresor de la Langue Francaise'의 편찬을 비롯한 각종 데이타베이스의 개발

5) 캐나다

◆워털루 대학의 옥스퍼드 영어 사전의 전자 사전화 및 하이퍼텍스트 시스템 연구 개발

6) 노르웨이 ◆오슬로 대학의 LOB계획에 의한 현대 영어 낱말 빈도 조사 연구

7) 스위스 ◆INFORTERM(유네스코 후원)의 전문 용어 은행

2-2. 국내 현황

1) 연세대 한국어 사전 편찬회

한국어 사전 편찬회(각 분야 교수 375 동인)가 1986년 발기되어 십 수차의 연구 발표

회를 통해 사전 편찬 이론과 실제, 컴퓨터에 의한 언어 정보 처리, 전자 사전 등 한국어 사전 편찬 분야에 대한 지식을 축적, 3차에 걸친 공동 연구를 수행하여 사전 편찬학 연구 논문집 3차 발간.[7]

2) 연세대 한국어 사전 편찬실과 전자 사전 개발실

한국어 사전 편찬실이 1989년 연세대 부설 연구 기관으로 설치되고 전산과에는 전자 사전 개발실을 설치하여 편찬학, 전산학, 정보 검색학, 계산 언어학, 인지 심리학, 기타 관련 학자들이 공동 참여하여 장기적 계획을 가지고 연구 진행중이며 어휘 뭉치의 서지학적 차원의 표본 선정[11]을 통해, 전산 처리에 필요한 어휘 뭉치 분류 정보를 삽입하여 3천만 어휘를 구축하였고, 언어 정보 처리 도구 개발을 위해 국어 대사전의 표제어와 관련 문법 정보에 관한 데이타베이스를 구축해 놓았다.

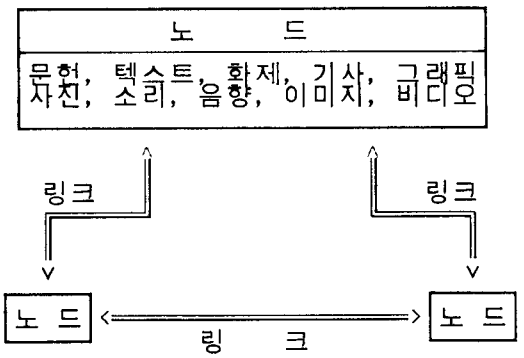
3) 과학 기술원, 그리고 서울대, 인하대 그외 몇몇 대학에서 주로 기계 번역과 관련해 한국어 분석 처리에 대한 연구가 꾸준히 진행중이다.[2~6,15],

3. 이론적 기초

3-1. 하이퍼텍스트

1) 하이퍼텍스트(hypertext)와 하이퍼미디어(hypermedia)

부시(Bush)가 Memex 시스템(1945년)에서 처음으로 제안한 개념으로, 일반적으로 이산적 데이타 사이에 링크의 생성과 표현을 하이퍼텍스트라 하며, 그 데이타가 텍스트, 숫치뿐만 아니라 그래픽, 소리, 음향 등이 될 수 있을 때 그 결과 구조를 하이퍼미디어라고 한다. 개념적으로는 텍스트와 데이타가 노드로 표현되어지는 의미 망조직(semantic networks)의 착상과 유사하다.[21] 실제적 측면에서 보면, 기민한 전후 참조(active cross-reference) 기능을 갖고서, 사용자가 요구할 때 데이타베이스의 다른 부분으로 손쉽고 신속한 이동을 가능케하는, 레코드와 화일사이에 무제한적인 링크를 지닌 일종의 비정규형(Non-First Normal Form) 모델인 데이타베이스라고 볼 수 있다.[12,16,20,21]



링크는 문헌 내부에 위치하거나 그래픽, 비디오 상의 부분으로 삽입되던가 문헌 끝에 나열 혹은 색인안에 포함되어질 수 있다.

