한자표기 용어로부터 바른 한글음의 자동생성

An Automatic Generation of Correct Hangul from Hanja Term

최석두, 이지영
이화여자대학교 문헌정보학과

Suk Doo Choi, Ji Young Lee
Dept. of Lib. & Inf. Sci., Ewha Womans Univ.

漢字로 기술된 용어를 한글로 자동변환하게 되면 생성된 한글이 한글의 음운구조에 맞지 않는 일이 생기게 된다. 이런 문제가 생기는 원인은 해당 한자에 대응하는 정확한 한글을 입력하여 변환하지 않았을 때와 해당 한자의 음이 없거나 한자와의 음운구조가 달라 생기는 때로 대별할 수 있다. 본고에서는 이를 해결하기 위하여 KS C 5601 표준코드(이하 표준코드라 한다)를 기준으로 漢字의 多音字를 조사하고 이 多音字가 포함되어 있는 용어를 이용하여 매핑파일을 구축함으로써 바른 한글음을 자동생성할 수 있는 방안을 논한다.

1 서론

많은 문헌의 서명이나 全文데이터가 漢字를 포함하고 있다. 이 데이터를 이용하여 키워드를 자동생성하게 되면 생성된 한글키워드가 바른 표기로 변환되지 않는 경우가 생기게 된다. 이것은 대부분 입력시 다른 音의 한글을 입력하여 漢字로 변환한 결과이다. 다른 한자코드가 입력되는 원인을 보면 크게 네 가지로 나눌 수 있다.

첫째, 일반적으로 입력자는 한자 폰트의 획득이 목적일 뿐, 역으로 생성될 한글음에는 관심이 없다.
둘째, 정확한 한글표기를 몰라서 다른 한글음을 이용하여 변환한다.
세째, 해당 한자의 음이 없어서 다른 음의 한자를 빌려서 입력한다.
네째, 한자와 한글의 음운구조가 달라 생성된 한글음이 맞지 않는다.

특히 이미 한자로 된 원고를 입력하는 경우가 직접 입력하는 원고에 비해 다른 한글음으로 입력하여 변환할 확률이 높아진다. 全文텍스트, 그리고 古書까지도 컴퓨터시스템에 의존하는 경우가 늘어나고 있으며, 문서작성자의 교육이나 홍보만으로 해결되지 않는 면이 있기 때문에 실용시스템에서 대단히 심각한 문제로 대두되고 있다.

본고에서는 이를 해결하기 위하여 표준코드를 기준으로 漢字의 多音字를 조사하고 이 多音字가 포함되어 있는 용어를 이용하여 매핑파일을 구축함으로써 다른 한글음을 이용하여 변환된 한자용어에 대하여 바른 한글음과 한자를 생성할 수 있는 방안을 논한다.

2 틀린 한글음의 생성원인

한자에서 한글음을 생성할 때 문제가 되는 경우는 다음 몇 가지로 나눌 수 있다. 다음의 각 예에서 기호←의 왼쪽은 바른 음이며 오른쪽은 변형이다.

1) 다른 음을 사용한다 : 해당음이 있지만 한자 폰트만 얻으면 되므로 다른 음을 사용한 경우이며, 가장 많다. 이 때는 후술하는 多音字파일을 이용하여 해당 용어의 변형을 자동생성할 수 있다.

예) 樂山(요산) ← 樂山(악산, 낙산, 락산)
殺到(쇄도) ← 殺到(살도)

2) 해당 음이 없어서 다른 음을 빌려쓴다 : 한자폰트는 있으나 해당음이 없어서 다른 음을 빌려온 경우이다. 입력자에게는 이 방법밖에 없다.

