다중 언어 인명 검색에 관한 연구

조 영 화† · 송 재 용†† · 류 근 호†††

요 약

이 논문은 로마자 표기 한글 인명의 효율적 검색 방안의 제시와 규칙기반 다중 언어 인명 검색 시스템의 설계에 관한 것이다. 기존의 서지 정보 검색 시스템이나 논문 검색 시스템에서 사람마다 제각기 표기되고 있는 로마자표기 한글 인명을 효율적으로 검색하는데 상당한 어려움이 따랐다. 예를 들어, 한글 인명 "홍길동"의 로마자 표기는 "Hong, gildong", "Gildong Hong", "Hong kil dong", "Hong kil-dong"등 철자나 표기 양식이 매우 다양하다. 이 논문에서는 이렇게 다양하게 표기되는 로마자 표기 한글 인명을 효율적으로 검색하기 위해 규칙기반을 이용한 질의 확장법을 제안하고 질의분류기, 예외처리기, 질의확장기, 질의실행기, 예외규정집, 규칙기반으로 구성된 다중 언어 인명 검색 시스템을 설계한다. 또한 인명생성기의 구현과 실행을 통해 규칙기반 질의 확장법의 효율적 검색 가능성을 보이도록 한다.

A Study on a Multilingual Name Retrieval

Young Hwa Cho† · Jaeyong Song†† · Keun Ho Ryu†††

ABSTRACT

In this paper, we propose a method to retrieve english-written korean names efficiently and design a multilingual name retrieval system. It is very difficult to retrieve english-written korean names in typical IR systems. For example, "홍길동" is written in english as various forms such like "Hong, gildong", "Gildong Hong", "Hong kil dong", "Hong kil-dong" and so on. We not only propose a rule-based query expansion method to retrieve english-written korean names efficiently but also design a multiligual name retrieval system which is consisted of query classifier, exception handler, query expander, query executor, exception list and rulebase. Finally we will try to show that english-written korean names could be efficiently retrieved with rule-based name generator.

1. 서 론

정보 검색 연구 분야중의 하나로 다중 언어 정보 검색(multilingual information retrieval)이 있다. 다중 언어 정보 검색은 문서나 질의의 언어와는 상관없이 사용자의 질의로 정보검색을 수행할 수 있는 것을 의미한다[3]. 현재까지 연구된 정보 검색 기술의 많은 것들이 영어를 위주로한 것이었다. 그러나 영어위주의 정보 검색 기술은 비영어 정보 검색에 대한 적용과 도입에 많은 제한과 제약이 있고 이러한 문제를 해결하기 위해 다중 언어 정보 검색이 연구되어 왔다.

이 논문은 로마자로 표기되는 한글 인명의 효율적인 검색을 위해 다중 언어 인명 검색 시스템 모듈들의 설계와 인명 질의 확장을 위한 인명 생성기 구현을 그 내용으로 한다. 사용자는 질의어로써 한글 인명을 입

* 이 연구는 1997년도 연구개발정보센터의 연구비 지원에 의하여 수행되었음.

† 정 회 원 : 연구개발정보센터 부장
†† 준 회 원 : 충북대학교 대학원 컴퓨터과학과
††† 종신회원 : 충북대학교 컴퓨터과학과 교수

논문접수 : 1998년 3월 10일, 심사완료 : 1998년 7월 2일

력하면 인명생성기에서 자동으로 표기 가능한 로마자 표기 인명을 생성하여 초기 인명 질의를 확장 검색할 수 있게 하는 것이다. 한글 인명의 로마자표기는 근래에 들어 거의 일상화되어 있다고 해도 과언이 아니다. 국내 영자 신문이나 각종 신용카드 그리고 국내의 논문 등에도 빠짐없이 로마자 표기 한글 인명이 쓰이고 있다. 하지만 이들 로마자 표기 한글 인명을 살펴보면 사람마다 또는 언론사나 재계 및 학계 등에서 쓰이고 있는 표기법이 모두 제각기 임을 쉽게 알 수 있다. 예들 들어, "홍길동"의 로마자 표기는 "Hong, gildong", "Gildong Hong", "Hong kil dong", "Hong kil-dong"등 매우 다양하다.

이렇게 다양하게 표기되는 로마자표기 인명은 기존의 정보 검색 시스템에서 효율적으로 검색하기 어려웠다. 질의와 문헌간의 언어적인 차이에서 발생하는 검색의 문제를 해결하기 위해 지금까지 제시된 연구 방법으로는 크게 시소러스를 이용하는 방법[2,8], 기계번역을 이용하는 방법[3], 코퍼스를 이용하는 방법[10] 등이 있다. 그러나 이 논문에서는 인공지능 분야에서 추론의 도구로 사용되어온 규칙기반으로 사용자 질의를 확장하는 방법을 이용하였다. 로마자 표기 인명은 표기자마다 표기법이 매우 다를 수 있지만 거의 대부분이 발음을 중심으로 표기되고 있어 일련의 표기 규칙들을 쉽게 발견할 수 있고 기존의 시소러스나 기계번역 등에서 요구되는 정보구조체의 구축에 상당한 시간과 자원을 줄일 수 있다.

이 논문의 구성은 다음과 같다. 2장에서는 지금까지 사용되어온 로마자표기 표준안들과 질의 확장법을 소개하고 3장에서는 이 논문에서 제안하는 로마자표기 인명의 효율적 검색 방안에 대해서 소개한다 그리고 4장에서는 다중 언어 인명 검색을 위한 시스템의 설계와 규칙기반을 정의하고 C로 구현된 인명 생성기로 한글 인명을 로마자 표기 인명으로 변환 생성하는 것을 보인다. 마지막으로 5장에서 결론과 향후 연구 방향을 기술한다.

