제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

한국어 자모 혼동행렬 기반 유사 외래어 표기 검출 기법*

권순호, 권혁철
부산대학교 컴퓨터공학과
e-mail:soonhok7@pusan.ac.kr, hckwon@pusan.ac.kr

Equivalent Writing of Loanwords Detection Method based on Korean Alphabet Confusion Matrix

Soonho Kwon, Hyuk-Chul Kwon
Dept of Computer Science & Engineering, Pusan National University

요 약

최근 한국어 문서에는 한국어뿐만 아니라 외래어 표기 등이 혼합되어 사용되고 있다. 외래어 표기는 한 단어에 대해 한 개만 존재하는 것이 아니라 여러 개의 다른 표기로 사용되고 있다. 이러한 표기상 불일치는 하나의 단어가 다른 개념으로 인식되어 정보검색 시스템의 성능 저하의 원인이 된다. 따라서 정보검색 시스템의 성능 향상을 위해 여러 외래어 표기를 같은 개념으로 인식하는 시스템이 필요하다. 본 논문에서는 한국어 자모 혼동행렬을 기반으로 한 유사 외래어 표기 검출 기법을 제안한다. 제안한 기법에 따라 유사 외래어 표기를 검출해줌으로써 정보검색 시스템의 성능을 향상할 수 있다.

1. 서론

최근 한국어 문서에는 한국어뿐만 아니라 영어 등의 외국어 문자 표기와 그에 해당하는 외래어 표기 등이 혼합되어 사용되고 있다. 이러한 현상은 이들 단어를 포함한 문서에 대한 검색을 어렵게 만드는 요인이 된다. 특히 외래어 표기는 한 단어에 대해 한 개만 존재하는 것이 아니라 여러 개의 다른 표기가 통용되는 것이 보편적이다. 예를 들어 영어 “data”에 대해 “데이타”, “데이터” 등의 표기는 모두 같은 개념을 표현하지만, 표기상 불일치로 말미암아 다른 개념으로 인식되어 정보검색 시스템의 성능 저하가 일어난다[3]. 따라서 정보검색 시스템의 성능 향상을 위해 여러 외래어 표기를 같은 개념으로 인식하는 시스템이 필요하다.

본 논문에서는 유사 외래어 표기 데이터에서 자모 단위의 혼동행렬을 구성하고, 이를 이용하여 외래어 표기의 유사도를 비교하여 유사 외래어 표기를 검출하는 기법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 영어와 한국어의 음성적 유사도를 비교하는 기존 연구 및 혼동행렬(Confusion Matrix)을 사용하는 검색 방법에 대한 연구를 살펴본다. 3장에서는 본 논문에서 제안하는 유사 외래어 표기 검출 기법을 자세히 소개한다. 4장에서는 실험을 통해 N-gram 방법과 본 논문에서 제안한 기법을 비교하여 성능을 평가한다. 마지막으로 5장에서는 결론과 향후 연구 과제를 논하고자 한다.

* 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2007-0054887).

2. 관련 연구

N-gram 방법[8]은 간단하고 언어에 독립적으로 적용할 수 있는 두 문자열의 비교 방법이다. 두 문자열의 N-gram들이 많이 일치할수록 더 유사하다고 판단한다. 그러나 N-gram 방법은 단어의 음성적 요소를 고려하지 않아 유사 외래어 표기를 검출하는 데는 적합하지 않다.

Soundex 알고리즘[5]은 영어 단어의 음성적 유사도를 구하는 대표적인 알고리즘으로 다섯 가지의 과정으로 이루어진다.

Step 1. 첫 번째 문자를 유지한다.

Step 2. 나머지 문자가 A, E, I, O, U, H, W, Y이면 ‘0’으로 치환한다.

Step 3. 문자를 다음의 숫자로 치환한다.

B, F, P, V: 1

C, G, J, K, Q, S, X, Z: 2

D, T: 3

L: 4

M, N: 5

R: 6

Step 4. 중복되는 숫자를 하나만 남기고 제거한다.

Step 5. ‘0’을 모두 제거하고, 네 자리 코드로 변환한다.

KODEX 알고리즘[1]은 soundex 알고리즘을 한글에 적용한 것으로, 초성과 종성에 코드번호를 부여하여 한글 단어의 음성적 유사도를 구한다. KODEX 알고리즘은 초성 이응 제거, 중복 종성 제거, 초성 대표 자음화, 코드 치환, 연속 중복 코드 제거의 다섯 단계의 처리 과정으로 구성된다.

