WWW에서의 한국어 표준 음가생성 시스템 구현

○
임재걸 이계영 남중구
동국대학교 컴퓨터학과

An Implementation of A Korean Standard Ponetic Value Generating System

Jaegeol Yim Gyeyoung Lee Joonggoo Nam
Dept of Computer Science, Dong-guk University
{yim, leegy, junggu}@wonhyo.dongguk.ac.kr

요 약

본 시스템은 한글 발음 교육 사이트 개발 프로젝트의 일부인 음가 생성에 관한 컴포넌트로써 한국어 표준발음 테이블에서 음운 변동값을 추출하고 해당 음성과 입모양을 출력하는 WWW상의 자바 애플릿 프로그램 개발에 관한 연구이다. 본 논문에서는 형태소 분석에 선행되어야 할 전처리 과정, 예외처리, 음가 생성부에서 고려하여할 점과, 시스템의 애플릿 구현 등에 대해 중점적으로 기술하였다.

1. 서론

문교부에서는 표준 발음법을 제정해 1998년 1월 19일 고시하였다[1]. 그러나 그 조항이 복잡할 뿐만 아니라 중복되거나 충돌하는 항들이 있어 그것을 숙지하고 사용한다는 것은 어려운 일이다. 본 논문에서는 접근성이 용이한 WWW에서 사용할 수 있도록 JAVA 애플릿으로, 표준 음가를 생성하는 시스템을 구현하였다.

본 논문은 2절에서 기존의 연구 및 배경지식을, 3절에서 시스템의 설계와 구현을 소개하고, 4절에서 결론과 향후과제에 대해 설명한다.

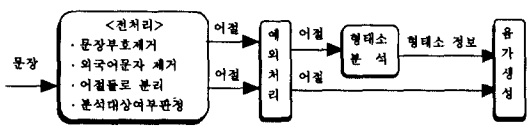
2. 기존의 연구 및 배경 지식

지금까지 if-else로 자모와 형태소를 비교해 음운변동을 처리하여 표준발음 음가를 생성시키고자 하는 연구가 있어왔다[3]. 이러한 방법은 많은 비교문을 사용하고 그만큼 많은 프로그램 코드를 필요로 한다.

본 시스템에서는 행렬을 이용해 자모의 코드값을 인덱스로 하여 직접 변동값을 테이블에서 추출할 수 있도록 구현하였다.

3. 시스템의 설계 및 구현

본 시스템은 (그림 2)에서와 같이 전처리부, 예외 처리부, 형태소 분석기, 음가 생성부로 나누어진다. 또한 이 시스템은 한글 발음의 교육을 목적으로 만들어졌으므로, 사용자의 입력처리, 형태소 분석, 음가생성, 음성 및 동영상 출력뿐만 아니라 수많은 html문서와의 통신이 필요하게 된다. 이러한 모든 요소들을 처리할 수 있으면서도 웹 상에서 동적인 프로그래밍이 가능한 Java를 시스템 구현에 이용하였고 SUN-SPARC상에서 JDK1.2를 사용하였다.



(그림 2) 전체 시스템의 구조

3.1. 전처리 과정

시스템의 입력단위는 문장으로써 전처리 과정에서 문장부호와 한글이 아닌 문자를 제거하고, 문장을 어절들

로 분리하며, 분석대상여부의 판정을 하게 된다. 여기서 분석대상여부의 판정이란 분리된 각각의 어절이 형태소 분석을 필요로 하는지를 미리 검사해 형태소분석에서 사전탐색 등의 시간을 줄이기 위해 하는 처리를 말한다. 규칙에서 형태소정보가 필요한 조항들은 대부분 모음으로 시작될 것을 요구하고 있고, 접미사 '히'인 경우와 ㄱ, ㄷ, ㅅ, ㅈ일 경우가 각각 하나씩 있으므로 이러한 것을 포함하지 않는 어절일 경우 형태소 분석기를 거치지 않게 함으로써 효율을 높일 수 있다. <표 1>에 각각의 경우에 대한 조항에 대해 언급하였다.