2. 관련연구

2.1 한글의 로마자표기 표준안

한글 인명의 로마자 표기가 표기자마다 매우 다양한 것은 지금까지 일관적이고 통일되지 못한 표준안의 부재를 한가지 원인으로 들 수 있다. 한글의 로마자 표기 표준안의 대표적인 것으로는 문교부고시 제 84-1호로 공표되어 교과서나 정부 간행물, 지명 및 도로명

〈표 1〉 주요 로마자 표기 표준안 비교
〈Table 1〉 Comparison of a few romanization rules

표준안 / 자음	국어의 로마자 표기법	M-R안	공업진흥청안	표준안 / 모음	국어의 로마자 표기법	M-R안	공업진흥청안
ㄱ	k,g	k,g	k	ㅏ	a	a	a
ㄴ	n	n,l	n	ㅑ	ya	ya	ya
ㄷ	t,d	t,d	t	ㅓ	ŏ	ŏ	eo
ㄹ	r,l	l,r,m	r,l	ㅕ	yo	yŏ	yeo
ㅁ	m	m	m	ㅗ	o	o	o
ㅂ	p,b	p	p	ㅛ	yo	yo	yo
ㅅ	s,sh	s,t,n	s	ㅜ	u	u	u
ㅇ	ng	ng	zero,ng	ㅠ	yu	yu	yu
ㅈ	ch,j	ch,j	c	ㅡ	ŭ	ŭ	eu
ㅊ	ch′	ch′	ch	ㅣ	i	i	i
ㅋ	k′	k′,k	kh	ㅐ	ae	ae	ae
ㅌ	t′	t′,t	th	ㅒ	yae	yae	yae
ㅍ	p′	p′,p	ph	ㅔ	e	e	e
ㅎ	h	h	h	ㅖ	ye	ye	ye
ㄲ	kk	kk	kk	ㅚ	oe	oe	oe
ㄸ	tt	tt	tt	ㅟ	wi	wi	wi
ㅃ	pp	pp	pp	ㅘ	wa	wa	wa
ㅆ	ss	ss	ss	ㅝ	wo	wŏ	weo
ㅉ	tch	tch	cc	ㅙ	wae	wae	wae
				ㅞ	we	we	we

〈표 2〉 인명 표기 유형 분류
〈Table 2〉 Classification of name writing type

유형	표 기 예	설 명
1	Hong gil dong	성과 이름의 순서로 음절별로 띄어 쓴다.
2	Gil dong Hong	이름과 성의 순서로 음절별로 띄어 쓴다.
3	Hong gildong	성과 이름의 순서로 성과 이름을 띄어 쓴다.
4	Hong gil-dong	성과 이름의 순서로 성과 이름을 띄어 쓰고 이름사이에 "-"를 넣어 쓴다.
5	Hong, gildong	성과 이름의 순서로 쓰고 성과 이름은 띄우고 사이에 코마를 넣는다.
6	Hong, gil-dong	유형 4와 5의 혼합형태이다.
7	Gildong Hong	이름과 성의 순서로 쓰고 이름과 성을 띄운다.
8	Gil-dong Hong	유형4와 같은 형태로 이름과 성의 순서를 바꾸어 쓴다.

등에 사용된 국어의 로마자 표기법이 있고 외국인이 만든 것으로 주한 미군이나 영자 신문에 주로 사용되었던 M-R(McCune-Reischauer)안 그리고 ISO(International Standardization Organization)에 제출을 위해 고안된 것으로 컴퓨터, 타자기 및 텔렉스 등과 같은 기계사용을 고려한 공업진흥청안 등이 있다[14]. 이들의 한글 자모별 표기예를 비교하면 <표 1>과 같다.

국어의 로마자 표기법과 M-R안은 표기법이 거의 유사하고, 공업진흥청안은 기계사용의 편의를 위해 만들어졌기 때문에 어깨점과 같은 특수기호를 사용하지 않았다. 또한 인명 표기와 관련하여 국어의 로마자 표기법에서는 성과 이름을 띄어쓰고 이름 사이에는 '-'를 넣을 수 있다고 밝히고 있다. 이러한 여러 가지 표기법과 사용자들의 인명 표기 사례 분석을 통해 성과 이름이 3자인 한글 인명의 로마자 표기 유형은 대략 8가지로 구분해 볼 수 있다. 그 내용은 <표 2>와 같다.

2.2 질의 확장 기법

질의 확장 기법(query expansion methods)으로는 구축 방법에 따라 수작업에 의한 방법(manual query expansion), 반자동적인 방법(semi-manual expansion), 자동적인 질의 확장 방법(automatic query expansion)으로 나눌 수 있고 확장 기법에 따라 구문정보에 기반한 방법, 관련 정보에 기반한 방법, 자동적인 용어 분류에 의한 방법, 문헌분류에 의한 방법, 개념에 기반한 방법, 어구에 기반한 방법등으로 분류할 수 있다[13]. 구문 정보에 기반한 방법은 용어간의 관계를 언어적인 지식과 용어들이 어느 정도로 같이 쓰이느냐 하는 통계치에 따라서 설정하고 문법과 사전을 이용해서 용어들을 추출해 내고 이들 용어들을 질의어에 추가시키는 방법이다[4,7]. 그러나 이 방법은 확장 전의 질의보다 큰 효과를 얻기 힘들다[9]. 두번째로 연관 정보에 기반한 방법은 의사시소러스(pseudothesaurus)[9,10]와 같은 전체적인 정보구조체나 최소신장트리[11]를 만들기 위해 관련정보(relevance information)를 이용하는 것으로 이렇게 만들어진 정보구조체에 의해 질의를 확장한다. 그러나 이 방법은 관련정보를 사용하는데 제약이 많다는 단점이 있다. 세번째로 자동적인 용어 분류에 의한 방법은 연관가정(association hypothesis)에 기반해서 용어들간의 유사성을 먼저 계산한 다음 용어들을 유사성 초기값(similarity threshold value)[5,12]을 설정해 분류하고, 질의어들을 포함하는 모든 용어들을 추가시키는 방법으로 질의 확장을 한다. 그러나 이 방법은 실효성이 없다는 문제가 있다[5,6,12]. 네번째로 문헌 분류를 이용한 방법은 문헌에 대해 우선적으로 분류한 다음 드물게 사용되는 용어들을 동일한 용어클래스(thesaurus class)[1]에 집단화 시키고 문헌들과 질의들의 색인은 시소러스 클래스를 이용해 용어를 교체하거나 색인 데이터에 시소러스를 추가하여 수행하는 것이다. 단점으로는 몇가지 변수에 의해 검색 성능의 영향이 심하고[2] 비용이 많이 든다는 것이다. 다섯번째의 유사어 시소러스(similarity thesaurus)는 용어 집합의 색인에 따른 용어-용어 유사 메트릭스라 할 수 있다. 벡터공간모델(vector space model)에서 질의의 용어유사도의 확률을 평가하는데 확률적 방법이 쓰이고 이 기법이 쓰인 예로는 SMART를 들 수 있다. 여섯번째의 어구에 기반한 방법은 텍스트내의 어구들을 결정하고 그 어구들을 표현하는 용어들로 용어-어구관계

