한국어 환경에서 XML을 이용한 다국어정보 입력

정휘웅 윤애선
부산대학교 인지과학협동과정, 부산대학교 불어불문학과
{hwjeong, asyoon}@pusan.ac.kr

Multilanguage data input in Korean environments using XML

Hwi woong Jeong Aesun Yoon
Dept. of Cognitive Science, Dept. of French, Busan National University

요 약

최근 인터넷의 보급은 사용자들에게 많은 다국어 정보를 제공하게 되었다. 그러나 정작 각 국가의 언어를 입력하기 위해서는 자주 자판세트를 변경해야만 하며, 각 국가별 자판 세트가 다르기 때문에 많은 입력 오류를 감수해야 한다. 이를 위해 본 연구진에서는 과거 한국어 환경에서 다국어 지원을 위한 많은 보조 환경을 구축하였으나, 언어 코드의 특성으로 인해 상세한 환경 설정은 전산 전문가의 도움을 통해야 했고, 언어 환경 구축 및 자판 세트 교정에 많은 어려움을 겪었다. 이러한 문제점을 해결하기 위해 본 연구에서는 XML을 이용하여 일반 윈도우기반 컨트롤에서 다국어 정보를 손쉽게 입력할 수 있는 XML DTD와 입력 보조 클래스를 개발하였다. 본 연구결과물을 이용할 경우 일반 언어전문가들이 자신만의 자판 입력세트를 손쉽게 구성할 수 있으며, 이를 운영하는 시스템의 크기도 매우 줄어들어, 전체적인 컴퓨터 운영 효율성을 상승시키는 효과를 거둘 수 있다.

1. 서론

최근 컴퓨터 운영체제의 발달과 인터넷의 보급으로 인해 컴퓨터 사용자들은 과거에 비해 많은 다국어정보를 접하고 있다. 정부기관은 글로벌화에 맞추어 다양한 외국어 사이트를 제공하고 있으며, 인터넷 신문 역시 다양한 외국어 정보를 제공하고 있다. 이에 따라 다국어 정보를 효율적으로 입력하기 위한 개발자의 요구도 급격히 증가하고 있다. 그러나 Windows 95/98 환경에 비해 다국어 입력지원 환경이 많이 개선되기는 했으나, 현재 Windows 2000/ME/XP 환경에서 다국어를 입력하기 위해서는 잦은 자판조합 변경과 각 언어별로 상이한 자판을 익혀야 한다. 실례로 불어의 A키와 Q키, 독어의 Z키와 Y키는 그 위치가 상이하여 신속한 다국어 입력에는 많은 어려움이 따른다. 또한 해당 자판세트의 독특한 조합 기법 (전위표기 및 특수문자 입력방식)은 해당 자판에 익숙하고, 언어의 특성을 이해하는 사용자가 아니면 빠른 다국어 입력은 매우 어려워 사전 개발 및 대용량 데이터베이스 구축을 위해 초급 기술자를 이용한 문서 입력은 거의 불가능하다.

이러한 문제를 해결하기 위하여 본 연구진에서는 Windows 95/98 환경에 기반한 후위표기 기반 다국어 정보 입력 인터페이스 및 다국어를 지원하는 전자사전 개발도구 DiET를 설계하였으나, 자판세트와 입력 로직을 설계하기 위해서 많은 프로그래밍 과정과 이진코드에 기반한 코드 정리 작업이 필요하여, 언어전문가들이 자신의 기호에 맞는 입력 알고리즘 및 인터페이스를 설계할 수 없는 단점이 있었다. 본 연구에서는 이러한 단

점을 개선하고, 보다 손쉽게 다국어 입력 인터페이스를 설계할 수 있는 XML DTD를 설계하였다.

2장에서는 윈도우 환경에서 다국어 입력 환경에 대해 소개하며, 3장에서는 XML 언어를 이용한 다국어 입력 DTD를 소개하겠다. 4장에서는 XML DTD를 이용한 다국어 입력 클래스 알고리즘에 대해 소개하겠다.