예) 五六月(오뉴월) ← 五六月(오륙월, 오육월)

木瓜(모과) ← 木瓜(목과)
盟誓(맹세) ← 盟誓(맹서)
白魚(뱅어) ← 白魚(백어)
分錢(푼전) ← 分錢(분전)
林巨正(임꺽정) ← 林巨正(임거정)

3) 사이시옷을 표현할 수 없다 : 한자는 옳게 입력되어 있으나 생성되는 한글이 달라지는 경우이다. 즉, 한자에는 사이시옷을 표기할 수 없기 때문이다. 그러나 해당 문자가 모두 한자로 표기할 수 있는 용어는 다음 예1의 여섯 가지밖에 없으므로 이 용어만 포함시키면 된다(미승우, 1993). 다만 한글과 한자가 섞여서 하나의 용어로 사용되는 경우가 있다. 이 중 한글이 선행하는 경우에는 문제가 되지 않는다. 한글이 한자의 음에 영향을 주지 않기 때문이다. 예2의 "귓병(귓病)"과 같은 경우이다. 그러나 예2의 "탯줄(胎줄-태줄)"과 같이 한자가 앞에 오는 경우에는 한자의 음에 영향을 주는 것이 있기 때문에 이를 처리하여야 한다.

예1) 庫間(곳간) ← 庫間(고간)
貰房(셋방) ← 貰房(세방)
數字(숫자) ← 數字(수자)
車間(찻간) ← 車間(차간)
退間(툇간) ← 退間(퇴간)
回數(횟수) ← 回數(회수)
예2) 귓病(귓병)
胎줄(탯줄) ← 胎줄(태줄)

4) 梵語 등의 영향으로 다른 음을 사용한다 : 특히 불교용어에 빈번하다. 대부분 해당 음이 없어서 다른 음을 빌려오게 된다.

예) 初八日(초파일) ← 初八日(초팔일)
婆羅(바라) ← 婆羅(파라)
陀羅尼(다라니) ← 陀羅尼(타라니)

5) 습관상 사용한다 : 원 글자의 뜻에는 영향을 미치지 않으나 습관상 특별하게 음을 붙인 경우이다. 대부분 해당 음의 漢子가 없어서 다른 음을 빌려오게 된다.

예) 內人(나인) ← 內人(내인)
司僕寺(사복시) ← 司僕寺(사복사)
奉常寺(봉상시) ← 奉常寺(봉상사)
宮商角徵羽(궁상각치우) ← 宮商角徵羽(궁상각징우)

6) 1988년 한글맞춤법 개정안과 다르다 : 1988년 1월 19일 문교부가 새롭게 개정고시하여 1989년 3월 1일부터 시행하도록 한 "標準語規定" 가운데 "標準語査定原則"(제11, 13항)에 규정되어 있는 내용과 상이한 경우이다. 다음 예와 같이 漢子 "着"은 모음의 발음변화를 인정하여 "책"으로, "句"는 단어의 일부가 될 때 "구"로 통일한다. 다만 "귀글", "글귀"에서는 "귀"를 쓴다.

예) 주책(主着) ← 주착(主着) "11항"
구절(句節) ← 귀절(句節) "13항"
문구(文句) ← 문귀(文句)
글귀(글句) ← 글구(글句) "귀를 쓴다"

3 매핑파일의 구축

3.1 표준코드의 분석

표준코드내에서 복수의 음을 갖는 漢字를 분석할 필요가 있다. 이춘택(1991) 교수가 조사한 결과를 보면 표준코드에 있는 한자로서 두 가지 이상의 음을 갖는 문자의 수는 다음과 같다. 이 수치는 동아출판사 발행의 "漢韓大辭典"(李家源, 權五恂, 任昌淳 監修)을 기본으로 "大字源"(張三植 著)을 참고하였으며, 그 외 일부를 추가한 것이다.

2音字 805자
3音字 139자
4音字 28자
5音字 7자
6音字 1자
7音字 1자
계 981자

본 연구에서는 상기 조사결과를 이용하였으나 多音字는 사용하는 한자사전에 따라 다소 달라지는 경향이 있다. 따라서 차후 多音字파일을 만들 때 어느 사전을 중심으로 할 것인가는 대상 데이터에 따라 달라질 것이다.

또한 표준코드에 있는 漢子로서 두 가지 이상의 코드를 갖는 문자의 수(즉 두 가지 이상의 음을 표준문자내에서 갖는 문자; 重出字)는 다음과 같다(이춘택, 1991). 따라서 표준코드의 한자는 4,888자 중에서 268자를 제외한 4,620자를 수록한 셈이 된다. 그외에 正字와 略字가 동시에 사용된 문자로서 암(岩, 巖)과 만(万, 萬)의 두 종류가 있으나 바른 한글음의 생성에는 영향을 주지는 않는다.