사전(term-vs-phrase assocication thesaurus)를 만든다. 이 방법은 용어-용어, 용어-명사용어, 용어-동사용어, 용어-형용사, 부사 관계 사전과 같은 모든 종류의 용어기반관계 사전의 기능을 갖는다. 질의어는 관계사전으로부터 질의와 가장 관계 깊은 어구들을 추가해 질의를 확장한다. 이 방법도 상당한 검색 성능의 향상을 이룰 수 있다.

3. 로마자 표기 인명 검색 방법

이 장에서는 인공지능 연구 분야에서 추론의 도구로 이용되고 있는 규칙기반을 이용하여 질의 확장 방법으로 로마자 표기 한글 인명을 검색하는 방안을 소개한다. 로마자로 표기된 한글 인명의 검색을 위해 기존의 시소러스와 같은 인명사전을 이용하는 방법도 고려될 수 있지만 한글 인명의 로마자표기 규칙을 이용함으로써 시소러스 구축 및 탐색에 대한 부담을 줄일 수 있는 장점이 있다.

3.1 규칙기반의 구축

다중 언어 인명 표기 규칙기반은 질의 확장시 참조되어 인명 질의를 확장하는데 이용된다. 규칙기반을 구성하는 각 규칙들은 조건부와 확장부로 나뉜다. 조건부에는 확장가능한 문자들이 기술되고 확장부에는 조건부의 문자들을 확장하는 문자들로 구성된다. 이러한 규칙들의 기본 구조는 다음과 같다.

R# : A [예약어 B] => C

R#는 규칙의 식별자이고 A는 조건부의 문자열, B는 A의 조건을 나타내는 문자열, 그리고 C는 확장부의 문자열이다. 다중 언어 인명 검색을 위한 규칙들은 조건부의 형태에 따라 2가지로 분류된다. 하나는 조건없이 확장되는 단순 확장 규칙이고 또 다른 하나는 선후행 문자에 따라 다르게 확장되는 연관 확장 규칙이다. 단순 확장 규칙은 기본적인 규칙의 형태와 같고 연관 확장 규칙은 <표 3>의 예약어를 사용하여 조건부를 구성한다.

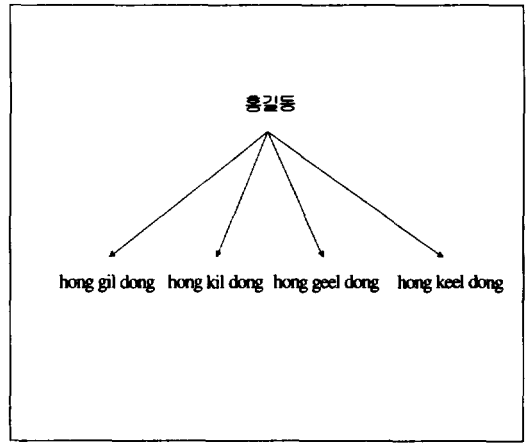
이러한 규칙들은 다시 한-영 인명 확장 규칙, 영-한 인명 확장 규칙의 2가지 유형으로 구분해 볼 수 있다. 한-영 인명 확장 규칙은 사용자 질의가 한글 표기 인명일 때 로마자 표기로 확장하는 규칙이고, 영-한 인명 확장 규칙은 로마자 표기 인명을 한글 표기 인명으로 확장하는 규칙이다. 이들 규칙의 적용시 질의가 한글 표기 질의일 때는 한-영 인명 확장 규칙만을 적용시켜 로마자표기 인명으로 확장시키고 질의가 로마자

〈표 3〉 규칙기반 구축을 위한 예약어
〈Table 3〉 Reserved word for rule-base

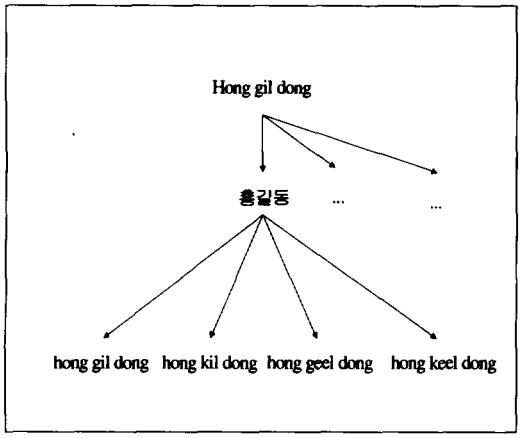
예약어	
before	후행문자를 검사할 때
after	선행 문자를 검사할 때
asTop	문자가 초성으로 쓰일 때
asBottom	문자가 종성으로 쓰일 때
VOWEL	모음
CONSONANT	자음
NULL	묵음

