(제 10회 한글 및 한국어 정보처리 학술대회)

# 한·영 혼용 문에서 조사오류 검출 및 교정

정 규 철 · 정 민 수 · 조 원 홍

군산대학교 컴퓨터과학과
전북 군산시 미룡동 산 68번지
우: 573-701
{kcjung, kihong, wonhong}@cs.kunsan.ac.kr

# A Josa-Errors Detection and Correction from Korea-English Mixed Sentences

Kyu-Chol Jung, Min-Su Jung, Wonhong Cho
Department of Computer Science, Kunsan National University

**요약**

전문 분야의 세분화로 인한 신조어 발생이 늘어나고 있다. 또한, 이러한 단어를 우리말로 표현이 불가능한 경우 우리 발음으로 풀어 기록하지 않고 그대로 적는 경우도 늘어나고 있는 추세이다. 특히 전문 서적일수록 두드러진다. 그러나 한글과 영어를 혼용하여 기록할 경우 부적절한 조사의 쓰임으로 인하여 매끄럽지 못함을 가끔 볼 수 있다. 본 논문에서는 영단어의 발음특성정보를 이용하여 한글 조사의 오류를 정확하게 검출하고 교정을 할 수 있다.

## 1. 서론

최근에 들어 전문 분야의 다양화를 이루면서 외국 주요기술의 이식이 늘어나고 있다. 이에 따라 외국 문서의 번역 또한 크게 늘어나는 추세이고 신종외국어의 번역이 불가능해지게 되어 한글과 영어를 혼용하여 표기한다. 그런데, 영어의 발음적 특징을 살펴보면 받침을 나타내는 것은 " l, m, n"(받침문자)이며 받침을 나타낼 수도 있고 나타내지 않을 수 도 있는 것은 "b, c, e, g, k, t"(혼용문자) 가 있고 나머지는 받침을 나타내지 않는 것들(받침 없는 문자)임을 알 수 있다. 또한, 한글 조사는 앞 글자가 받침발음을 하는 문자일 경우에는 "은, 이, 을,…"처럼 초성이 무성음으로 시작하고 받침이 없는 경우에는 "는, 가, 를,…" 처럼 유성음으로 나타낸다.
이러한 한글과 영어의 발음적인 특성을 이용하여 조사 오류를 검출 하고자 한다.
이는 한글과 영어가 혼용될 경우 영어는 체언으로 쓰이는 경향이 있기 때문에 조사와 결합하는 경우가 많다. 그러나 현존하는 맞춤법 검사 방식으로는 이에 대한 조사 결합 오류를 정확하게 검출이 불가능하므로 본 논문에서는 영어의 발음적인 특성과 한글 조사의 특징을 이용하여 오류를 정확하게 검출하고자 한다.

본 논문의 구성은, 2장에서 받침을 나타내는 영문자, 받침을 나타내지 않는 영문자와 받침을 나타낼 수도 있고 받침을 나타내지 않을 수 도 있는 영문자에 대해 알아본다. 3장에서는 체언과 결합하는 조사의 종류와 받침의 유무에 따라 변하는 조사를 알아본다. 4장에서는 조사가 정확히 사용되었는지를 검사하는 알고리즘을 기술하고 오류검출에 필요한 몇 가지 규칙에 대해 알아본다. 5장에서는 본 논문에서 제안한 시스템의 우수성을 가리기 위한 실험 및 기존 방법과의 비교 평가를 한다. 마지막 6장에서는 본 논문에서 제시하는 시스템의 문제점과 향후 개발 사항을 기술한다.

## 2 받침 문자의 검색

### 2.1 받침만 나타내는 문자

표 1 받침문자의 사용 예

| 문자 | 사용 예 |
|---|---|
| l | school(스쿨), mail(메일), well(웰), handball(핸드볼), … |
| m | Tom(톰), humanism(휴머니즘), gram(그램), cream(크림), … |
| n | man(맨), carton(카턴), chain(체인), heroin(헤로인),… |

받침만 나타내는 경우의 영문자는 "l, m, n"으로 확연히 드러남을 볼 수 있다. 그 사용 예는 표 1과 같으며 꼭 받침을 나타냄으로 받침문자로 처

리하여야 한다.