2회 257종 514자 → 257자
3회 4종 12자 → 8자
4회 1종 4자 → 3자
계 262종 530자 → 268자

3.2 多音字 파일의 작성

파악된 표준코드내의 모든 多音漢子에 대하여 다음 예와 같이 각 音(표준코드에 없는 音도 포함한다)을 조사하였다. 따라서 多音字파일에 수록된 문자수는 두 가지 이상의 음을 갖는 문자 1,867자와 重出字 530자를 포함하여 도합 2,543자가 된다.

多音字파일에서 첫 번째 숫자(356 및 10)는 글자의 그

룹을 나타내는 일련번호이며 같은 숫자는 같은 문자군을 나타낸다. 이 숫자정보는 용어변형의 자동생성시 사용한다. 두 번째 숫자는 1이면 표준코드에 없는 음이라는 것을 나타낸다. 각괄호는 해당하는 한자의 코드(16진코드)이다.

예1) 樂 낙 356 (E3E2)
樂 락 356 (E5A5)
樂 악 356 (ED55)
樂 요 356 (EF9B)
예2) 假 가 10 (CAA3)
假 하 10 1

3.3 기본 매핑파일의 자동생성

어휘에서 多音字에 의해 생길 수 있는 변형을 조사하고 조사결과를 이용하여 변형 중 처리대상에서 제외시킬 수 있는 상식적인 규칙을 찾아낼 수 있을 것이다. 예를 들면, 다음과 같이 "살인, 여성, 희노애락, 김천" 등의 용어를 "쇄인, 녀성, 희노애요, 금천" 등으로 읽어 漢字로 변환할 확률은 매우 낮으며, 특히 "金"이라는 문자는 고유명사이면 거의 "김"으로 읽기 때문이다.

예) 살인(殺人) 쇄인(殺人)
여성(女性) 녀성(女性)
희노애락(喜怒哀樂) 희노애요(喜怒哀樂)
김천(金泉) 금천(金泉)

그러나 상기 방법으로 각 변형에 해당하는 문자열을 하나 하나 생성한다는 것은 노동집약적인 일이며, 인간의 많은 지적 노력을 필요로 한다. 또한 고유명사와 보통명사 양쪽으로 사용되어 구별이 안되는 용어도 있으며 예외를 처리할 수 없는 단점이 있다. 따라서 매핑파일이 방대하게 늘어나고 사용될 확률이 매우 낮은 문자열이 포함되더라도 자동으로 생성하는 것이 보다 효율적일 것이다.

자동생성의 방법은 다음과 같다. 多音字가 포함되어 있는 "여성 女性"이란 용어의 예를 들어 보자. 多音字파일에

女 여 503 (EDFC)
女 녀 503 (E443)

가 있을 때, "여성 女性"이라는 용어가 입력되면 漢字부분만 한 문자씩을 잘라 多音字파일을 탐색한다. 즉 "女"(여)와 "性"(성)을 차례로 탐색하게 된다. 우선 "女"(여)가 多音字파일의 "女 여 503"과 매칭되면 "여성 女性"을 같은 그룹에 있는 모든 문자로 각각 대체시킨다. 따라서 "여성 女性"과 "녀성 女性" 두 용어가 생성된다. "性"(성)은 매칭되는 문자가 없으므로 생성된 용어의 변형은 두 가지가 된다. "樂山樂水"(요산요수)의 예를 들면, "樂"은 3.2 예1)과 같이 "낙, 락, 악, 요"의 4가지 음을 가지므로 첫 번째의 "樂"(요)에 의해 다음의 4가지 용어를 생성한다.

낙산요수 樂山樂水 락산요수 樂山樂水
악산요수 樂山樂水 요산요수 樂山樂水

두 번째 이후의 문자부터는 그 그룹에서 만들어진 용어 전체를 대상으로 대체한다. 3번째 "樂"(요)를 앞에서 만든 네 가지 용어에 대하여 적용시키게 되며 결과는 다음과 같이 16 가지가 된다. 매핑파일을 만들 때 바른 용어는 제외시킬 수도 있고 포함시킬 수도 있다.