〈표 4〉 규칙의 예
〈Table 4〉 Example of Rules

규 칙
R26 : "ㅅ" asBottom => NULL R133: "ng" after VOWEL => ㅇ R8 : "ㅇ" asTop => NULL R27 : "ㅇ" asBottom => ng R129: "yo" => ㅕ, ㅛ R130: "yu" => ㅕ, ㅠ R131: "x" => NULL R132: "z" => NULL R39: "ㅅ" before "ㅣ" => sh



(그림 1) 한글 인명 표기의 확장
(Fig. 1) Expansion of korean name



(그림 2) 로마자 표기 인명의 확장
(Fig. 2) Expansion of english-written korean name

표기일 때는 영-한 인명 확장 규칙을 적용시켜 한글인명으로 확장한 다음, 다시 확장된 한글표기 인명을 한-영 인명 확장 규칙을 적용시켜 로마자표기 인명을 확장한다. 이에 대한 예는 (그림 1)과 (그림 2)와 같다.

한-영 인명 확장 규칙과 영-한 인명 확장 규칙의 예는 <표 5>와 같다.

〈표 5〉 규칙의 종류
〈Table 5〉 Sort of rules

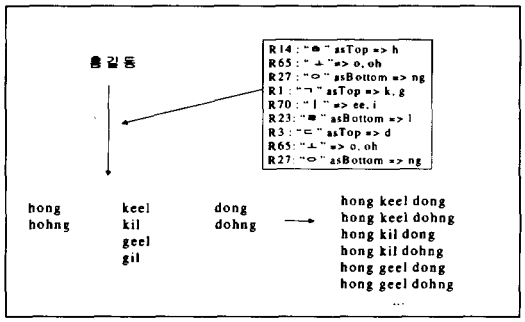
한-영인명 확장규칙	영-한인명 확장규칙
R61: "ㅏ" => a, ah	R82: "ae" => ㅐ
R62: "ㅑ" => ya	R83: "a" => ㅏ
R63: "ㅓ" => ur, o, u, eo	R98: "k" => ㄱ

3.2 질의의 생성과 확장

사용자로부터의 인명 질의는 질의의 분석을 통해 규칙기반에서 정의된 규칙에 따라 자동으로 생성 확장된다. 질의의 생성과 확장은 다음과 같은 일련의 과정을 통해 수행된다.

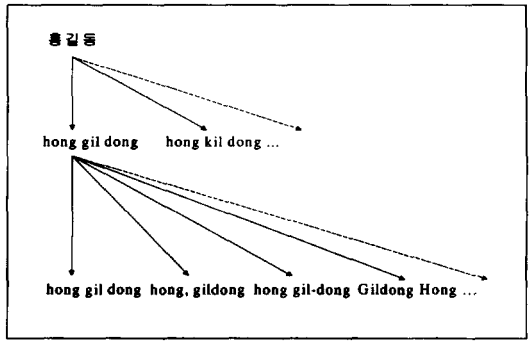
단계1: 사용자에 의한 초기 질의의 분석
단계2: 인명 질의 추출
단계3: 규칙기반을 이용한 인명 질의 확장
단계4: 확장 질의 구성

예를 들어, 사용자 질의가 "저자 = '홍길동'이고 논문 ='로마자표기법'"이라고 하면 질의어 분석을 통해 인명 질의 "저자 = '홍길동'"을 추출하고 '홍길동'은 한글 인명을 로마자로 변환 확장하는 규칙기반을 이용하여 로마자표기 인명으로 확장된다. (그림 3)은 한글 인명이 로마자로 변환되는 과정이다.



(그림 3) 한글 인명의 확장 과정
(Fig. 3) Expand process of english-written korean name

위의 질의처리로 생성된 질의는 다시 인명표기 유형별로 확장이 수행된다. (그림 4)는 한글 표기 인명이 유형별로 확장되는 과정이다.



(그림 4) 인명 표기 유형별 확장
(Fig. 4) Expansion of name by type

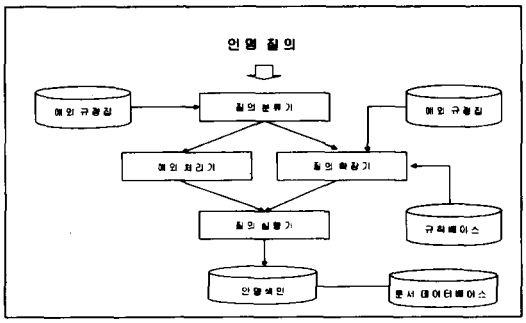
(그림 4)와 같이 유형별 로마자 인명 표기까지 인명 질의가 확장이 되면 최종적으로 질의를 생성하게 된다. 따라서 초기 질의 "저자 = '홍길동'이고 논문 ='로마자표기법'"은 다음과 같은 확장 질의로 구성될 수 있다.

AND('로마자표기법', OR('홍길동','hong gil dong', 'hong, kildong','kildong hong')

4. 로마자 표기 인명 검색을 위한 다중 언어 인명 검색 시스템 및 규칙기반

4.1 시스템의 구성

다중 언어 인명 검색 시스템은 기존의 정보 검색 시스템에서 인명 질의 처리 부분을 확장한 것으로써 질의 분류기, 질의 확장기, 예외 처리기, 질의 실행기, 규칙기반, 예외 규정집으로 설계하였다. (그림 5)에서 보여 주는 바와 같이 사용자가 원하는 인명을 질의하면 다중 언어 인명 검색 시스템은 예외 규정집과 규칙기반을 이용하여 관련된 정보를 사용자에게 제공하게 된다.