2. 윈도우 환경의 다국어 입력 환경

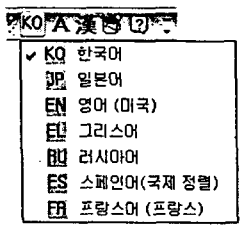
OS의 발전과 소프트웨어의 글로벌화에 대한 요구가 증가함에 따라 과거 각 국가별 버전에서 한정적으로 지원되던 자국언어입력 기능이 점차 보편화되고 있다. 그러나 다국어 입력 환경이 각 국가별 입력자들을 위한 고유 자판세트로 구성되어, 각 국가 언어를 모국어로 사용하지 않는 사용자들이 원하는 문자를 입력하기 위해서는 많은 어려움이 따른다. 따라서 본 절에서는 한글 환경에서 다국어를 입력하는 방법과 그 문제점에 대해 소개하겠다.

2.1. 윈도우에서 한글환경과 유니코드

지금까지 아래아 한글과 같은 워드프로세서에서 제한적인 다국어 입력 기능이 지원되었던 것은 KSC5601 코드에서 지원되던 일부 다국어 코드(러시아어, 그리스어, 일본어, 특수문자군)와 사용자 요구에 따라 자체적인 외곽선 글꼴로 지원되는 일부 문자세트의 도움에 따른 것이었다. 따라서 각 문자를 표현하기 위해 워드프로세서 독자 코드방식을 따라야 했으며, 이는 워드프로세서간 자료 호환성을 가로막는 요인으로 작용하였다.

윈도우 환경에서 다국어 정보를 표현하기 위해서는 입력과 출력, 저장문제가 해결되어야 한다. 현재 윈도우 환경에서 입력은 극동권 문자를 제외하고는 키세트를 변경하는 방식으로 입력할 수 있으며, 극동권 문자는 중국어(간체, 번체), 일본어, 한국어를 위한 개별적인 IME(Input Method Editor)를 제공하고 있다.[그림 1] 출력은 트루타입(Truetype) 글꼴의 개선된 형태인 오픈타입(OpenType) 기술을 이용하여 하나의 트루타입 글꼴이 여럿의 유니코드 언어코드페이지를 지원하는 방식으로 구성되어, 다국어를 표현하기 위해서 각 문자 글립[1] 세트만 존재하는 경우 어렵지 않게 표현할 수 있다.[2] 저장의 경우 유니코드뿐만 아니라 로컬코드(Local Code)로 저장된 문자열들을 자동으로 변환 해 주는 시스템 기반 함수 기능이 지원되고 있어, 언어간 코드 전환에도 커다란 문제점은 현재로서는 없다.

그러나 다양한 사용자의 요구를 만족시켜야 하는 워드프로세서 환경에서는 이러한 코드 호환성의 많은 부분이 제대로 지원되지 않아, 호환성을 고려해야 하는 경우 워드프로세서로 문서를 작성할 때 이용문자에 대해 많은 제약이 있다. 가령 아래아 한글에서 저장하고 있는 고어 및 특수 인쇄용 약물 등은 자체적인 코드를 이용하고 있어, MS-word나 훈민정음과 같은 이기종 워드프로세서와 호환되지 않는다. 또한 심벌형 글꼴의 경우에도 독자적인 글꼴 포맷을 가지고 있어 다른 워드프로세서간 정보 공유가 매우 어렵다. 이는 다른 워드프로세서의 경우에도 그렇게 다르지 않다. 이를 해결하기 위해 RTF (Rich Text File Format)이라는 표준형 교환 정보가 있고, 최근 HTML을 이용한 문서간 호환성 지원도 어느 정도 시도되고 있으나, 각 워드프로세서의 독자적인 기능을 100% 전달하기에는 어려움이 많다.



