ISO/IEC 14651 틀에서 한글 간추리기 문제점에 대한 해결 방안

옥 제영*, 정 일동*, 김 경석**
* 부산대학교 전자계산학과
** 부산대학교 정보·컴퓨터 공학부

A solution for the problems of Collating Hangeul in the framework of ISO/IEC 14651

Jae-Young Ok, Il-dong Jung, Kyongsok KIM
Dept. of Computer Science, Pusan National University

요 약

국제 표준인 ISO/IEC 14651(International String Ordering)은 글자의 차례를 정하고 문자열(=글자떼)을 간추리는 틀에 관한 표준이다. ISO/IEC 14651을 사용하면, 글자의 차례를 바꾸기 위하여 프로그램을 바꾸지 않고, 공통 틀 표라고 하는 표만 수정하면 글자의 차례를 바꾸어 간추릴 수 있다. ISO/IEC 14651에서 한글과 다른 나라 글자가 섞여 있는 글자떼를 간추리면 제대로 된 결과가 나오지 않는다. 이 문제를 해결하기 위하여, 한글 글자마디를 첫소리, 가운데 소리, 끝소리 글자로 바꾼 뒤 비교하는 방안을 제안한다.

1. 서 론

일반적으로 문자열(=글자떼)을 간추리는 (ordering) 프로그램은 간추릴 때 사용하는 글자의 순서를 내장하고 있다. 그러므로, 새로운 간추리는 차례를 적용하려면 글자떼를 간추리는 프로그램을 고쳐야 한다. 특히 같은 글자라도 나라에 따라 차례가 다른 유럽에서는 각 나라 별로 글자떼를 간추리는 프로그램을 유지하고 배포해야 한다. 이런 불편을 줄이기 위하여, 글자떼를 간추리는 프로그램은 바꾸지 않고, 글자 차례를 나타내는 자료 파일만 바꿈으로써 각 나라의 글자 순서를 반영할 수 있는 방식이 바로 ISO/IEC 14651이다. 캐나다에서 처음으로 만든 안이 바탕이 되어 ISO/IEC 14651 이 나오게 되었다.

ISO/IEC 14651은 2000년 4월에 발표된 국제 표준이며, 정식 명칭은 'ISO/IEC 14651 - International String Ordering and Comparison - Method for Comparing Character Strings and Descriptions of the Common Template Tailorable Ordering' 이다 [1].

ISO 14651은 글자의 순서를 공통 틀 표 (Common Template Table) 라는 표에 정의하기 때문에 간추리는 (ordering) 프로그램은 바꿀 필요가 없으며, 공통 틀 표에 나타난 글자의 차례를 바꾸기만 하면 글자떼를 간추리는 차례를 바꿀 수 있다 [3]. 보기를 들어, 남북의 가나다 차례도 ISO 14651의 공통 틀 표를 이용해서 쉽게 바꿀 수 있는 장점을 가지고 있다.

문서 내의 글자떼들은 한글로만 이루어져 있지 않고, 한글과 여러나라 글자가 섞여서 같이 나오는 경우가 있다. 이러한 경우를 나타내면 다음과 같다.

1) 완성형 한글 + 여러나라 글자
2) 첫가끝 조합형 한글 + 여러나라 글자
3) 완성형 한글과 첫가끝 조합형 한글 + 여러나라 글자

위의 경우 중 3)은 그림 1에서 보듯이 같은 글자에 대해 다른 부호값을 가지고 있다.

```
가 (U+1100 1161) : 조합형 한글
가 (U+AC00)      : 완성형 한글
```

그림 1. 같은 글자마디에 대한 부호값

이러한 경우에 어느 한 방향의 한글로 바꾸지 않으면 간추리기가 잘 되지 않는다. 완성형 한글은 요즘 글자마디 11,172개만 나타낼 수 있는데 비해, 첫가끝 조합형 한글은 요즘 한글뿐만 아니라, 옛 한글까지 나타낼 수 있다. 그러므로, 완성형 한글은 첫가끝 조

합형 한글로 바꾸어 주어야 한다. 그래서 3)은 2)의 경우와 같아진다. 1)을 간추리면 그림 2와 같이 한글의 글자마디가 1개의 부호값으로 나누어져서 비교되므로 간추리기에 문제가 없다.

```
가      (U+AC00)
가 A    (U+AC00 0041)
각      (U+AC01)
가 方   (U+AC00 4E07)
```