(그림 5) 다중 언어 인명 검색 시스템의 구조
(Fig. 5) Structure of multilingual name retrieval system

질의 분류기(query classifier)는 사용자 질의를 입력받아 질의를 확장 가능한 질의와 그렇지 않은 질의로 분류하는 모듈로서 예외 규정집을 이용하여 질의 확장이 불가능한 것을 분류한다. 예외 처리기(exception handler)는 질의분류기를 통해 인명 질의 확장이 불가능한 것을 처리하는 것으로 고유명사화한 인명, 두자리 이상의 성을 가진 인명이나 외자 이름을 가진 인명 그리고 기타 완전하지 못한 인명 등을 대상으로 한다. 질의 분류기를 통한 로마자 표기 한글 인명은 질의 확장기(query expander)에서 질의의 확장이 이루어 진다. 질의어의 생성 및 확장은 질의어의 분석을 통해 미리 구축해 놓은 규칙기반을 이용하여 1차적으로 철자상의 확장이 이루어지게 되고 2차적으로는 인명 표기 유형별로 확장을 하게 된다. 질의실행기(query executor)는 예외처리기와 질의확장기를 통한 질의어를 인명 색인에 의해 검색을 수행하는 부분으로 확장된 인명 질의를 초기 질의에 추가하여 최종적으로 질의를 처리하는 모듈이다. 예외 규정집은 규칙기반과 같은 형식으로 구축되지만 규칙기반의 탐색이전에 탐색되어 규칙기반 탐색여부를 결정하게 된다. 규칙기반은 앞장에서 소개했던 것처럼 사용자에 의한 인명 질의를 자동으로 변환 확장할 수 있도록 하는 규칙들을 사전에 정의해 놓은 것으로 수작업으로 구축된다. 규칙들의 유형은 한-영 인명 확장규칙, 영-한 인명 확장규칙의 두가지 유형이 있다.

4.2 인명의 자동 생성을 위한 규칙기반

한글의 초성은 기본자음 14자와 쌍자음 5자의 19자, 중성은 기본 모음 10자와 복모음 11자의 21자, 그리고 종성으로 기본 자음 14자와 쌍자음 2자의 16자가 쓰이는데 <표 6>의 규칙기반은 200여개의 인명데이터의 분석을 통해 수작업으로 자음에 대한 한-영인명 변환규칙 39개, 모음에 대한 한-영인명 변환규칙 42개, 영-한인명 변환규칙 52개, 예외규정 9개를 정의하였다. 한글 자음은 대부분이 영문 자음에 대하여 1대1일로 대응이 되거나 2개까지 대응이 되기 때문에 규칙의 정의가 쉬운 반면 한글 모음은 선후행 자음에 따라 여러가지로 표기되기 때문에 규칙의 정의가 까다롭다. 예로 규칙 68번(R68: "ㅠ" => yu, yoo, you, u, ue)은 하나의 한글 모음에 대하여 5개의 로마자가 사용될 수 있다.

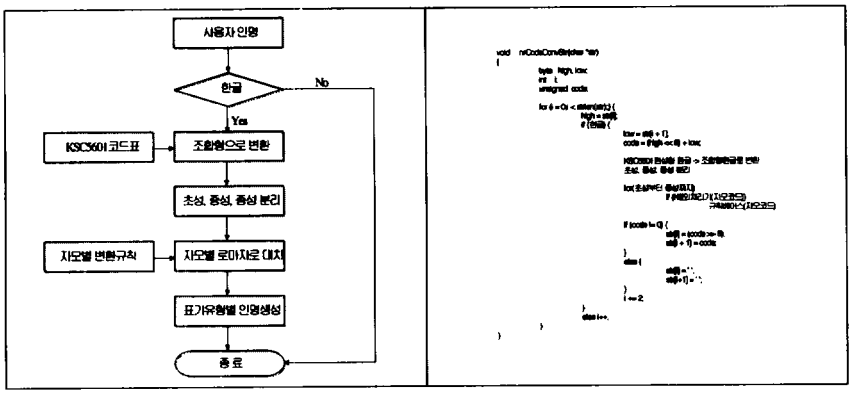
(그림 6)은 <표 6>의 규칙기반을 이용하여 한글 인명을 로마자표기 한글 인명으로 변환 생성하는 인명 생성 알고리즘이다. 사용자에 의한 초기 질의에서 추출된 인명은 문자열의 앞 첫 자부터 초성, 중성, 종성별로 로마자로 변환된다. 일반적으로 네트웍이나 대부분의 시스템이 완성형 한글을 사용하기 때문에 완성형 한글을 조합형으로 변환한 다음 한글 자모를 분리하게 된다. 이를 위해 KSC5601코드표를 완성형 한글을 조합형 한글로 변환하는데 이용하고 초성, 중성, 종성으로 분리된 한글 자모는 규칙기반을 이용하여 로마자로 자동 변환되어 최종적으로 인명 표기 유형으로 인명을 생성 확장하게 된다.

<표 6>의 규칙기반과 (그림 6)의 인명 확장 알고리즘을 이용하여 (그림 7)과 같이 인명 생성기를 구현 실행하였다. (그림 7)의 인명 생성기의 실행은 규칙기반 중에서 한-영인명 확장규칙과 예외규정을 이용한 것이다

성씨로 쓰이는 "이"나 "박"은 대부분이 "lee"나 "Rhee"