[그림 1] 윈도우 기반 다국어 입력환경

[1] Glyph : 트루타입 글꼴에서 자료구조적으로 저장되는 하나의 개별 문자의 단위. 글립은 문자의 전체를 표현할 수도 있으나, 한 문자의 일부 영역만을 저장하여, 최종 출력시 조합형식으로 표현할 수도 있다. 한글 글꼴의 크기가 작은 경우 글립 조합방식인 경우가 많다.

[2] 최근 윈도우 2000의 경우 다국어 지원을 위해 유니코드 기반 환경이 구축되어, 손쉽게 다국어 출력기능을 구현할 수 있다.

2.2. 다국어 입력

현재 윈도우 2000/XP 환경에서 다국어를 입력하기 위해서는 매번 자판 세트를 변경 해 주어야 한다. 그러나 이럴 때 마다 자판의 배열이 변경될 뿐만 아니라 조합 문자를 표현하기 위해서는 전위표기 기반 입력방식을 따라야 한다. 가령 불어 자판의 경우 많이 사용되는 é, è, à와 같은 문자는 2, 8, 0번 자판에 연결되어 한 번의 입력으로 원하는 문자를 입력할 수 있다. 그러나 â와 같은 문자는 [^] + [a] 키를 조합해야만 한다. 이 경우 처음 [^]키를 누를 경우 화면상에는 어떠한 문자도 나타나지 않으며, 후속 입력 키가 a나 i와 같이 특수조합 문자를 나타내는 경우는 특수문자를 반환하나, q, h와 같이 특수문자와 관련 없는 문자가 올 경우 ^ 문자와 입력한 문자를 합쳐 두 문자를 동시에 화면에 나타낸다.

이 방식은 다음과 같은 문제점을 포함한다. 첫째, 전위표기 입력시 자신이 어떤 악상을 선택했는지, 두 번째 문자를 입력하기 이전까지는 확인할 수 없어 입력 오류를 발생시킬 가능성이 매우 높다. 둘째, 두 번째 입력 문자에 따라 문자의 반환 수가 변경되기 때문에 사용자의 가독성을 떨어뜨리고, 불어 사용자가 아닌 외국어 사용자의 경우 쉽게 해당하는 언어를 입력하기 어렵다. 셋째, 규칙성을 가진 입력 방식이 아닌 자주 사용되는 문자는 1회 입력, 자주 사용하지 않는 문자는 2회 이상 입력을 해야 하므로, 다국어 입력작업을 수행하는 사용자의 입력 오류율을 상승시킬 가능성이 높다.

따라서 이러한 문제점을 해결하기 위해 본 연구에서는 과거 다국어 입력을 위해 후위표기에 바탕한 조합형 다국어 입력 알고리즘을 개발한바 있다. 후위표기의 경우 완전한 규칙 기반 다국어 입력 방식이기 때문에, 모국어 사용자에게는 추가적인 입력방식일 뿐만 아니라, 자주 사용하는 문자 역시 조합형으로 입력해야 하므로, 결과적으로는 자판 입력 횟수를 상승시키는 단점이 있다. 그러나 본 연구진에서 개발한 후위 표기는 각 언어를 모국어로 이용하지 않는 각 국가별 입력자들이 다른 언어정보를 쉽게 입력할 수 있도록 규칙 기반 입력 방식을 개발 한 것이며, 인간이 문자를 입력할 때 알파벳을 쓰고, 악상 관련 기호를 입력하는 방식과 유사하여, 해당 언어를 깊이 모르는 사용자라도 규칙에 의거하여 다른 국가의 언어를 손쉽게 쓸 수 있다는 장점이 있다. 또한 다국어뿐만 아니라 국제발음기호 및 특수 기호를 입력할 수 있는 입력 방식을 개발할 수 있어, 멀티미디어 컨텐츠 제작 및 특수 문자 입력환경을 개발하는데 이용될 수 있다.

