유니코드3.0의 CJK 한자 정렬

윤지헌°, 변정용
동국대학교 컴퓨터학과

A Sorting of Unicode 3.0 CJK Chinese Characters

Zi-heon Yoon, Jeongyong Byun
Dept. of Computer Science, Dongguk University

요 약

최근 많은 양의 문서가 전자화되어 컴퓨터에 저장되고 인터넷을 통하여 공유가 되고 있고, 그 범위를 고문헌에까지 넓혀가고 있다. 그러나 한자 문화권의 고문헌은 대부분 2만에서 3만여 자의 한자로 작성되어 있어서 한자 입력 시 코드문제가 뒤따른다. 하지만 유니코드 3.0에서는 27,786자의 한자를 코드화 하여 놓아서 한자 문화권 나라에 많은 도움을 주고있다. 하지만 한중일 3개국에서 많이 쓰이는 한자를 대상으로 하여 부수, 획수 순으로 정렬하여 국내 실정에 맞지않고 그나마 유니코드 한자를 입력할 수 있는 환경도 MS Word 2000 정도로 제한적이다. 본 논문에서는 유니코드 3.0 한자 입력기에서 기본 한자 코드로 사용될 CJK 한자 영역에 배정된 한자를 정렬하는 방안을 제안하고 운영체제 독립적인 한자 입력 시스템에 활용한다.

1. 서론

컴퓨터와 인터넷의 발달로 수많은 문서들이 전자화되어 컴퓨터에 저장되고 인터넷을 통하여 공유할 수 있게 되었다. 하지만 우리나라나 중국, 일본처럼 한자 문화권의 나라에서는 한자 입력에 많은 어려움이 따른다. 특히, 고문헌에 사용되는 한자의 경우 한자의 수가 2만에서 3만여자[12]에 달하고 있어서 더욱 어렵다. 하지만 한자 문화권의 여러 나라에서는 많은 수의 한자를 컴퓨터에서 이용할 수 있도록 노력하여 왔고, 그 결과 각국에 현실에 맞는 한자코드를 제정하게 되었다. 하지만 한자 문화권 나라에서 동일한 모양의 한자를 사용하여도 자체적으로 한자 코드를 제정하여 사용하고 있어서 많은 문제가 생긴다. 하지만 새로운 만국 공용 코드 표준인 유니코드3.0[1]을 사용하게 되면 전세계 어디에서나 동일한 코드를 사용하기 때문에 해당국의 글꼴만 설치되어있다면 문서에 쉽게 접근이 가능하다. 이 유니코드 3.0에는 한국, 중국, 일본에서 많이 사용하는 한자 27,786자를 CJK영역에 배정해 두었지만 정작 유니코드3.0 한자를 입력하는 방법은 현재로선 찾아보기 힘들다. 또한, 한자들이 부수, 획수 순으로 정렬되어 있어서 국내실정에 맞지 않다. 현재까지 유니코드3.0 한자를 입력하는 방법은 MS Windows 환경에서 MS Word2000[2]에서만 가능한데, 이것은 운영체제나 응용 프로그램에 종속적이어서 누구나 사용하기가 곤란하다.

본 논문에서는 유니코드3.0 한자의 정렬방법을 제시하고, 유니코드 CJK영역의 한자를 모두 입력할 수 있고 사용자의 운영체제 환경에 독립적이면서 범용적으로 작동하는 유니코드3.0 한자 입력 시스템에

적용하고 이를 통하여 검증한다.

2. 기존 연구 현황

국내에서 가장 많이 쓰이는 워드프로세서인 HWP의 경우 자체 코드인 확장조합형 코드에 한자를 배정하여 사용하고 있다. 하지만 국제표준코드가 아닌 자체 구현한 코드를 사용하고있기 때문에 비표준 문제가 야기되어 정보교환 시 수신측에서 HWP를 가지고있지 않다면 한자정보가 모두 유실된다.

MS Word2000의 경우 유니코드 한자를 지원하는 대표적인 입력 시스템이다. 하지만 MS Word2000은 윈도우 환경만을 지원하고 있어서 모든 운영체제 환경에서 사용하는 것은 불가능하다.