### 2.2 혼용 발음 문자

영어 문자를 발음함에 있어서 받침 문자와 무 받침 문자로 만 구분되어 지는 것이 아니라 경우에 따라 변하는 문자들이 있다.
이에 해당하는 문자는 "b, c, e, g, k, t"이며 사용 예는 다음 표 2와 같다.
표 2에서 보듯이 혼용 문자에 해당하는 문자들은 경우에 따라 다른 결과를 볼 수 있다. 하지만 위의 경우를 상세히 보면 몇 가지의 법칙을 유추할 수 있다.

표 2 혼용문자의 사용 예

| 문자 | 사용 예 |
|---|---|
| b | Arab(아랍), club(클럽), bulb(벌브) |
| c | graphic(그래픽), Arc(아크) |
| e | image(이미지), cycle(사이클), dance(댄스), guide(가이드), megaphone(메가폰) |
| g | jitterbug(지터버그), diving(다이빙) |
| k | kick(킥), disk(디스크), network(네트워크), crank(크랭크) |
| p | stamp(스탬프), up(업) |
| t | accent(악센트), bat(배트), carat(캐럿), capet(카페트) |

① "b","c"의 경우는 앞문자가 모음이면 받침을 나타낸다.

② "p"의 경우 특이하게도 앞문자가 받침 문자인 경우만 받침발음을 갖지 않는다.

③ "e"의 경우는 독자적인 발음을 나타내지 못하고 앞문자가 받침을 나타내는 문자냐 아니냐에 따라 좌우된다.

④ "k"의 경우를 보면 앞 문자가 "c"인 경우에만 받침을 갖고 이를 제외하고는 화자에 따라 받침을 나타낼 수 도 있고 나타내지 않을 수 도 있다.

⑤ "t"의 경우는 앞 글자에 관계없이 받침 발음과 무받침 발음 모두를 허용함을 볼 수 있다.

⑥ "g"의 경우에는 앞문자가 "n"인 경우를 제외하고 거의 무받침 발음을 갖는다. 즉 "ng"를 이룰때만 받침을 갖는다.

### 2.3 받침을 나타내지 않는 문자

위의 두 가지 경우를 제외한 철자들(a,d,f,h,i,j,o,q,r,s,u,v,w,x,y,z)은 받침을 갖지 않는다.
예) acacia, academy, canvas, computer, box, card, dash, giro, …

## 3. 조사

조사는 스스로 자립할 수 없고 자립성을 갖은 다른 말(주로 체언)에 부쳐서 다른 말과의 관계를 나타내거나, 뜻을 더해 주는 단어이다. 조사는 체언 뿐 아니라 체언 상당어구와 어울려 사용된다.[1,2,3]
조사는 크게 격조사와 보조사로 나누어진다.

### 3.1 조사의 종류

격조사는 다시 7가지로 나눠지고, 보조사는 2가지로 나눈다.[4]
<격조사>
① 주격조사 : -이, -가, -께서, -에서, -서

② 목적격조사 : -이, -가

③ 관형격조사 : -의

④ 호격조사 : -야, -아, -여, -시여, -이시여, -이야, -이여

⑤ 부사격조사 : -에게, -같이, -한테, -으로, -로, -에서, -처럼, -대로, -에, -한테서, -에게로, -으로서, -으로써, -으로서, -로서, -로써, -더러, -보고, -처럼, -보다. -만큼, -라고, -이라

⑥ 공동격조사 : -과, -와, -하고

⑦ 인용격 조사 : -라고, -고

<보조사>
⑧ 통용보조사 : -곧, -까지, -ㄴ랑, -나, -는, -는커녕, -도, -든지, -라도, -랑은, -랑, -마다, -마저, -만, -부터, -서껀, -야, -야말로, -온, -은랑, -은커, -조차

⑨ 종결보조사 : -마는, -그려, -요 이다.