낙산낙수 樂山樂水 락산낙수 樂山樂水
낙산락수 樂山樂水 락산락수 樂山樂水
낙산악수 樂山樂水 락산악수 樂山樂水
낙산요수 樂山樂水 락산요수 樂山樂水

악산낙수 樂山樂水 요산낙수 樂山樂水
악산락수 樂山樂水 요산락수 樂山樂水
악산악수 樂山樂水 요산악수 樂山樂水
악산요수 樂山樂水 요산요수 樂山樂水

하나의 용어내에 세 가지의 多音字가 있는 경우에도 동일한 방법으로 처리한다. 즉, 두 번째 처리완료 데이터가 다음 세 번째 처리의 대상정보가 된다. 표준코드에 없는 음인 경우에도 多音字파일에 漢字코드가 있으면 매칭된 음을 취하고 같은 문자군에 있는 다른 음과 漢字로도 대체시킨다.

다만 표준코드에 없는 음이 있을 때 어느 음에 해당하는 한자를 사용했는지 알 수 없는 경우가 생기게 된다. 다음과 같은 경우를 보자.

什 십 459 (ED37)
什 집 459 (F49C)
什 습 459 1

상기 예에서 "什"(습)은 표준코드에 없는 음이므로 "什"(십)의 음에 해당하는 한자를 따랐는지 "什"(집)의 음에 해당하는 한자를 따랐는지 알 수가 없다. 이를 때는 두 번째 숫자 1(표준코드에 없는 음)을 이용하여 해당 문자군의 모든 한자를 대체하면 된다. 그러나 이와 같은 일은 하나의 漢字가 세 가지 이상의 음을 갖고 있으면서 적어도 두 가지는 표준코드에 있는 경우에만 발생하며 일반 명사에서는 거의 일어나지 않는다. 다만 고서관련 자료인 경우에는 가능하므로 두 번째 숫자(1: 표준코드에 없음)를 준비하고 있다.

이와 같은 과정으로 생성된 모든 변형에 대하여 바른 한자·한글쌍을 대응시켜 매핑파일을 만든다. 전자국어사전을 이용하여 모든 기입어의 한자를 모두 多音字파일과 비교함으로써 多音字가 포함되어 있는 용어를 찾아낼 수 있다. 본 연구에서는 공개된 전자국어사전이 없어서 이화여자대학교 중앙도서관이 사용하고 있는 한자음절변환사전(항목수 약 5만어)을 중심으로 하고 大漢韓辭典(1985)

의 용례로 보완하여 실험하였다.

만들어진 한자·한글쌍이 각각을 구별할 수 있는 하나의 유일한 코드가 된다. 매핑파일의 예를 들면 다음 표 1과 같다. 생성된 변형이 →의 왼쪽이 되며 입력된 용어가 오른쪽이 된다. 또한 변형표기와 대체표기의 "(한자)"부분의 영문자는 그 위치에 있는 "바른 표기"의 한자임을, 그 중 첨자가 붙은 영문자는 코드가 다른 한자임을 의미한다.

표 1 한자·한글쌍의 매핑파일

바른표기(한자)	변형표기(한자)		대체표기(한자)
세도 (歲到)	살도 (A_1B)	→	세도 (AB)
시월 (十月)	심월 (AB)	→	시월 (AB)
오뉴월(五六月)	오육월(AB_1C)	→	오뉴월 (ABC)
	오륙월(ABC)	→	오뉴월 (ABC)
요산 (蓼山)	락산 (A_1B)	→	요산 (AB)
	낙산 (A_2B)	→	요산 (AB)
	악산 (A_3B)	→	요산 (AB)
곳간 (庫間)	고간 (AB)	→	곳간 (AB)
나인 (內人)	내인 (AB)	→	나인 (AB)
초파일(初八日)	초팔일(ABC)	→	초파일 (ABC)

3.4 기본 매핑파일의 보완

전자국어사전을 이용하여 기본매핑파일을 만들어도 모든 용어가 망라될 수는 없다. 추가로 여러 가지 전문용어사전을 이용하여 용어를 수집할 필요가 있다. 이 때 가장 효용성이 있는 것은 전문용어사전과 시소러스가 될 것이다. 전문용어사전이나 시소러스에 바른 한자코드가 입력되어 있을 필요는 없다. 최소한 바른 한글음이 한자와 함께 기술되어 있으면 된다.