3. XML을 이용한 한국어 및 다국어 입력

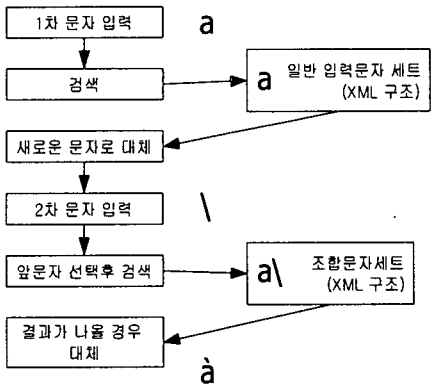
앞 절에서 소개하였듯이, 본 연구진에서는 다국어 입력을 위한 알고리즘, 그리고 이를 이용한 다국어 입력 인터페이스를 개발하였었다. 그러나 본 연구진에서 이미 개발했던 다국어 입력 인터페이스는 자판세트를 설정하기 위해서 이진코드 기반의 조합 세트를 만들었기 때문에, 각 언어별 코드세트를 만들기가 어려웠으며, 입력 오류(반복 정의, 중복 정의) 탐지 및 코드세트의 확장을 위해서는 많은 이진코드 작업과 소스코드 수정 작업이 요구되었다.

이러한 문제점을 해결하기 위해서 본 연구에서는 XML에 기반하여 다국어 입력 알고리즘을 손쉽게 구현하고 다른 시스템에도 적용이 용이한 DTD와 이를 이용하는 다국어 입력 환경을 개발하였다. 다국어 입력기를 XML 기반으로 구성할 경우, 자판 세트의 저장을 XML 엔진에 위탁함으로써, 소스코드의 크기를 대폭 줄일 수 있으며, 원하는 문자를 빠르게 검색할 수 있다.

3.1. 자판 입력 알고리즘

문자의 자판 입력은 후위표기 기준을 따르며, 처음 문자가 입력되었을 때 자신과 대조되는 문자세트를 저장한다. 처음 문자는 ASC 코드를 기준으로 검색하며, 그에 상응하는 코드를 attribution 형태로 출력하여 입력된 문자를 대체한다. 두 번째 입력되는 문자는 자신의 앞에 입력된 문자를 다시 읽어 조합 문자 세트에 저장된 정보와 비교한다. 이 때 새로운 문자정보가 반환될 경우

이전 문자와 입력된 문자를 무시하고 새롭게 검출된 문자를 대체한다.



[그림 2] 후위표기 입력 알고리즘 (예: 불어)

이러한 구조는 키보드에서 입력된 문자를 hooking 한 뒤, 입력 컨트롤로부터 전달되는 제한된 정보를 이용하여 손쉽게 문자 정보를 변환할 수 있기 때문에 간단한 기능을 제공하는 일반 입력창에 손쉽게 알고리즘을 적용할 수 있는 장점이 있다. 또한 조합 횟수를 반복시킬 경우 조합 기능을 요구하는 문자들(Greek Polytonic Character, Gujarat character 등)을 쉽게 입력할 수 있는 장점이 있다.

3.2. XML용 DTD 정의

XML은 전산환경 전반에 있어 많은 변화를 가져오고 있다. 더욱이 전산 전문가가 아닌 일반 정보구조 전문가들이 쉽게 자신의 정보 구조를 설계할 수 있는 길을 열어 준 것은 큰 의의가 있다. 그러나 XML이 시스템과 연동되어 시스템 운영에 영향을 주는 설계는 아직까지 많지 않다. 그러나 본 연구에서는 XML을 시스템 설계 과정에 도입하였으며, 실제로 기존에 개발했던 다국어 입력 환경에 비해 그 크기를 1/5수준으로 줄일 수 있었다.
다국어 입력을 지원하기 위한 XML DTD는 attribution 기반 접근법을 이용하여 정의하였다. 이는 본 XML 코드 정보의 depth가 크지 않으며, 서술형 정보가 거의 없기 때문이다. 다음은 XML문서의 정의 DTD다.

```
<!ELEMENT GKM (KM, CKM)>
<!ELEMENT KM (K+)>
<!ELEMENT CKM (CK+)>
<!ELEMENT K>
<!ELEMENT CK>
<!ATTLIST KM
        lang    id      #required
        ef      cdata   #required
        kf      cdata   #required>
<!ATTLIST CKM
        lang    id      #required>
<!ATTLIST K
        ka      id      #required
        kc      id      #required>
<!ATTLIST CK
        fc      id      #required
        sc      id      #required
        rc      id      #required>
```

각 태그와 속성에 대한 설명은 [표1]과 같다.