웹 브라우저로 많이 쓰이는 Netscape이나 Internet explore의 경우 한국, 중국, 일본에서 사용되는 한자를 표현할 수 있지만 자국에서만 쓰이는 코드를 화면에 표시만할 수 있고 입력기능은 운영체제에 종속적이다.

그 외 국내외의 많은 한자 입력시스템에서 한자를 지원하고 있지만 이러한 것들은 대부분 자국의 코드만을 지원하고 유니코드를 지원하지 않고 있다. 국내의 대부분의 한자를 지원하는 문서편집 소프트웨어들도 운영체제에서 지원하는 완성형 한자만을 지원하고 있다.

3. 한자 입력 시스템 설계

3.1. 요구사항

한자 입력 시스템에 있어서 요구사항은 다음과 같다.

한자 변환 사전의 한글음 순 정렬
모든 유니코드 한자 표현
단어단위의 한자변환 가능
두음법칙 등의 한자 표기규칙 지원
운영체제 독립적 실행환경 제공
유니코드 글꼴 포함

3.2. 한자 변환 사전을 위한 유니코드 3.0 CJK 한자 정렬

유니코드 한자 입력기의 가장 중요한 기능은 한글로 입력된 음절을 한자로 바꾸는 것이다. 이를 위해서 한자음을 한글로 변환하는 변환표가 필수적으로 요구된다. 기존 연구[3]에서는 유니코드 컨소시움에서 제공하는 'Unihan.txt'파일을 변환해 사용하였다. 'Unihan.txt'에는 유니코드 한자를 한자 문화권 나라의 발음, 자국 코드, 사전 위치 등 59개의 항목으로 분류하여 놓고 있는데, 변환표에서 사용하여야 할 항목은 한자의 한글 음을 로마자로 표기하고있는 kKorean이다. [3]에서 kKorean항목을 변환한 변환표는 한글 음별로 오름차순정열이 되어있지 않고 파일 포맷이 복잡하여 확장에 어려움이 따른다. 또한, [3]에서 구현된 입력기는 한자 변환 모듈이 입력 모듈과 구분되어있지 않아 자바 빈즈(Java Beans) 등의 콤포넌트로 재구성시에 많은 어려움이 뒤따르고, 한글 처리 모듈의 부재로 한국어정보처리 시 한글과 한자를 따로 처리해야 하는 불편함이 따른다.

본 연구에서도 [3]에서와 같이 'Unihan.txt'를 변환하여 사용하였지만, 파일을 오름차순정열 하였고 확장성을 위해 파일포맷을 변경하였다. 그 결과 [3]보다 더욱 높은 변환속도와 안정성을 보여주었고, 단어단위의 한자변환을 위한 사전의 추가가 용이하게 되었다.

3.3. 유니코드 3.0 한자 입력 시스템 설계

입력 받은 한글 음을 한자로 변환하기 위해서는 한글/한자 문자열 처리 클래스를 생성하여야 한다. 이 문자열 처리 클래스에서 두음법칙이나, 이음어 ( ) 등을 처리하여 한자 변환표에서 해당하는 한자로 바꾸게 된다.

두음법칙 등을 적용한 한자음을 처리하기 위해서 한글 정보 처리용 모듈이 반드시 필요하다. 이를 위해서 [4]에서 C++용으로 연구된 정음형 한글 코드 처리 모듈을 자바로 변환하여 탑재하였다. 이때,

[4]에서 개발된 C++용 모듈은 아스키 코드(ASCII Code)를 입력 받아 정음형 한글 코드로 변환하여 처리한 후 아스키 코드로 되돌려주는 방법을 취하고있지만, 자바에서는 기본 처리 코드가 유니코드이기 때문에 반드시 모든 모듈에서 아스키 코드가 아닌 유니코드를 처리하게 변환을 하여야 한다.

이렇게 변환된 한자를 화면에 출력하기 위해서는 유니코드 글꼴과 자바1.2를 실행할 수 있는 JRE(Java Runtime Environment)를 시스템에 설치한다. 이때, JRE에서 사용하는 글꼴은 기본적으로 아스키 글꼴을 사용하게 설정되어있는데, 이 아스키 글꼴을 대신해서 유니코드 글꼴을 사용하게 하기 위해서 기존연구[3, 5]에서 처럼 '$JAVEHOME/jre/lib/font.properties'를 변경한다.