그림 2. 완성형 한글 + 여러나라 글자

그런데 2)와 3)을 간추리면 한글 글자마디가 2개의 부호값 또는 3개의 부호값으로 비교되므로 한글 다음에 어떤 부호값의 글자가 오는가에 따라 간추리는 결과가 달라진다. 이러한 경우에 대한 보기를 들면, 다음과 같다.

```
ㄱ ㅏ     (U+1100 1161)
ㄱ ㅏ A   (U+1100 1161 0041)
ㄱ ㅏ ㄱ  (U+1100 1161 11A8)
ㄱ ㅏ 万  (U+1100 1161 4E07)
```

그림 3. 여러나라 글자가 섞인 글자떼

그림 3에서 각 글자떼의 간추리는 차례를 결정하는 방법은 글자떼들 ('가', '가A', '가方', '각') 의 각 글자들의 부호값들을 차례로 비교해서 크기 순서로 결정한다. 여기에서 각 글자들의 'ㄱ', 'ㅏ'에 대한 부호값은 같지만, 다음 글자에서 부호값의 차이가 있으므로 부호값의 크기로 간추리기를 하면 기존 남한의 한글 가나다 차례와 다른 '가', '가A', '각', '가方'과 같은 순서로 된다.

이 문제에 관한 연구가 현재까지 진행된 바가 없으며, 따라서 남한의 한글의 가나다 차례에 맞는 간추리기를 위해서는 이에 대한 연구가 필요하다.

본 논문에서는 첫가끝 조합형 한글과 다른 나라 글자가 섞여 있는 글자떼를 간추릴 때 발생하는 문제점을 지적하고, 그 해결 방안을 제안한다.

2장에서는 ISO 14651 틀을 써서 간추리는 방법을 정리하고, 3장에서는 ISO 14651에 관한 기존 연구들을 살펴본다. 4장에서는 ISO 14651 틀에서 첫가끝 조합형 한글과 다른 글자가 섞여 있을 때의 간추리기 문제점을 지적하고, 해결 방안을 제안한다. 5장에서는 결론 및 향후 연구 과제를 제시한다.

2 ISO 14651 틀을 써서 글자떼 간추리기

ISO 14651은 글자떼의 차례를 비교하기 위하여 각 글자떼마다 key를 만드는데, level은 보통 4이다. 한글의 경우 level 1만 쓰면 되는 것으로 보인다. 그러므로 level 1에서 필요한 용어외의 전문적인 용어들의 설명과 level 1을 제외한 나머지 level들에 관한 구체적인 내용은 본 논문에서 다루지 않는다. ISO 14651에 따른 정확한 방식은 [1]을 참조하도록 한다. 이 논문에서는 level 1에 UCS 부호값을 편의상 쓴다.

ISO 14651 틀을 써서 첫가끝 조합형 한글을 간추리는 보기를 들면 다음과 같다.

```
        level1
거  <U1100><U1165>
기  <U1100><U1175>
구  <U1100><U116E>
```

그림 4. 첫가끝 조합형 한글 간추리기

그림 4에서 첫가끝 조합형 한글 '거', '기', '구'의 level 1은 공통 틀 표에 있는 'ㄱ', 'ㅏ', 'ㅣ', 'ㅜ'의 글자들을 조합한 것이다.

한글의 글자마디에 대한 간추리는 차례는 첫소리 글자, 가운데 소리 글자, 끝소리 글자 순서로 부호값들의 크기를 비교해서 결정하는데 그림 4에서 첫소리 글자의 부호값은 같기 때문에 가운데 소리 글자의 부호값을 비교한다. 글자마디 '거'의 가운데 소리 글자 부호값이 비교하는 다른 글자마디의 가운데 소리 글자 부호값보다 작기 때문에 '거'가 가장 먼저 나오고, '구', '기' 차례로 간추리기가 된다.

3. 관련연구

[4]는 첫가끝 한글과 완성형 한글이 섞여 있는 글자떼를 간추리면 첫가끝 조합형 한글이 완성형보다 먼저 나오는 결과가 올바르지 않음을 지적하고, 완성형 한글에 대해 첫가끝 조합형 한글을 나타낼 것을 제안하였다. 이 방법은 ISO 14651의 기본 틀을 벗어나지 않으면서 문제를 해결하고 있지만, 한글과 여러 나라 글자가 섞여 있는 상황은 고려하지 않았다.

이 논문에서는 ISO 14651 틀 안에서 첫가끝 조합형 한글과 여러 나라 글자가 섞여 있을 때, 글자떼를 제대로 간추릴 수 있는 방법을 제안한다.