이들 중 받침문자인지 아닌지에 따라 변하는 조사는 표 3과 같다.
이 조사들은 앞단어의 끝 글자에 의해 서로 짝을 이루어 변하게 된다.

표 3 체언에 따라 변하는 조사

| 받침이 있을 경우 | 받침이 없을 경우 |
|---|---|
| 이 | 가 |
| 을 | 를 |
| 아 | 야 |
| 이시여 | 시여 |
| 으로 | 로 |
| 으로써 | 로써 |
| 으로부터 | 로부터 |
| 이라고 | 라고 |
| 이랑 | 랑 |
| 이며 | 며 |
| 와 | 과 |
| 이든지 | 든지 |
| 이라고 | 라고 |

## 4. 오류검출 알고리즘

### 4.1. 오류검출을 위한 적용규칙

그림1에서와 같이 오류를 판단할 수 있으며 적용규칙은 다음과 같다.

① 영문 스트링이 특수한 경우인지 확장사전을 검색하고 사전에 없을 경우에 아래의 규칙을 적용한다.

예) 10kg, 5cm, 10db,…

② 받침문자는 "l,m,n"이다.

예) beam, bell, champion,…

③ 혼용문자는 "b, c, e, g, k, p, t"이다.

④ ②와 ③ 이외의 문자는 받침 없는 문자이다.

⑤ "t"를 제외한 혼용문자는 앞 문자에 의해 받침 여부가 결정되며, "t"는 받침처리에 관계없이 오류가 없는 것으로 간주한다.

예) contact(콘택트=콘택), credit(크레디트=크래딧)

⑥"b","c"는 앞문자가 모음이면 받침문자로 처리하고 "b"앞에 "m"이 오면 "b"는 묵음이 되므로 받침문자로 처리한다.

예)bomb(봄), Arab, graphic

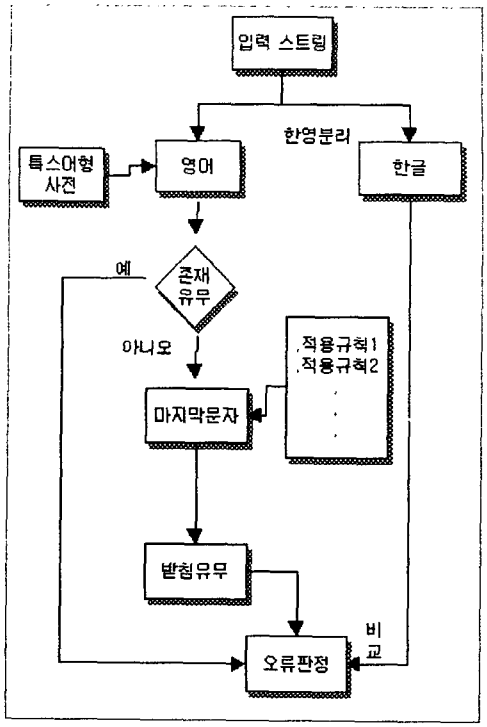
⑦ "g"는 앞문자가 "n"이면 받침문자로 처리한다.

예)casing(케이싱), catalog(카달로그)

⑧ "e"는 앞문자가 받침문자이면 받침문자로 처리하고, 앞문자가 "k, t"이면 모두 맞는 것으로 인식하며, 나머지는 받침 없는 문자로 처리한다.

예)case(케이스), rice(라이스),
chocolate(초콜릿=초코리트),
concrete(콘크리트=콘크릿),
cake(케익=케이크)

그림 1 오류검출 알고리즘




⑨ "k"는 앞문자가 "c"인 경우에만 받침문자로 처리하고, "s"인 경우엔 받침 없는 문자로 처리하고, 나머지는 모두 허용한다.