그림 1. 하이퍼텍스트의 구성

2) 하이퍼텍스트상 지식 표현의 유형

하이퍼텍스트는 노드와 링크로 구성된 망 구조로서 다음과 같은 유형이 있다.[21]

가) 노드 유형

텍스트 노드, 그림 노드, 음향 노드, 혼합 매체 노드, 버튼성 텍스트 노드,
색인된 텍스트 노드, 색인 노드, 대상체 노드, 규칙 노드

나) 링크 유형

◆항해 링크: 이동 링크, 확대 링크,축소 링크, 뷰 링크

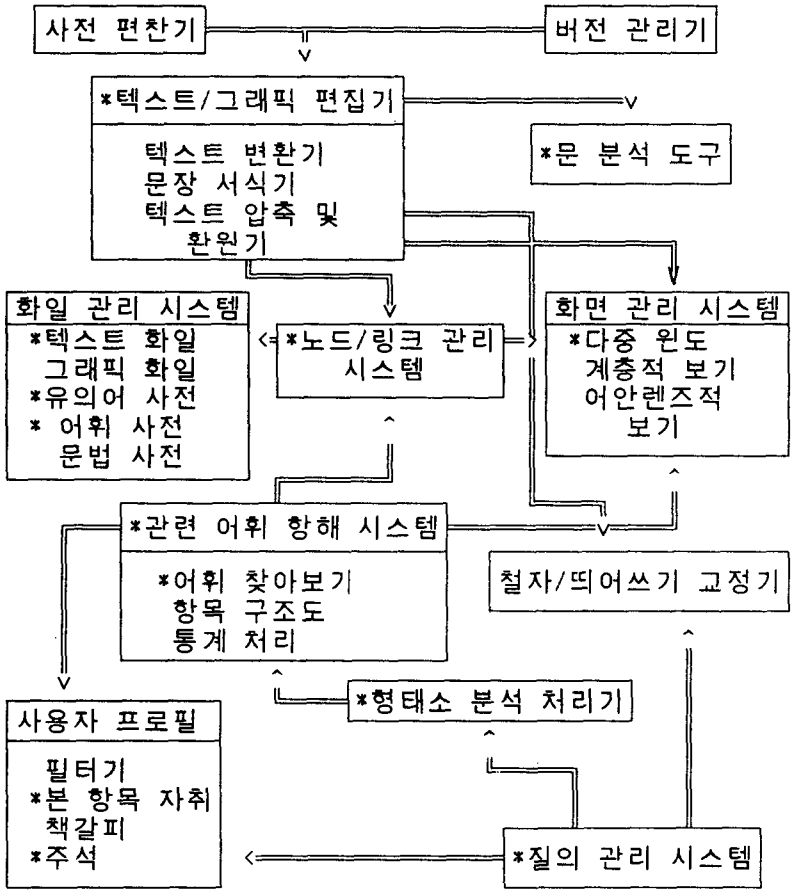
◆조직,추론 링크: 색인 링크, 피포함(is-a)링크, 포함(has-a) 링크,
함축의미 링크, 수행 링크

3-2. 전자 사전

보편적 정의는 일반적으로 낱말에 관한 각종 언어 정보를 주로 CD(Compact Disk)-ROM에 기록하여 온 라인 데이타베이스화함으로써 낱말에 관련된 모든 정보를 보다 빠르게, 편리하게, 값싸게 그리고 흥미롭게 이용할 수 있는 시스템을 말한다.[14,18,22]

3-3. 하이퍼텍스트화된 전자사전

1) 하이퍼텍스트화된 전자 사전의 총체적 구조



위 그림에서 별표 항목이 본 시스템에서 설계 구현한 부분이고 나머지는 구현 도구(tool)만 구현하였다.