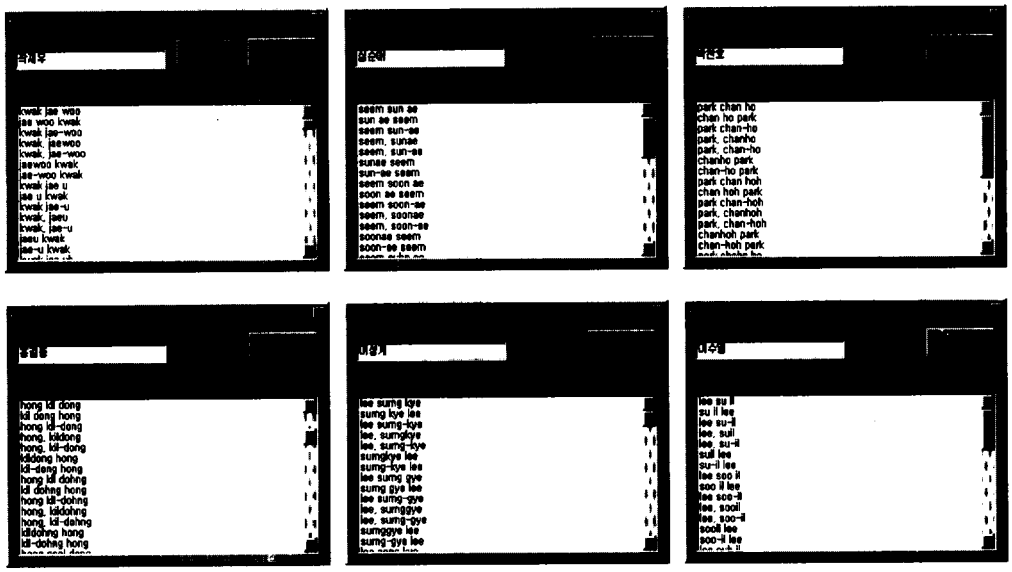
〈표 6〉 변환규칙
〈Table 6〉 Transition rules

규 칙 기 반			예외규정집
한-영인명 변환규칙(자음)	한-영인명 변환규칙(모음)	영-한인명 변환규칙	
R1: "ㄱ" asTop =〉 k, g	R40: "ㅏ" after "ㅇ" =〉 a, ah	R82: "ae" =〉 ㅐ	E1: "이" asTop =〉 lee, rhee
R2: "ㄴ" asTop =〉 n, lh	R41: "ㅑ" after "ㅇ" =〉 ya	R83: "ah" =〉 ㅏ	E2: "박" asTop =〉 park
R3: "ㄷ" asTop =〉 d	R42: "ㅓ" after "ㅇ" =〉 ur, o, u, eo	R84: "a" =〉 ㅏ	E3: "이순신" =〉 Yi Sun-shin
R4: "ㄹ" asTop =〉 r, l	R43: "ㅕ" after "ㅇ" =〉 yu, you, yeo, yo	R85: "b" =〉 ㅂ	E4: "이승만" =〉 Syngman Rhee
R5: "ㅁ" asTop =〉 m	R44: "ㅗ" after "ㅇ" =〉 o, oh	R86: "ch" =〉 ㅊ, ㅈ	E5: "lee" asTop =〉 이
R6: "ㅂ" asTop =〉 b, p	R45: "ㅛ" after "ㅇ" =〉 yo	R87: "d" =〉 ㄷ	E6: "rhee" asTop =〉 이
R7: "ㅅ" asTop =〉 s	R46: "ㅜ" after "ㅇ" =〉 woo, u, uh	R88: "eui" =〉 ㅢ	E7: "park" asTop =〉박
R8: "ㅇ" asTop =〉 NULL	R47: "ㅠ" after "ㅇ" =〉 yu, yoo, you	R89: "ee" =〉 ㅣ, ㅢ	E8: "Yi Sun-shin" =〉 이순신
R9: "ㅈ" asTop =〉 j, ch	R48: "ㅡ" after "ㅇ" =〉 eu	R90: "ei" =〉 ㅔ	E9: "Syngman Rhee" =〉 이승만
R10: "ㅊ" asTop =〉 ch	R49: "ㅣ" after "ㅇ" =〉 i	R91: "eo" =〉 ㅓ	
R11: "ㅋ" asTop =〉 k	R50: "ㅐ" after "ㅇ" =〉 ae	R92: "eu" =〉 ㅡ	
R12: "ㅌ" asTop =〉 t	R51: "ㅒ" after "ㅇ" =〉 yae	R93: "f" =〉 NULL	
R13: "ㅍ" asTop =〉 p	R52: "ㅔ" after "ㅇ" =〉 NULL	R94: "g" =〉 ㄱ	
R14: "ㅎ" asTop =〉 h	R53: "ㅖ" after "ㅇ" =〉 NULL	R95: "h" =〉 ㅎ	
R15: "ㄲ" asTop =〉 NULL	R54: "ㅢ" after "ㅇ" =〉 NULL	R96: "i" =〉 ㅣ	
R16: "ㄸ" asTop =〉 NULL	R55: "ㅟ" after "ㅇ" =〉 uy, wi	R97: "j" =〉 ㅈ	
R17: "ㅃ" asTop =〉 NULL	R56: "ㅝ" after "ㅇ" =〉 ee, eui	R98: "k" =〉 ㄱ	
R18: "ㅆ" asTop =〉 NULL	R57: "ㅘ" after "ㅇ" =〉 wa	R99: "lh" =〉 ㄴ	
R19: "ㅉ" asTop =〉 NULL	R58: "ㅙ" after "ㅇ" =〉 wo, weo	R100: "l" =〉 ㄹ	
R20: "ㄱ" asBottom =〉 k	R59: "ㅞ" after "ㅇ" =〉 NULL	R101: "m" =〉 ㅁ	
R21: "ㄴ" asBottom =〉 n	R60: "ㅙ" after "ㅇ" =〉 NULL	R102: "n" =〉 ㄴ	
R22: "ㄷ" asBottom =〉 d	R61: "ㅏ" =〉 a, ah	R103: "oo" =〉 ㅜ	
R23: "ㄹ" asBottom =〉 l	R62: "ㅑ" =〉 ya	R104: "oh" =〉 ㅗ	
R24: "ㅁ" asBottom =〉 m	R63: "ㅓ" =〉 ur, o, u, eo	R105: "oi" =〉 ㅚ	
R25: "ㅂ" asBottom =〉 p	R64: "ㅕ" =〉 yu, you, yeo, yo	R106: "o" =〉 ㅓ	
R26: "ㅅ" asBottom =〉 NULL	R65: "ㅗ" =〉 o, oh	R107: "p" =〉 ㅂ, ㅍ	
R27: "ㅇ" asBottom =〉 ng	R66: "ㅛ" =〉 yo	R108: "q" =〉 NULL	
R28: "ㅈ" asBottom =〉 NULL	R67: "ㅜ" =〉 u, oo, uh	R109: "r" =〉 ㄹ	
R29: "ㅊ" asBottom =〉 NULL	R68: "ㅠ" =〉 yu, yoo, you, u, ue	R110: "sh" =〉 ㅅ	
R30: "ㅋ" asBottom =〉 NULL	R69: "ㅡ" =〉 eu	R111: "s" =〉 ㅅ	
R31: "ㅌ" asBottom =〉 NULL	R70: "ㅣ" =〉 ee, i	R112: "t" =〉 ㅌ, ㄷ	
R32: "ㅍ" asBottom =〉 NULL	R71: "ㅐ" =〉 ae, yae	R113: "ue" =〉 ㅠ	
R33: "ㅎ" asBottom =〉 NULL	R72: "ㅒ" =〉 yae	R114: "uh" =〉 ㅜ	
R34: "ㄲ" asBottom =〉 NULL	R73: "ㅔ" =〉 e, ei	R115: "ur" =〉 ㅓ	
R35: "ㄸ" asBottom =〉 NULL	R74: "ㅖ" =〉 ye	R116: "uy" =〉 ㅟ	
R36: "ㅃ" asBottom =〉 NULL	R75: "ㅚ" =〉 oi	R117: "u" =〉 ㅓ, ㅜ, ㅠ,	
R37: "ㅆ" asBottom =〉 NULL	R76: "ㅟ" =〉 uy, wi	R118: "v" =〉 NULL	
R38: "ㅉ" asBottom =〉 NULL	R77: "ㅢ" =〉 ee, eui	R119: "weo" =〉 ㅝ	
R39: "ㅅ" before "ㅣ" =〉 sh	R78: "ㅘ" =〉 wa	R120: "wa" =〉 ㅘ	
	R79: "ㅝ" =〉 wo, weo	R121: "wo" =〉 ㅝ	
	R80: "ㅞ" =〉 NULL	R122: "w" =〉 ㅇ	
	R81: "ㅙ" =〉 NULL	R123: "x" =〉 NULL	
		R124: "yeo" =〉 ㅕ	
		R125: "yoo" =〉 ㅠ	
		R126: "you" =〉 ㅕ, ㅠ	
		R127: "ya" =〉 ㅑ	
		R128: "ye" =〉 ㅖ	
		R129: "yo" =〉 ㅕ, ㅛ	
		R130: "yu" =〉 ㅕ, ㅠ	
		R131: "x" =〉 NULL	
		R132: "z" =〉 NULL	
		R133: "ng" after VOWEL =〉 ㅇ	