<표 1> 형태소분석을 필요로 하는 조항이 요구하는 조건별 분류

- 모음으로 시작 : 12.4, 13, 14, 15, 17, 29, 30.3항
- 접 미 사 '히' : 17항 붙임
- 어미 ㄱ,ㄷ,ㅅ,ㅈ : 24, 25항

3.2 예외 처리부

표준 발음법에는 예외가 있어서 예외에 해당하는 어절에 대해서는 형태소 분석과정 없이 바로 음운변동의 결과를 얻어낼 수 있다. 예외로 처리되거나 예외를 포함하는 조항은 <표2>와 같다. 26항과 29항은 형태소분석이나 사전검색보다 예외로 처리하는 것이 효율적인 경우이다. 26항의 처리를 위해 사전검색 방법을 선택할 경우, 수 십만개에 해당하는 명사에 대해 한자어인지 아닌지에 대한 정보를 기록하여야 하는 문제가 있다. 국어사전을 통해 'ㄹ'받침 뒤에 'ㄷ, ㅅ, ㅈ'이 연결되는 한자어 1826개를 예외사전에 수록하고 예외로 처리하였다.

28항은 '눈동자'와 같이 '눈의 동자'라는 관형격 기능을 가지는 복합어의 경우인데 이 조항의 처리를 위해서는 의미 분석의 단계를 거쳐야 하므로 예외로 처리하는 것이 효율적이다.

<표 2> 예외 처리된 조항들

- 예외를 포함하는 항 : 10항, 16항, 20항, 29항 단서
- 예외로 처리하는 것이 효율적인 항 : 26항, 28항, 29항
  - 제 26항 : 한자어에서 'ㄹ'받침 뒤에 연결되는 'ㄷ, ㅅ, ㅈ'은 된소리로 발음한다.
  - 제28항 표기상으로는 사이시옷이 없더라도, 관형격 기능을 지니는 사이시옷이 있어야할(휴지가 성립되는) 합성어의 경우에는, 뒤 단어의 첫소리 'ㄱ, ㄷ, ㅂ, ㅅ, ㅈ'을 된소리로 발음한다.
  - 제 29항 : 합성어 및 파생어에서, 앞 단어나 접두사의 끝이 자음이고 뒤 단어나 접미사 첫 음절이 '이,야,여,요,유'인 경우에는 'ㄴ'음을 첨가하여[니,냐,녀,뇨,뉴]로 발음한다.

29항은 조항에서 예로 제시한 담요(毯-)의 경우를 보면 담(毯)자는 담요 '담'자로써 '담요, 모포, 털로 짠 깔개'라는 뜻을 가지고 있다. 이 예에서 앞 단어 '담'의 끝이 자음이고 뒤 단어 '요'가 '이, 야, 여, 요, 유' 인 경우이므로 '요'에 'ㄴ'음을 첨가하여 '뇨'가 되었다. 그러나 여기서 한 글자로 이루어진 한자어를 어떻게 단어로 인식할 것인가가 문제가 된다. 해결책으로 대용량의 한자어 데이터베이스를 구축하는 방안을 생각해 볼 수 있는데 이 방법은 다음과 같은 몇 가지 문제점이 있다.

- 그것이 한자어로 쓰였는지 순 우리말로 쓰였는지 명확하게 구분하기가 어렵다.
- 검열(檢閱)과 같은 경우에는 [검녈/거멸]과 같이 규칙의 적용을 받게 되지만 간이(簡易), 강요(强要), 강유(剛柔 : 강함과 부드러움)와 같은 경우는 규칙의 적용을 받지 않는다.
- 대용량의 한자어 데이터베이스를 구축하여 어렵게 탐색에 성공한다 하더라도 그것으로 처리가 종결되는 것이 아니라 다른 명사 사전을 같이 병행해 검색해야 하고, 복합어이므로 연결되는 다른 단어에 대해서도 같은 처리를 해 주어야 한다.
- 성공한 단어가 해당규칙에 적용되는 한자어인지와 규칙에 해당되는 형태를 가지는지를 확인해야 한다.
- 많은 붙임 조항과 단서 처리를 병행해야 한다.

3.3 형태소 분석

표준발음법의 적용을 위한 형태소분석은 그 범위가 크므로 별도의 논문에서 다루기로 한다. 형태소 분석을 통해 어절은 형태소단위로 분리되고 그것은 표준발음법의 적용을 위해 필요한 정보인 용언활용형, 조사, 어간, 어미, 접미사, 실질형태소, 피동/사동의 접미사 '기', 사이시옷 등의 정보를 가지게 된다.