그림 2. 하이퍼텍스트화된 전자사전 구조도

2) 하이퍼텍스트화된 전자사전의 장점[17,21]

◆ 인간의 연관 기억과 연관적 사고에 충실한 편리한 검색

◆ 다양한 계층구조로 관련 정보를 한 곳으로 쉽게 모아 줌

◆ 비선형 텍스트를 제공

◆ 다중 매체를 사용하여 인간의 이해를 증대

◆ 화면에 보여줄 자료 크기와 디스플레이 방식 조정

◆ 대화식 방식으로 자유롭고 손쉽게 동의어,반의어,용례의 참조로 문장 작성시 온 라인 상에서 사전의 효율적 사용

◆ 다른 지식 구조를 구현하기 위해 사용되어질 수 있고, 삽입되어질 수 있는 유동성

4. 하이퍼텍스트화된 전자 사전 주요 구성 요소 구현

본 하이퍼텍스트 데이타베이스는 사전의 각 표제 서술 내용을 텍스트 노드 유형으로 하고 가변 크기의 레코드이며, 링크는 관련된 노드로 단순히 이동시키는 이동 링크 유형으로서, 표제어와 표제 서술 내용으로 구성되어 사전 검색 방식이 참조 문헌 검색 시스템과 유사[10]하기 때문에 표제 서술 문장의 어절 자체에서 형태소 분석을 통해 추출해 낸 기본형을 가지고 색인표를 이진 탐색하기 때문에 사전 편집자는 전혀 링크의 개념을 염두에 둘 필요가 없다. 기본적으로 화일 구조는 색인된 가변 레코드 순차 화일 구조이며, 편리한 사전 이용을 위해 어느 시점에서도 상태 의존 도움말 기능이 가능하며 전체적으로 풀다운 메뉴로 운용된다.

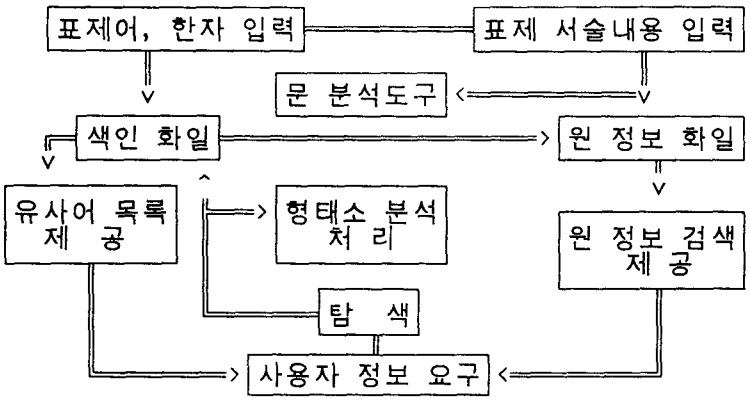


그림 3. 전자 사전 원형 시스템 구성도

4-1. 연세 말꾸러미(전자사전) 편찬기

1) 연세 말꾸러미 편집기(Yed)

국어 사전만의 독특한 품사나 불규칙 활용 표시등을 위해 폰트 편집기를 사용하여 90여자를 만들어, 편집기 화면의 '약호,약어 기호' 윈도 상에서 토글 키와 숫치 키를 사용하여 입력할 수 있게 하였으며 한자는 행망용 카드의 4888자를 그대로 호출하여 사용했다.

표제어와 표제 서술 내용이 입력될 때 자동적으로 색인된 순차 화일이 생성되며, 유사어 사전(thesaurus)도 비슷한 방식으로 생성시킨다.

주 메뉴/약호, 약어 기호 윈도	
표제 내용 기술 도움 윈 도	서술 내용 편집 화면겸 상태 의존 도움말 윈도
주요 키 사용 / 한자 보이기 윈도	

그림 4. Yed 화면 구성도

2) 연세 문 분석 도구(YOCP)