[표 1] XML 각 속성값의 의미

값	유형	설명
GKM	E	전체 XML 문서의 루트
KM	E	1회 입력 키맵
CKM	E	조합 키맵
K	E	각 1회 입력키
CK	E	각 조합 입력키
lang	A	언어값
ef	A	영어명칭
kf	A	한글명칭
ka	A	아스키코드 키보드 입력값
kc	A	실제 출력코드값
fc	A	첫번째 입력 문자값
sc	A	두번째 입력 문자값
rc	A	결과값

* E: Element, A: attribute

여기서 lang값은 조합키맵과 1회 입력 키맵을 연관시키는 매개체로서, 만약 두 개의 태그가 동일한 언어임을 설명하기 위해서는 언어정의 코드가 동일해야 한다. 각 언어정의 코드는 사용자가 자유롭게 정의할 수 있으며,

영어 명칭이나 한글 명칭 또한 모국어를 달리 하는 사람들에 의해 자유롭게 변경할 수 있다. 다음은 [표1]에 정의된 XML DTD를 이용하여 정의된 키세트의 예다.

```
<GKM>
        <KM lang='FRN' ef='French' kf='불어'>
                <K ka='32' kc=' ' />
                <K ka='33' kc='!' />
                <K ka='34' kc='"' />
                ......
        </KM>
        <CKM lang='FRN'>
                <CK FC='a' SC='^ ' RC='â' />
                <CK FC='E' SC='`' RC='È' />
                <CK FC='e' SC='`' RC='è' />
                <CK FC='E' SC='/' RC='É' />
                ......
        </CKM>
</GKM>
```

문서의 시작은 XML 문서이기 때문에 하나의 root 태그를 정의하며, 내부 문서의 KM 태그와 CKM 태그는 순차적으로 나열된다. CKM 태그가 KM 태그 이전에 올 수는 없으나, CKM 태그가 나열된 순서는 반드시 KM 태그의 나열 순서와 일치할 필요는 없다. 가령 KM 태그의 키세트가 영어-불어-독어 순으로 되고 CKM 태그의 키세트가 불어-영어-독어 순으로 되어도 문서 정의 규칙에는 어긋나지 않는다.

4. XML DTD를 이용한 다국어 정보 입력

앞 절에서 소개한 XML 문서에 정의된 문자를 입력하기 위해서 본 연구에서는 최소한의 정보를 바탕으로 다국어정보 입력환경을 구현할 수 있는 입력보조 엔진을 구축하였다. 다국어 입력기는 앞 절에서 소개한 XML 기반 키보드 코드 관리기와, 입력되고 있는 문자에 대한 정보를 저장하는 언어모드 관리기로 분류된다. 본 엔진이 구동되기 위해서는 원본 컨트롤이 자신의 현재 커서 위치와 입력 키보드 값, 발생 이벤트에 대한 정보를 전달 해 주어야 한다. 전달된 값을 바탕으로 입력 처리 엔진은 선별적으로 반응하여, 컨트롤에서 발생하는 이벤트를 조작하여 새로운 문자값을 반환한다.