3.4 음가 생성부

음가 생성부에서는 규칙을 저장하고 있는 3차원 행렬을 탐색하게 되는데 표준발음법의 조항들을 어떻게 정확하고 포괄적으로 테이블에 수용했으며 어떻게 생성되었는지에 대해서는 별도의 논문에서 다루기로 한다.

표준 발음 음가테이블의 일부를 <표 3>에서 보였는데 효율적인 탐색을 위해 모든 규칙을 재적용 시킬 필요 없이 한번의 테이블 탐색을 통해 음운 변동값을 추출할 수 있도록 설계되어져 있다. 여기서 인덱스는 유니코드 한글 자음의 부코드(subcode) 값이어서 음절들을 분리해 그것을 바로 테이블의 인덱스로 사용하고, 테이블에서 추출된 값들을 자모 조합을 통해 음절들로 쉽게 합성할 수 있는 이점이 있다.

테이블의 구조를 간략히 설명하면, 세로축은 종성의 코드값을, 가로축은 다음글자의 초성 코드값을 나타낸다. 가로축과 세로축이 만나는 각 셀에는 종성과 다음글자의 초성이 만났을 때의 음운변동 결과값이 기록되어 있다. 이것을 표현하면 다음과 같이 나타낼 수 있다.

- 테이블 = {rule[i][j][k] = 자음의 음가를 나타내는 부코드 값}
  (j=종성의 코드값, k=다음음절 초성의 코드값,
  i가 0일때 : rule결과값은 변화된 종성의 코드값,
  i가 1일때 : rule결과값은 변화된 다음 음절 초성의 코드값,)

종성의 변화값은 셀의 아래쪽에(i=0), 초성의 변화값은 셀의 위쪽(i=1)에 명시하였다. 즉, 2행 0열의 경우 종성 ㄲ은 ㄱ으로, 초성 ㄱ은 ㄲ으로 변함을 의미한다.

<표 3> 표준 발음법의 음운 변동을 표현한 3차원 행렬표

종성 \ 다음음절의 초성		종=0 초=1	ㄱ	ㄲ	ㄴ	…	어말	모음조사 어미 접미	모음 어미 접미	ㅏ ㅓ, ㅗ ㅜ, ㅟ 실질 형태소	ㅣ조사접미사	허접미사
			0	1	2		19	20	20	21	22	20
-	0	초						ㅇ11 -0		ㅇ11 -0	ㅇ11 -0	ㅎ18 -0
		종										
ㄱ	1	초	ㄲ1					ㄱ0 -0		0ㄱ -0	ㅇ11 ㄱ1	ㅎ18 ㄱ1
		종			ㅇ21							
ㄲ	2	초	ㄲ1					ㄲ1 -0		ㄱ0 -0	ㅇ11 ㄱ1	ㅎ18 ㄱ1
		종	ㄱ0	ㄱ 1	ㅇ21		ㄱ 1					
ㅆ	3	초	ㄲ1					ㅆ10 ㄱ1		ㄱ0 -0	ㄲ1 ㄱ1	ㅎ18 ㄱ1
		종	ㄱ0	ㄱ 1	ㅇ21		ㄱ 1					
ㄴ	4	초						ㄴ2 ㅇ21		ㄴ2 -0	ㅇ11 ㄴ4	ㅎ18 ㄴ4
		종										
(중략)												
어간받침 ㄹ	30	초	ㄲ1 ㄹ8	ㄲ1 ㄹ8	ㄴ2 ㄹ8		(-1) ㄹ8	ㅇ11 ㄹ8		ㅇ11 ㄹ8	ㅇ11 ㄹ8	ㅎ18 ㄹ8
		종										
어간받침 ㄴㅈ	28	초	ㄲ1 ㄴ4	ㄲ1 ㄴ4	ㄴ2 ㄴ4		(-1) ㄴ4	ㅇ11 ㄴ4		ㅇ11 ㄴ4	ㅇ11 ㄴ4	ㅎ18 ㄴ4
		종										
어간받침 ㅁㅃ	29	초	ㄲ1 ㅁ16	ㄲ1 ㅁ16	ㄴ2 ㅁ16		(-1) ㅁ16	ㅇ11 ㅁ16		ㅇ11 ㅁ16	ㅇ11 ㅁ16	ㅎ18 ㅁ16
		종										
어간받침 ㄹ	30	초	ㄲ1 ㄹ8	ㄲ1 ㄹ8	ㄴ2 ㄹ8		(-1) ㄹ8	ㅇ11 ㄹ8		ㅇ11 ㄹ8	ㅇ11 ㄹ8	ㅎ18 ㄹ8
		종										