우리말 '큰 사전' 이후 기존 여타 국어 사전은 그 사전에서 약간씩 첨가, 삭제하여 편찬할 뿐이지 독자적으로 광범위하게 조사, 수집하여 편찬하지 못하고 있다. 또한 사전의 용례는 실 문장에서 발췌한 인용문이기보다는 용례를 위해 인위적으로 만들어진 것이 태반이다. 옥스퍼드 사전 편찬에는 5백만이나 되는 보기 인용문을 수 년에 걸쳐 실 문장에서 수집했다는 것을 유념할 필요가 있다.[7,8,11,13] YOCP는 우리말 상용 2바이트 조합형 문자을 입력으로 받아 어휘 빈도, 색인 작성, 출처 색인된 어휘 문맥 리스트 등을 자동으로 생성해 주는 프로그램인데, 옥스퍼드 계산 서비스 센타에서 메인 프레임상의 OCP Version 2(Oxford Concordance Program, 1986~1987년)를 IBM PC 호환 기종판으로 수정한 Micro-OCP[24]를 가지고, 한글 처리가 가능토록 명령어 한글 편집기를 내장한 전·후 처리기로서 명령어에 대한 한글 데이타 사용이 가능토록 하여, 어휘 빈도수 조사를 통한 표제어 선별과 출처 색인된 어휘 문맥 목록을 통해 용례 인용문 발췌에 이용된다.

4-2. 연세 말꾸러미

1) 화면상 항해 시스템(Ydic)

우리말 전자 사전은 일반인들이 특별한 지식없이도 손쉽게 이용할 수 있도록 'See and point interfaces' 방식[16]을 택하여 전체 사전 구조를 사용자가 전혀 의식할 필요가 없도록 하였으며 이를 위해 표제어가 굴절, 활용된 상태에서도 화면상에서 직접 찾아볼 수 있도록 전자 사전 전용 형태소 분석(기본형 찾기) 시스템을 두었다. 또한 동철이의어가 있을 경우 '유사어 목록' 윈도에서 한자로 선택할 수 있게 하였다. 지나온 자취를 항상 추적할 수 있도록 화면상에 '둘러 본 항목 자취' 윈도를 두었으며 유사어(thesaurus)를 직접 찾아볼 수도 있다.[16,23]

주 메뉴 윈도 및 항해 경로 추적 윈도	
유사어 목록/ 동철 이의어 윈 도	전자사전 항해 및 상태 의존 도움말 윈도
주요 키 사용 윈도	

그림 5. Ydic 화면 구성도

2) 형태소 분석 시스템(Ylex)

전자 사전 본질상 형태소 분석에 필요한 정보는 색인 화일과 표제 서술 내용 화일에 전부 수록되어 있어, 형태소 분석을 위한 시스템 전용 사전 크기는 고려 대상이 되지 않는다.

기본 알고리즘[1,3,4,6,8]은 다음과 같다.

가) 한 어절을 분리해와 표제어 시험

나) 굴곡형(체언,부사,용언 + 조사)이면 조사를 분리하고 표제어 시험

다) 파생어(접두어+자립형+접미어)이면 접사 분리하고 표제어시험

라) 활용형(어간+어미)이면 용언 처리한 후 표제어 시험

마) 위의 모든 경우가 아니면 위의 수행 결과로 생긴 임시 기본형의 첫음절부터 최장 일치법에 의해 색인 화일에서 검색해내고 불완전 처리임을 신호

◆ 표제어 시험: 그 어절의 첫음절, 초성에 해당하는 색인 사전을 이진 탐색하여 성공하면 수행 완료.[15]

◆ 어절의 기준은 화이트(white) 문자로 구분되는 각 문자열내에서 한글이나 한자 부분만 취한 것이다.