4. 한자 입력 시스템 구현

본 연구는 구현에서의 어려움 보다 설계에서의 어려움이 상대적으로 크다. 왜냐하면 자바 언어는 언어 자체에서 유니코드를 지원하기 때문에 구현에서는 단순히 유니코드를 처리하고 출력만하면 되기 때문이다. 하지만 현재까지 직접적으로 유니코드를 처리하는 방법과 출력하는 방법이 많이 알려지지 않아 설계 시 많은 시행착오가 반복되었다. 또한, 유니코드에서 지원하는 완성형 한글 11172음절을 한글 정보 처리용으로 사용하기 위해서는 항상 반복되는 부수적인 작업 때문에 시스템에 많은 부하가 걸린다. 따라서 한글 정보 처리에 적합한 코드인 정음형 코드를 내부 처리에 이용하였다.

4.1. 구현 환경

3.1의 요구사항을 만족하기 위해서 개발환경으로 Window2000에서 자바1.2를 사용하였고, 유니코드 글꼴은 일반에 공개되어있는 Bit Stream사[6]의 CyberBit글꼴[7]을 사용하였다. 또한, 한글 정보 처리를 위해 [4]에서 연구된 C++용 정음형 한글 처리 클래스를 자바용으로 변환하여 사용한다.

4.2. 한자 및 한글 정보처리 클래스

한자 변환은 크게 몇 부분으로 나뉠 수 있으나 가장 중요한 부분은 하나의 한글 음절을 한자로 변환하는 부분이다. 하나의 음절 변환 시 가장 주의하여야 하는 부분은 두음법칙 등의 적용으로 한자음이 변하는 경우[8]이다. 이 경우 한글 정보 처리 모듈에서 한글 음의 두음법칙 정용 여부를 파악해 원래 음의 한자를 추출하는 것이 중요하다.

두개이상의 한글 음절을 한자단어로 변환하는 과정에서 가장 중요한 부분은 한자 단어 사전의 구축이다. 본 연구에서는 PC 통신망상에 공개되어있는 한자단어사전을 입수하여 적절한 포맷으로 바꾸어 사용하였다. 또 하나 주의하여야 할 부분은 한자단어사전에 포함되어있지 않은 한글 어절에 대해서는 일대일 변환을 거쳐야 한다. 이때도 하나의 음절 변환시와 동일한 규칙을 적용하여야 정확한 한자변환이 가능[8]하다.

표 1. HanjaConverter 클래스의 대표적 Method

HanjaConverter(String)	생성자
HanjaConverting()	한자 변환 처리
DooEumProcessing()	두음법칙 처리
GetAnother()	이음어 처리
CreateHanjaConvertWindow(String)	한자 버튼 생성

4.3. 한자 변환 및 한자 변환 버튼 생성

추출된 한자가 화면에 표시되어 사용자의 선택을 요구하기 위해서 추출된 한자를 모아 버튼으로 화면에 출력하여야 한다. 사용자가 한자변환 윈도우에서 해당 한자 버튼을 선택하면 선택된 한글 음을 한자로 대체하고 윈도우를 닫는다. 이렇게 동적으로 버튼을 생성하고 출력하기 위해서 필요한 과정은 아래와 같다.

변환된 모든 한자의 개수를 구한다.

한자의 수만큼 new연산자를 이용하여 버튼 배열을 생성한다.

버튼의 텍스트를 n번째 한자로 변환한다.

각 버튼에 ActionListner를 추가하여 입력 이벤트에 대응하게 한다.