예)kick(킥), back(백), truck(트럭), disk(디스크), work(워크=웍)

⑩ "p"는 앞 문자에 받침문자(l,m,n)가 올 경우일 때만 받침 없는 문자로 처리한다.

예) cap(캡), stamp(스탬프)

⑪ 모든 문자가 대문자일 경우에는 규칙 1)만 적용한다. 왜냐하면, 대문자로 쓰이는 경우는 약자를 나타내므로 거의 알파벳 발음대로 읽는 경향이 많기 때문이다.

예)I.M.F, O.E.C.D, ADM,…

⑫ 영문자 다음에 바로 조사가 나오지 않고 '('가 오게 되면 ')'까지는 괄호 앞의 내용을 설명하는 설명문이므로 무시하고 바로 다음에 나오는 한글과 비교한다.

예)World Wide Web(WWW)은

⑬ 영문 단독으로 쓰이더라도 이후의 문자열이 조사에 해당하는 문자열(표2)이면 조사로 간주하여 영문과 부쳐 주고 위의 규칙1)~12)을 적용한다.

예)school 는 학교이다. =>
school은 학교이다.

### 4.2. 특수 어형 사전 구축

위의 적용규칙에 맞지 않는 특수한 경우를 종종 볼 수 있다.

예) me(미), bag(백), ballet(발레), bariquant(바리캉), cantabile(칸타빌레), …외에 측정단위포함

이와 같은 경우 미리 특수어형사전을 구축하고 한·영 분리 후 바로 검색하여 사전에 존재하면 바로 오류판정으로 들어간다.
이 사전에는 영어 이외의 문자들도 포함하여 위의 발음 규칙에 어긋나는 내용을 첨가한다.

## 5. 실험 및 평가

기존의 시스템(한글 워드프로세서)[5]에서도 이와 비슷한 내용이 있지만 그 시스템은 음운을 가지고 처리하지 않는 것으로 보인다.

예) Smith, pentium, Intel, web,…

예에서처럼 고유명사인 경우나 사전에 존재하지 않는 단어를 입력할 경우엔 오류를 검출하지 못한다.
자세히 파악을 할 수는 없지만 이는 자체적인 알고리즘을 가지고 있거나 사전입력시 미리 등록을 해 놓았을 것으로 보인다. 하지만 위의 방법들은 새로운 단어를 입력 할 때 처리가 미비할 뿐만 아니라 메모리에도 커다란 부담이 아닐 수 없다.
본 논문에서 제안한 시스템은 도스용 Turbo C++에서 작성하여 Intel Pentium 150Mhz에서 시뮬레이션 하였다.
대상단어는 외래어 표기 용례에 나온 1775단어, 영어 표기식 지역명 115단어와 로마자 표기 인명[1,2,3] 중 알파벳으로 이루어진 276단어들이다.
실험 방법은 두 가지 경우를 모두 입력하여 오류검사를 한 후의 실패 확률을 표4에 기록하였다. 단, 위의 규칙에서 모두 허용하는 경우는 맞는 것으로 간주하였다. 특히, network등과 같은 경우 작성자의 특성에 따라 발음 될 수 있는 네트워크와 네트웍은 같은 것으로 인정하였다.
표 4에서는 기존 시스템[5]과 비교한 결과로 오류검출률이 아주 높음을 볼 수 있다. 특히 지명이나 인명처럼 사전에 등록되어 있지 않는 경우엔 거의 완벽에 가깝게 검출해낸다. 외래어 표기 용례에서 실패한 경우는 "pão, appliqué, chou a la créme,…" 등과 같이 알파벳이 아닌 특수 문자들이 입력되어 있을 경우와 알파벳을 사용하지만 영어 발음이 되지 않는 몇 가지 단어들(특수 어형사전에 등록되지 않은 단어)을 제외하고는 모두 성공함을 볼 수 있다.
본 시스템을 일반적인 맞춤법 검사에 포함하여 활용할 경우 보다 부드러운 문장을 만들 수 있으며 영한 번역시 신조어를 단어 그대로 표현할 때도 특별한 알고리즘 없이 바로 적용할 수 있는 장점이 있다.
또한, 약자를 많이 사용하는 여러 전문분야의 논문이나 전공서적을 편찬할 때 사용자가 일일이 조사의 옳고 그름을 비교할 필요가 없어지게 되어 보다 효율적인 편집작업을 도와 줄 수 있을 것이다.