◆ 굴곡형을 먼저 시험하는 이유는 굴곡형은 전체 어절 사용 빈도에서 48.0%를 차지하고 활용형은 32.8%이기 때문이다.[6]

5. 결론

전자사전 개발실에서 3백만 어휘 뭉치에 대한 언어 정보 처리 도구와 어휘 데이타베이스 구축 및 전자 사전 개발에 관해 장기적 계획을 가지고 연구중이고,한국어 사전 편찬실에서는 효율적 국어 사전 편찬을 위해 또한 지속적 연구중이다.[14]

그 연구 과제 일환으로 본 논문은 하이퍼텍스트화된 전자 사전 개발상에 필요한 다양한 우리말 처리 도구를 연구,. 설계하여 제반 문제를 시험대에 올려 놓고 그 원형을 구현해본 것이며, 계속적 연구를 통해 최종 시스템이 완성되면 우리말의 각종 언어 정보에 관한 정밀한 지식을 토대로 타 학문(특히 국어학)의 이론 정립에 크게 이용될 수 있겠고 한글 표준화 문제, 기계 번역 시스템, 전문 용어 은행, 각종 전자 백과 사전과 전문 분야의 전자 사전 개발 등에 지대한 파급효과가 있으리라 기대된다.

참 고 문헌

[1] 김대식, "Lexical Database 구축을 위한 어절 분석 도구에 관한 연구", 연세대 대학원 석사학위 논문, 1990.

[2] 김창년, "한국어 기계 번역을 위한 사전 구성", 청주대 산업대학원 석사학위 논문, 1987.

[3] 박세영, 김권양, "자연어 처리를 위한 한국어 syntax 구조에 관한 연구", 정보과학회 논문지, 1985.

[4] 손석원, "기계번역을 위한 한국어 용언의 분석",인하대 석사학위 논문, 1987.

[5] 송춘환, 강재우, 김연배, "한글 철자 및 띄어쓰기 검사기", 정보과학회 가을 학술발표 논문집, 1989.

[6] 윤면기, "자연어 처리를 위한 형태소 분석", 인하대 대학원 석사학위 논문, 1984.

[7] 이상섭, 남기심 외, 새 한국어 사전 편찬을 위한 사전 편찬학 연구, 제1집, 제2집, 탑 출판사, 1988.

[8] 이상섭, "뭉치 언어학: 사전 편찬의 필수적 개념", 한글 및 한국어 정보처리 학술발표 논문집, 1989.

[9] 장유미, "한글 띄어쓰기 검사기의 개발에 관한 연구", 연세대 대학원 석사학위 논문, 1990.

[10] 정영미, 정보 검색론, 정음사, 1986.

[11] 정찬섭, "한국어 어휘뭉치의 표본 선정 기준", 한글 및 한국어 정보처리 학술 발표 논문집, 1989.

[12] 정희영, "하이퍼스페이스 상에서 효율적인 탐색을 지원 하는 하이퍼텍스트 시스템", 연세대 대학원 석사학위 논문,1990.

[13] 조재수, 국어 사전 편찬론, 과학사, 1984.

[14] 최윤철, 송만석, "한국어 전자 사전 개발의 현황과 과제", 한국정보과학회 국어정보처리 춘계워크샵 학술 발표논문집, 1990.

[15] 홍종화, "초성 테이블을 이용한 한글 문헌 정보 검색 시스템의 설계 및 구현", 한양대 산업대학원 석사학위 논문, 1986.

[16] Ben Shneiderman and Greg Kearsley. Hypertext Hands-On!. Addison-wesley, 1989.

[17] D.R. Raymond and F.W. Tompa. "Hypertext and the Oxford English Dictionary," Comm. of the ACM, July 1988.

[18] Electronic Dictionary Research Project(EDR), Electronic Dictionary Research Inst, Japan. 1988.

[19] Gerard Salton. Automatic Text Processing. Addison-Wesley, 1989.

[20] Jeff Conklin. "Hypertext: An Introduction and Survey," IEEE Computer 2, 9, Sept. 1987.

[21] Kamran Parsave, Mark Chignell, Setrag Khoshafian and Harry Wong. Intelligent Databases. Wiley, 1989.

[22] Sinclair. Looking Up-An Account of the COBUILD Project, Collins ELT, 1987.

[23] Oxford English Dictionary on Compact Disc, User's Guide Version 4.10, 1987.

[24] Oxford University Computing Service. MICRO-OCP. Oxford University Prsss,1988.