(그림 6) 인명 확장 알고리즘
(Fig. 6) Algorithm of name expansion

그리고 "Park"으로 동일한 표기법을 쓰기 때문에 확장 규칙을 이용하지 않고 예외 사항으로 두었다. 실행 결과 인명 확장의 정확도는 자음보다는 모음의 확장 규칙에 많이 의존되는 것을 알 수 있었다. <표 7>은 인명생성기에 의한 결과로서 6개의 한글 인명 대하여 각각 32, 8, 4, 3, 12, 24개의 인명을 생성하였다. 실험에 사용된 규칙은 단모음과 장모음의 구별이 없기 때문에 중성의 확장이 많아져 결과적으로 많은 수의 인명 표기가 생성되었다. 예로 "홍"이나 "길"과 같은 경우 "hohng"보다는 "hong"이, "keel"보다는 "kil"이 더 일반적인 표기법이라 할 수 있고 장모음과 단모음을 구별하여 규칙을 정의하면 생성되는 기본형의 인명수도 32개에서 4개로 줄일 수 있다.



(그림 7) 인명 생성기의 실행 화면
(Fig. 7) Screen shot of name generator

〈표 7〉 인명 확장 결과
〈Table 7〉 Result of name expansion

확장 / 예	성			가운데 이름			마지막 이름			확장결과(수)
	초성	중성	종성	초성	중성	종성	초성	중성	종성	
홍길동	h	o, oh	ng	g,k	i, ee	l	d, t	o, oh	ng	32
이성계	lee, rhee로 예외처리			s	ur,o,u,eo	ng	k, g	ye	null	8
박찬호	park로 예외처리			ch	a, ah	n	h	o, oh	null	4
이수일	lee로 예외처리			s	u,oo,uh	null	i	l	null	3
심순애	s, sh	i, ee	m	s	u,oo,uh	n	null	ae	null	12
곽재우	k, g	wa	k	j, ch	ae, yae	null	null	woo,u,uh	null	24

5. 결 론

이 논문에서는 로마자로 표기되는 한글 인명의 효율적 검색을 위해 규칙기반 질의 확장법을 제시하였고 다중 언어 인명 검색 시스템을 설계하였다. 또한 인명 질의 확장을 위한 규칙으로 자음에 대한 한-영인명 변환규칙 39개, 모음에 대한 한-영인명 변환규칙 42개, 영-한인명 변환규칙 52개, 예외규정 9개를 정의하였다. 로마자로 표기되는 한글 인명은 사람마다 표기법이 매우 다양하지만 발음을 위주로 표기된다는 특성이 있고 일련의 표기 규칙성이 있기 때문에 규칙에 기반한 질의 확장법을 이용하여 로마자표기 한글 인명을 검색할 수 있는 가능성을 보였다.