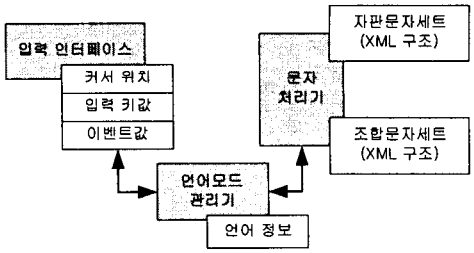


그림 3 다국어 입력기 구조

이 때 일반적인 프로그래밍 기법처럼 message hooking 기법을 이용하여 이벤트를 낚아채는 방식을 이용하지 않은 것은, 다른 운영체제는 상이한 API 환경을 지원하기 때문에 코드 이식성이 떨어지기 때문이다. [그림4]는 이러한 구조에 바탕 하여 기존 윈도우 텍스트 입력기를 응용한 다국어 입력 인터페이스를 나타내고 있다. 본 입력 컨트롤은 기존 텍스트 입력기를 객체지향 프로그래밍 모델의 Inheritance 특성을 이용하여 내부 입력창과 언어 선택 메뉴를 추가하고, 사용자의 편의를 위해 마우스 오른쪽 버튼 메뉴에 자판세트 보기와 마우스의 클릭 방식을 통한 문자 입력 기능을 지원하도록 구성하였다.

그림 4 다국어 입력 인터페이스

본 입력 컨트롤이 한글 환경에서 구동 될 경우, 모든 IME 환경들은 (한국어, 일본어, 간체 중국어, 번체 중국어) 하나의 문자 혹은 문자열(일본어의 경우)이 완성되기 이전에는 시스템으로 새로운 글자가 입력되었다는 메시지를 발송하지 않기 때문에, 한글 입력과정에서 후위표기 혹은 전위표기로 인해 문자 입력 알고리즘상 문제를 발생시키지 않는다. 따라서 IME 모드에 따라 다국어 입력 모드는 켜짐과 꺼짐이 자동으로 전환된다.

그리고 초보 입력자를 위해 별도의 입력 자판셋을 나타내는

5. 결론 및 향후 연구

본 연구에서는 XML이 웹을 중심으로 하는 정보의 표현뿐만 아니라 시스템 구축을 보다 용이하게 하고, 코드의 수를 줄일 수 있으며, 손쉬운 관리기능을 제공해 주는데도 이용될 수 있음을 보였다. 본 연구의 결과는 커서의 위치를 설명하는 포인터가 지원되는 입력창이나 인터페이스만 있을 경우 어느 시스템과도 연동될 수 있어, 웹 환경이나 운영체제를 초월하여 손쉽게 구현될 수 있으며, 적은 용량으로 시스템 부하를 혁신적으로 줄여준다.

그러나 이러한 연구 결과에도 불구하고, 본 연구는 다음과 같은 부분에 있어 좀 더 심도 있는 연구가 필요하다. 비록 많은 초급 입력자들을 통해 짧은 시간내에 숙달된 입력이 가능함을 관찰할 수 있었으나, 이것이 기존 모국어 자판 입력자들에 비해 어느 정도의 효율성을 가지고 있는지 검증이 되지 않아, 이에 대한 연구가 필요하며, 아직까지 윈도우 환경에만 국한된 XML 기반 검색 알고리즘으로서, Java나 일반 웹 기반 스크립트 언어와 손쉽게 연동될 수 있는 방안을 강구해야 할 것으로 본다.

6. 참고 문헌

[1] Microsoft corp, Global software development

[2] Kano, Nadine, Developing International Software for Windows 95 and Windows NT, Microsoft Press, 1995

[3] The Unicode Consortium, The Unicode Standard-version 2.0, Addison Wesley, 1996

[4] premium.microsoft.com _ International Support in Window NT 5.0(/msdn/labrary/conf/pdc97/intl_supnt_5.htm)

[5] www.w3c.org - XML Specification

[6] www.unicode.org _ Unicode organization

[7] 김경석, 컴퓨터 속의 한글 이야기, 영진출판사, 서울, 1995

[8] Lauren, Simon, LT, XML Primer, IDG Books, New York, 1998

[9] Pardi, William J., XML in Action, Microsoft Press, Redmond, 1999