그러나 漢字의 다른 음을 사용한 경우를 제외하고 상기 규칙을 그대로 적용해버리면 실제 상황에서는 사용하지 않는 용어가 다수 만들어지는 일이 생기게 된다. 또한 일정한 규칙에 의해 변형이 생기는 것이 아니라 경우에 따라 달라지는 음들이 있어서 자동생성시 에러가 발생할 우려가 있다. 이와 같이 용어의 수를 줄이고 불규칙적인 용어를 정확하게 처리하기 위하여 다른 음을 사용하는 경우를 제외하고는 매뉴얼로 입력하여 보완하여야 한다. 예를 들면 "內, 寺"의 音은 "내, 사" 밖에 없으므로 특별한 용례에서의 音인 "나, 치"는 多音字파일에 등록되지 않는다. 이와 같은 예는 기본매핑파일에 별도로 입력한다.

이 때 대체표기에서 해당 음의 한자가 없고, 빌려와야 할 문자후보가 두 종류 이상일 때는 어느 것을 사용해도 좋다. 예를 들면, 표 1에서 "오륙월, 오육월"의 대체표기인 "오뉴월(五六月)"에서 "六"은 "륙"의 음을 따르고 있으나 "육"의 음을 따라도 좋다.

또한 한 문자로 하나의 단어가 되는 경우도 있으므로 상기 多音字파일의 모든 문자를 매핑파일에 포함시키는 것이 바람직할 것이다.

4 바른 한글음의 생성

교정처리의 방법은 우선 색인대상이 되는 한자용어의 한글·한자쌍을 생성하게 된다(표 1의 변형표기). 처리대상 용어는 시스템에 따라 수동, 자동 어느 쪽으로도 선정할 수 있다. 다만 생성된 한글·한글쌍은 그 용어를 식별할 수 있는 유일한 코드라고 가정하며 다음 경우 중의 하나가 된다.

첫째, 맞는 경우
둘째, 한자는 맞으나 한글이 틀린 경우
세째, 한글, 한자가 다 틀린 경우

둘째, 세째의 경우에는 매핑파일과 비교하여 매칭이 되면 대응되는 바른 표기의 코드쌍(표 1의 대체표기)으로 간단히 대체하게 된다. 첫째는 코드쌍이 본래 바르게 표기된 경우이며 매핑파일에 없으므로 그대로 사용한다.

그러나 모든 경우를 다 해결할 수 있는 것은 아니다. 다음과 같이 의미에 따라 다르게 사용하는 경우에는 문제가 있다. 즉, "금슬(琴瑟)"이란 용어는, 악기를 나타낼 때는 "금슬(琴瑟)"이 맞으며, 부부의 사랑을 나타낼 때는 "금실(琴瑟)"이 맞다. 어느 쪽 한글을 생성할 것인지를 결정해야 한다.

예1) 금슬(琴瑟) ← 금슬(琴瑟) "악기"
금실(琴瑟) ← 금슬(琴瑟) "부부의 사랑"
예2) 차간(車間) ← 차간(車間) "차간거리"
찻간(車間) ← 차간(車間) "타는 곳"

이 부분에 대해서는 색인자에게 필요정보를 제공하여 선택하게 하거나 자연언어처리시스템에서 문맥을 파악하여 유추하는 길밖에 없을 것이다.

5 결론

漢字표기용어로부터 바른 한글음의 자동생성방법에 대하여 논하였다. 바른 한글음의 자동생성을 위해서는 多音字파일의 구축과 변형이 가능한 용어를 조사하는 일이 중요하며 어려운 일이다. 조사된 용어는 근본적으로 망라적일 수 없으며, 실제 시스템에서 사용하면서 보완되어야 할 것이다. 또한 한자가 확장되거나 처리대상의 분야가 확대되면 부수적으로 多音字과 변형이 가능한 용어의 조사가 실시되어야 할 것이다.

참고문헌

大漢韓辭典(1985). 張三植 著. 서울: 三榮出版社.
미승우(1993). 새맞춤법과 교정의 실제. 증보판. 서울: 어문각.
이춘태(1991). 韓日 國家規格漢字코드의 統合硏究. 중앙대학교 대학원 도서관학과 자료조직전공 박사학위논문, 미간행