제안된 다중 언어 인명 검색 시스템은 질의분류기, 질의확장기, 예외처리기, 질의실행기, 규칙기반, 예외규정집으로 구성된다. 질의분류기는 사용자 질의에서 인명 질의를 추출하여 확장 가능성에 따라 질의를 분류하고, 질의 확장기는 규칙기반을 이용하여 사용자 질의를 가능한 표기법들로 확장한다. 예외처리기는 질의 확장이 어렵거나 일반적인 표기법들을 처리하고, 질의 실행기는 확장된 질의나 예외처리기를 통한 질의를 평가하여 최종적인 질의를 구성하여 실행한다. 규칙기반은 한글 인명의 로마자 표기 규칙들을 정의한 것이고 예외규정집은 규칙에 기반하여 인명 질의를 확장하기 어려운 것을 기술해 놓은 것이다. 마지막으로 인명 생성기는 한글 인명의 로마자표기 규칙에 기반하여 한글 인명 질의를 로마자표기 인명으로 확장하는 것으로써 인명의 생성은 한글 자모별로 철자상의 확장을 통한 다음 다시 표기 유형별로 인명을 생성하도록 하였다. 연구된 검색 시스템의 응용은 기존의 시스템에 하나의 모듈로서 추가되거나 논문 검색 시스템과 같이 인명 검색이 요구되는 시스템을 기반으로 메타검색 시스템의 형태로 구현될 수 있다.

이 논문에서 구현한 인명 생성기의 성능 평가는 한글의 로마자표기에 대한 단일한 기준이 없고 일반적인 표기법이라는 주관적인 기준으로는 객관적으로 평가하기 어렵기 때문에 실제 검색 시스템에서 확장하지 않은 질의와 확장 질의를 적용시켜 얻은 결과에 대하여 비교 검토가 이루어져야 하고 인명생성기를 통해 생성되는 인명의 수를 줄일 수 있는 방안과 사용자의 판단에 도움을 줄 수 있는 순위부여 방식의 연구가 필요하다.

참 고 문 헌

[1] Croch, C.J., "An approach to the automatic construction of global thesauri", Information Processing and Management, 26(5), pp.629-40, 1990.

[2] Croch, C.J., Yong, B., "Experiments in Automatic Statistical Thesaurus Construction", SIGIR'92, pp. 77-87, June 1992.

[3] Douglas W.O., Bonnie J.D., "A Survey of Multilingual Text Retrieval", *UMIACS-TR-96-19 CS-TR-3615*, 1996.

[4] Grefenstette, G., "Use of syntactic context to produce term association lists for retrieval", SIGIR 92, pp.89-97, June 1992.

[5] Minker, J., Wilson, G.A., Zimmerman, B.H., "An evaluation of query expansion by the addiction of clustered terms for a document retrieval system", Information Storage and Retrieval, 8(6), pp.329-48, 1972.

[6] Peat, H.J., Willett, P., "The limitation of term co-occurrence data for query expansion in document retrieval system", J. of the ASIS, 42(5), pp.378-83, 1991.

[7] Ruge, G, "Experiments on linguistically-based term associations", Information Processing and Management, 28(3), p317-32, 1992.

[8] Salton, G., "Experiments in automatic thesaurus construction for information retrieval", Information Processing, 1, pp.115-123, 1971.

[9] Salton, G., "Automatic term class construction using relevance - a summary of work in automatic pseudoclassification", Information Processing & Management, 16(1), pp.1-15, 1980.

[10] Salton G., "Automatic Text Processing", Addison-Wesley Publishing Company, 1988.

[11] Smeaton, A.F., Van Rijsbergen, C.J., "The retrieval effect of query expansion on a feedback document retrieval system", The Computer Journal, 26(3), pp.239-46, 1983.

[12] Sparck-Jones, K., Barber, E.B., "What makes an automatic keyword classification effective?", J.of the ASIS, 18, pp.166-175, 1971.

[13] Qui Yonggang and Frei,H.P., "Concept based query expansion", SIGIR'93, pp.160-169, 1993.

[14] 김복문, "한·일 로마자 표기의 비교연구", 무역출판사, 1996

조 영 화

1977년 성균관대학교 통계학과(학사)

1990년 성균관대학교 정보처리학과(석사)

1991년~1993년 KIST 시스템공학연구소 단장

1993년~현재 KAIST 연구개발정보센터 부장

1996년 충북대학교 컴퓨터과학과 박사과정 수료

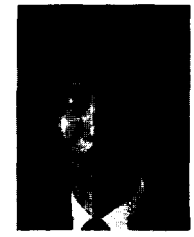
관심분야 : 정보검색, 데이터마이닝, 소프트웨어공학

송 재 용

1997년 충북대학교 경영정보학과(학사)

1997년~현재 충북대학교 대학원 컴퓨터과학과 석사과정

관심분야 : 시간지원 데이터베이스, 정보검색, 웹에이전트, 데이터마이닝



류 근 호

1976년 숭실대학교 전산학과(학사)

1980년 연세대학교 산업대학원 전산전공(공학석사)

1988년 연세대학교 대학원 전산전공(공학박사)

1976년~1986년 육군군수 지원사 전산실(ROTC 장교), 한국전자통신연구원(연구원), 한국방송통신대 전산학과(조교수) 근무

1989년~1991년 Univ. of Arizona Research Staff(TempIS 연구원, Tempral DB)

1986년~현재 충북대학교 컴퓨터과학과 교수겸 컴퓨터정보통신 연구소장

관심분야 : 시간지원 데이터베이스, 시공간 데이터베이스, 정보검색시스템, 객체 및 지식베이스 시스템