4. 첫가끝 조합형 한글 간추리기

4.1 문제점 지적

ISO 14651에서 첫가끝 조합형 한글과 여러나라 글자 (한자, 영문자, 기호 등) 가 있는 글자떼를 간추리면 간추리기에 문제가 있다. ISO 14651 틀을 사용한 보기를 들면 다음과 같다.

```
level1
기A : <U1100><U1175><U0041>
기草 : <U1100><U1175><U8349>
길   : <U1100><U1175><U11AB>
```

그림 5. 잘못된 간추리기 보기

그림 5에서 글자떼들의 level 1에 있는 각 글자들을 비교하면, 각 글자들의 'ㄱ', 'ㅣ'에 대한 부호값은 같지만, 다음 글자의 부호값에서 차이가 나기 때문에 간추리는 차례가 결정된다. 그런데, 끝소리가 있는 한글 글자마디의 끝소리 글자 ('길'의 'ㄹ') 와 다른 나라 글자 ('기草'의 '草') 가 비교 되어 남한의 한글 가나다 차례와는 다른 '기A', '길', '기草'와 같은 잘못된 간추리는 차례가 된다.

한글은 첫소리, 가운데 소리, 끝소리 글자로 이루어져 있으며, 한글 글자마디는 끝소리가 있거나 없을 수도 있다. 이에 비해 영문자, 한자, 일본어, 기호 등과 같은 글자들은 일반적으로 하나의 부호값으로 이루어져 있다. 그러므로, 첫가끝 조합형 한글과 다른 나라 글자들을 간추리면, 끝소리가 없는 글자마디 (보기: '가') 와 다른나라 글자 (한자, 영문자, 기호 등) 가 비교되는 경우가 있다. 보기를 들어, 그림 5에서처럼 '길'의 'ㄹ'과 '기A'의 'A'가 비교되는 것이다. 이런 상황에서, 첫가끝 조합형 한글 다음에 첫가끝 조합형 한글 부호값(U+1100 ~ 11FF)보다 큰 다른 나라의 글자가 오면 그림 5의 간추리기 결과와 같이 남한의 한글 가나다 차례와는 다른 잘못된 간추리기가 된다.

4.2 해결 방안 제시

문제점을 해결하는 방법은 첫소리, 가운데 소리, 끝소리 글자로 이루어진 한글의 특성을 고려해서 모든 한글 글자마디를 첫소리, 가운데 소리, 끝소리 글자로 나누어 비교하는 것이다. 끝소리가 없는 한글 글자마디들도 끝소리 채움을 포함해서 나눈다.

현재 ISO 14651에 있는 공통 틀 표는 간추리기 차례를 비교하는 단순한 기능을 가지고 있다. 그러므로, 한글 글자마디의 끝소리 채움과 같은 복잡한 것을 처리하지 못하기 때문에 한글 글자마디를 첫소리, 가운데 소리, 끝소리 글자로 나눌 때, 반드시 전처리를 이용한 뒤 ISO 14651을 적용해서 간추리는 차례를 결정해야 한다.

전처리를 이용해서 한글 글자마디를 나눌 때 각 글자에 대한 부호값이 주어진다. 그런데, 현재 한글 첫가끝 조합형 부호값에는 끝소리 글자 채움에 대한 부호값이 지정되어 있지 않고, ISO 14651에서도 해결 방안에 대해 제시하고 있지 않으므로 어떤 부호값을 지정할 것인지 고려해야 한다.

본 논문에서는 이러한 부호값으로 가상의 부호값을 지정하도록 한다. 가상의 부호값으로 첫가끝 부호값 U+11xx 영역에서 비어있는 16 개 부호값 (U+115A ~ 115E, U+11A3 ~ 11A7, U+11FA ~ U11FF) 중에서 1개를 활용할 것을 제안한다. 그런데 16개의 부호값 중에서 끝소리 글자 부호값(U+11A8 ~ 11F9)보다 큰 부호 값(U+11FA ~ U11FF)은 사용하지 못한다. 왜냐하면 끝소리 채움 부호값이 끝소리 부호값보다 크면 한글 가나다 차례와 맞지 않게 되므로 잘못된 간추리기가 된다. 보기를 들면 다음과 같다.

```
level1
가 <U1100><U1161><U11FA>
각 <U1100><U1161><U11A8>
```

그림 6. 잘못된 끝소리 부호값 지정

그림 6은 끝소리 글자 채움 부호값으로 'U11FA'를 지정한 경우이며, '각'의 끝소리 부호값(U11A8)보다 크기 때문에 남한의 한글 가나다 차례와 맞지 않게 된다. 그러므로 본 논문에서는 끝소리 채움 부호값으로 'U11A1'을 지정한다.