예를 들어 '닦다' 의 경우 <표 3>에서 2행 3열에 해당하는데 ㄲ이 ㄷ과 만났을 때 10항의 받침의 발음 규칙에 의해 종성 ㄲ은 대표음 [ㄱ]이 되고, 23항의 경음화 규칙에 의해 초성 ㄷ은 경음화 형상으로 ㄸ이 되는 것을 보이고 있다. 즉 rule[0][2][3] = 'ㄱ', rule[1][2][3] = 'ㄸ'이다. <표3>은 테이블에서 쉽게 음운 변동값을 추출하는 의사코드를 보이고 있다.

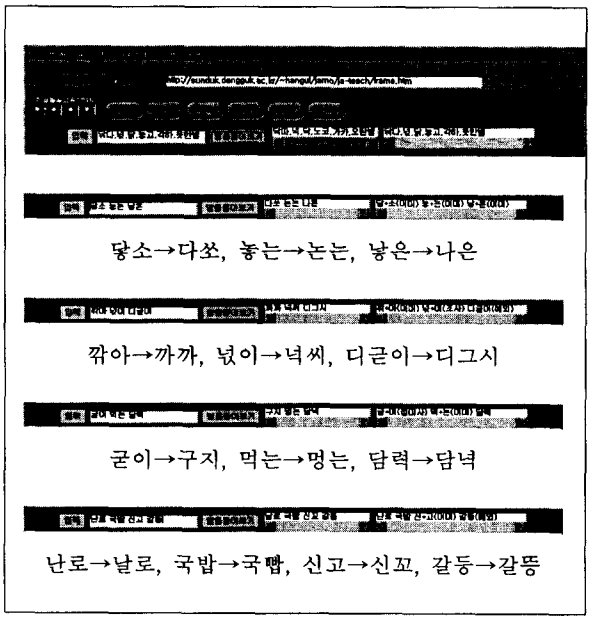
<표 4> 테이블에서 음운 변동값을 추출하는 의사코드

```
protected static int[] jamoConvert(int[] jamos){
    int jongsung, chosung;
    int jongsungPosition, chosungPosition;
    for(자모의 배열이 모두 처리될 때까지) {
        jamos[jongsungPosition]=rule[0][jongsung][chosung];
        jamos[chosungPosition] = rule[1][jongsung][chosung];
    }
    return jamos;
}
```

이렇게 추출된 자모는 음절들로 합성되고 그 음가에 해당하는 동영상과 음성파일을 로딩해 출력한다.

3.5 실험결과

표준발음법의 조항에서 예로 제시한 예제 단어들을 이용해 실험해 보았다. (그림 3)은 실험의 일부로써 조항들마다 하나씩 단어를 실험해보았다. 맨 위의 그림은 넷스케이프에서 실험한 것이고 나머지는 지면상 애플릿뷰어에서 실행한 결과이다. 본 시스템의 발음교육 대상이 유아이므로 초등학교 1학년에서 3학년까지의 교과서에 대해 실험한 결과 완전한 음운변동결과를 얻을 수 있었다



(그림 3) 실험결과

4. 결론

본 시스템은 한글 발음 교육 사이트 개발 프로젝트의 일부인 음가 생성에 관한 컴포넌트로써 한국어 표준발음 테이블에서 음운 변동값을 추출하고 해당 음성과 입모양을 출력하는 WWW상의 자바 애플릿 프로그램 개발에 관한 것으로써 표준발음법에 의해 정확한 음가를 생성할 수 있도록 구현하였다. 더욱 효율적인 발음교육을 위해 음운변동의 이유와 원리를 포함하는 것과 외국인이 학습할 수 있도록 국제 발음 기호로 표기하는 등의 연구가 필요하다.

참고문헌

[1] 문교부, "표준어 규정", 문교부 고시 제88-2 호, 제2부 표준어 발음법, 1988.
[2] URL, http://java.sun.com/docs/white/langenv/Intro.doc1.html#943
[3] 이계영, 음성 처리를 위한 한국어 자료 구성에 관한 연구, 1992