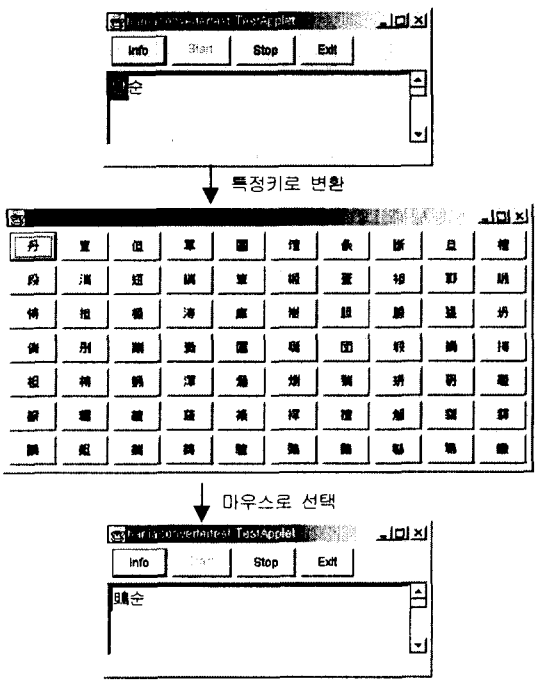

그림 2. 유니코드 한자 변환

5. 결론 및 향후 연구 방향

지금까지 본 논문에서는 유니코드 3.0의 CJK영역에 배정되어있는 한자의 정렬과 모든 유니코드 한자를 입/출력할 수 있으며 운영체제에 독립적인 유니코드3.0 한자 입력 시스템에 대하여 알아보았다.

본 연구에서 제안된 방법으로 유니코드 CJK 한자를 한글음 순으로 정렬한 결과 27,786자의 한자가 정확히 정렬되었고 유니코드를 지원하는 문서 편집기에서 텍스트 형식으로 변환한 한자 변환 사전을 읽은 결과, 한자가 한글음 순으로 정확히 출력되었다. 또한, 개발된 유니코드 한자 입력 시스템은 2만 7천여자의 한자를 자유롭게 입/출력할 수 있고 두음법칙이나 이음어 처리가 가능하여 정확한 한자의 입력할 수 있었다. 3.1의 요구사항 중 글꼴 포함 부분은 사용자의 컴퓨터에 유니코드용 글꼴이 포함되어 있지 않을 경우 CyberBit글꼴을 인터넷상에서 다운로드 가능하게 도와주는 모듈로 대체하였다.

현재는 한자 변환 사전이 속도보다 변환 효율에 중심이 맞추어져서 개발 되었기 때문에 속도 향상을 위하여 자료 구조의 연구가 필요하다. 또한 한국어 순으로만 한자 정렬이 가능하지만 앞으로 일본어와 베트남어 순으로도 한자가 정렬이 가능하게 연구하고, 자바 빈즈 등의 콤퍼넌트로 재구성하여 개발자로 하여금 재활용과 더욱 높은 생산성 향상을 꾀할 수 있게 하고 성능 향상을 통하여 사용자로 하여금 전혀 불편함이 없이 유니코드 한자를 입력할 수 있게 연구되어야 할 것이다. 또한 IIIMP규약[9]을 따라 전세계 어디서나 유니코드 한자를 입/출력할 수 있는 연구가 이루어져야 한다.

[참고문헌]

[1] 유니코드 컨소시움, http://www.unicode.org
[2] MS Word2000, "Word2000 새로운 기능 개요", http://www.microsoft.com/korea/office/enterprise/prodinfo/WordPEG.doc
[3] 윤지헌.변정용, "유니코드2.1기반의 한자 입력방안", 추계한국정보처리학회 논문모음 CD-ROM, 1999
[4] 윤지헌.변정용, "한글 문자열 처리를 위한 클래스 라이브러리", 봄 한국정보과학회 학술발표논문집, 26권, 2호, pp.366-368, 1999
[5] 자바소프트, "Adding Fonts to the Java Runtime", http://java.sun.com/products/jdk/1.1/docs/guide/intl/fontprop.html
[6] Bit Stream, http://www.bitstream.com
[7] Netscape FTP Site, ftp://ftp.netscape.com/pub/communicator/extras/fonts/windows/Cyberbit.zip
[8] 서영진, "한자-한글 변환의 문제 분석", http://kldp.org/KoreanDoc/html/Hanja2Hangul/Hanja2Hangul.html
[9] 자바소프트, "Java Input Method Framework", http://java.sun.com/products/jdk/1.2/docs/guide/intl/spec.html
[10]윤지헌.변정용, "유니코드3.0 한자입력시스템", 봄 한국정보과학회 학술발표논문 CD-ROM
[11]한인, 이용규, 이금석, 홍영식, 한보광, "유니코드 한자처리를 위한 입력기와 편집기의 설계 및 구현", 전자불전 기고논문, 1999
[12]윤용석, "고려대장경의 인테넷 검색 및 열람", 전자불전 특집논문 ,1999
[13]Ken Lunde, "CJKV Information Processing", O'Reilly, 1999