제안한 방법으로 ISO 14651 틀을 사용해서 첫가끝 조합형 한글 또는 완성형 한글과 여러나라 글자가 섞여 있는 글자떼를 간추리는 보기를 나타내면 다음과 같다.

```
level1
기A  <U1100><U1175><U11A1><U0041>
기草 <U1100><U1175><U11A1><U8349>
길    <U1100><U1175><U11AB>
```

그림 7. 제안한 방법으로 간추리기

그림 7에서 '기A'와 '기草'의 '기'가 'ㄱ(<U1100>)', 'ㅣ(<U1175>)', '끝소리 채움' 과 같이 첫소리 글자, 가운데 소리 글자, 끝소리 글자로 나누어지고, '길'도 'ㄱ(<U1100>)','ㅣ(<U1175>)','ㄹ(<U11AB>)'과 같이 3개의 글자로 나누어지므로 끝소리 글자가 없는 한글의 글자마디와 다른나라 글자가 비교되는 경우가 없어진다.

그러므로, 본 논문에서 제안한 방법으로 첫가끝 조합형 한글과 여러나라 글자들이 섞여 있는 글자떼를 간추리면, 한글 글자 다음에 한글의 부호값보다 큰 다른 나라의 글자가 오더라도 간추리기가 제대로 된다. 게다가 현재 ISO 14651의 공통 틀 표에서 제대

로 간추리지 못하는 불완전한 한글 글자를 처리할 수 있을 것으로 보인다.

5. 결 론

ISO 14651은 여러 나라 글자계 (script) 가 섞여 있을 때, 모든 글자의 차례를 정하고 간추리는 틀에 관한 표준이다. ISO 14651에서 끝소리가 없는 첫가끝 조합형 한글과 여러나라 글자가 섞여 있는 글자떼를 간추리면 한글의 가나다 차례와 다르게 간추리기가 되는 경우가 있다.

본 논문에서는 이러한 문제점을 지적하고 해결방안으로 전처리 (preprocessing) 를 이용해서 한글의 글자마디를 첫소리, 가운데 소리, 끝소리 글자로 표현할 것을 제안하였다. 그리고, 제안한 방법으로 간추리기에 문제가 일어나지 않는다는 것을 보였다.

향후 연구 과제로는 가상의 부호값으로 첫가끝 부호값 U+11xx 영역에서 비어있는 영역에 대한 부호값을 끝소리 채움 부호값으로 사용했지만, 이것은 임시방편에 지나지 않는다. 빈 자리가 없는 경우 ISO 14651 틀 안에서 해결 할 수 있는 방안이 현재 없는 것으로 보이는데 앞으로 이에 대한 연구가 필요하다. 그리고 옛 한글 또는 새로운 겹글자와 여러나라 글자가 섞여 있는 글자떼의 간추리기에 대한 문제점 및 해결방안에 대해 연구도 필요하다.

[참고 문헌]

[1] ISO/IEC, 'ISO/IEC 14651 - International string ordering and comparison', ISO/IEC JTC1/SC 22WG, 2000.

[2] The Unicode Consortium, " The Unicode Standard, version 3.0", Addison-Wesley, 2000.

[3] 한국표준협회, "국제 표준에 따른 남북 가나다 차례 지원 방안 연구", 한국표준협회, 2001.

[4] 김종휘, 김경석, 'ISO 14651에 의한 한글 ordering의 문제점과 그 해결 방안', 한국 정보과학회 2001 가을 학술 발표논문집(II), 제28권, 제2호, pp.187-189, 2001.

[5] 기술표준원, 국제문자부호계 KS규격의 국제규격 부합화 연구, 기술표준원, 2000.

[6] Kent Karlsson, 'Ordering rules for Hangul', Working Group Document, 2001.

[7] KIM Kyongsok 'New Canonical decomposition and composition processes for Hangeul' , Computer Standards & Interfaces Vol.24 pp69-82 2002

[8] Unicode Standard Annex #15, 'Unicode Normalization Form', Http://www.unicode.org/unicode/reports/ tr15

[9] 김경석, '한글 전산화의 관점에서 살펴본 거꾸로 간추린 한글 사전의 가나다 차례 원칙' , 한글 학회 , 한글 제 256권 pp191-241, 2002