## 6. 결론

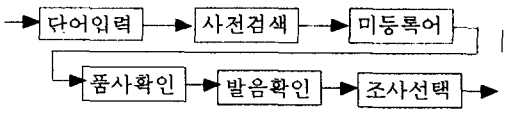
본 논문에서는 영어와 한글 조사의 결합을 자연스럽게 함을 목적으로 하고 있고 다음과 같은 특징을 가진다.

표 4 시스템 비교 결과(성공률 %)

| | 기존 시스템 | 받침을 이용한 시스템 |
|---|---|---|
| 외래어표기용례 (1775단어) | 47.93 | 99.39 |
| 지명 및 인명 (391단어) | 29 | 100 |

첫째, 단어의 옳고 그름에 관여하지 않고 오로지 발음상의 법칙으로만 처리되므로 신종어의 발생이나 약자입력에도 무리 없이 대처할 수 있다.
둘째, 이 간단한 알고리즘을 형태소 분석 이전에 수행함으로써 형태소 분석 오류를 최소한으로 줄일 수 있다.
특히 영한번역시 미 등록어에 대한 조사 처리를 효율적으로 할 수 있다.
영한번역에 적용과정을 알아보면 그림 2와 같다.

그림 2 영한번역시 적용방법




일반적인 번역 프로그램을 보면 미등록어일 경우 " Minsu은(는) ...."으로 수행됨을 볼 수 있다.
본 알고리즘을 사용하면 사전의 부담 없이 매끄러운 번역이 될 것이다.

마지막으로 본 시스템의 문제점은 첫째, 영어 알파벳 이외의 특수 문자를 입력할 경우 처리가 힘들다. 둘째, 모양은 영어 알파벳을 취하더라도 유럽어들 처럼 영어이외 국가의 발음을 그대로 수용할 경우, 처리에 오류를 일으킬 수 있다.

향후 발전 계획은 문자 중 알파벳이외의 문자가 단어에 포함될 경우 입력된 문장이 어느 나라 언어인지를 확인하고 그 나라 언어에 맞는 적용 규칙을 생성해 내는 것으로 문제를 해결 할 것이다.
또한 자주 사용되는 단어나 신조어의 경우 보다 정확한 결과를 얻도록 특별어형사전을 확장할 것이며 지금은 분량이 얼마 되지 않지만 확장되어 사전이 커질 경우 빠른 검색을 위하여 더블Trie 알고리즘[6]을 도입할 계획이다.

참고문헌

[1] 이희승, 안병희, 한글 맞춤법 강의, 신구문화사, 1994

[2] 최기호, 한글맞춤법 새 길라잡이, 보담, 1994

[3] 외래어 표기법, 1985년 12월 28일 문교부 고시

[4] 김재훈 외, 통합국어정보베이스를 위한 한국어 형태·통사 태그 설정, 한국과학기술원, 1996

[5] 호글97, 도서출판 호글과컴퓨터, 1997

[6] K. Morimoto, H. Iroguchi and J. I. Aoe, "A Retieval algorithm of Dictionaries by Using Two Trie Structures", 일본전자공학회논문집 D-II VOL. J76-D-11, No.11, pp.2374-2383, 1994