CKODEX 알고리즘[2]은 KODEX 알고리즘에 기반을 둔 것으로 첫음절의 모음 정보를 추가하고 Metaphone 알고리즘[6]의 개념을 도입하여 KODEX 보다 세분화한 규칙을 적용함으로써 한국어 외래어 음차표기의 유사도 비교 성능을 향상했다.

본 논문에서는 한국어 자모 혼동행렬을 기반으로 하여 외래어 표기의 유사도를 비교한다. Savitha 등[7]은 혼동행렬을 기반으로 한 음성 자료 검색 방법을 제안하였다. IBM의 음성 인식 시스템을 사용하여 US 영어의 음성 혼동행렬을 구성하고 Bayesian 확률 모델에 음성 혼동행렬 정보를 결합하여 음성 자료를 검색한다. 다음 장에서는 본 논문에서 제안하는 유사 외래어 표기 검출 기법을 자세히 소개한다.

3. 유사 외래어 표기 검출 기법

이 장에서는 초성, 중성, 종성에 대한 자모 혼동행렬의 구성방안과 이 혼동행렬을 이용하여 검색하는 방법, 그리고 검색 결과로 생성된 후보 외래어들의 유사도를 재검증하는 방법을 소개한다. 분석 데이터로는 우리말 배움터(http://urimal.cs.pusan.ac.kr)에서 서비스되고 있는 외래어-한글표기 상호변환기의 로그 데이터에서 1,662쌍의 유사 외래어 표기 데이터를 추출하여 사용하였다.

3.1 혼동행렬 구성

분석 데이터에서 길이가 같은 외래어 쌍을 자모 단위로 비교하여 초성, 중성, 종성에 대한 자모 혼동행렬을 구성하였다.

$$C_{ij} = P(i|j) \qquad (1)$$

수식(1)은 자모 i를 자모 j로 혼동한 확률을 나타낸다. 그림 1은 수식(1)을 이용해 구성한 초성 혼동행렬의 구조이다.

$$C = \begin{matrix} P(ㄱ|ㄱ) & P(ㄱ|ㄲ) & P(ㄱ|ㄴ) & \cdots & \cdots & P(ㄱ|ㅌ) & P(ㄱ|ㅍ) & P(ㄱ|ㅎ) \\ P(ㄲ|ㄱ) & P(ㄲ|ㄲ) & P(ㄲ|ㄴ) & \cdots & \cdots & P(ㄲ|ㅌ) & P(ㄲ|ㅍ) & P(ㄲ|ㅎ) \\ P(ㄴ|ㄱ) & P(ㄴ|ㄲ) & P(ㄴ|ㄴ) & \cdots & \cdots & P(ㄴ|ㅌ) & P(ㄴ|ㅍ) & P(ㄴ|ㅎ) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ P(ㅌ|ㄱ) & P(ㅌ|ㄲ) & P(ㅌ|ㄴ) & \cdots & \cdots & P(ㅌ|ㅌ) & P(ㅌ|ㅍ) & P(ㅌ|ㅎ) \\ P(ㅍ|ㄱ) & P(ㅍ|ㄲ) & P(ㅍ|ㄴ) & \cdots & \cdots & P(ㅍ|ㅌ) & P(ㅍ|ㅍ) & P(ㅍ|ㅎ) \\ P(ㅎ|ㄱ) & P(ㅎ|ㄲ) & P(ㅎ|ㄴ) & \cdots & \cdots & P(ㅎ|ㅌ) & P(ㅎ|ㅍ) & P(ㅎ|ㅎ) \end{matrix}$$

(그림 1) 초성 혼동행렬

표준 발음법, 외래어 표기법, 그림 1과 같은 자모 혼동행렬을 근거로 하여 혼동하기 쉬운 자모를 하나의 그룹으로 묶어 대표 자모를 만들었다. 표 1, 표 2, 표 3은 각각 초성, 중성, 종성의 대표 자모를 나타낸다.

<표 1> 초성 대표 자모

대표 초성	초성
ㄱ	ㄱ
ㄴ	ㄴ
ㄷ	ㄷ
ㄹ	ㄹ
ㅁ	ㅁ
ㅂ	ㅂ, ㅃ
ㅅ	ㅅ, ㅆ
ㅇ	ㅇ
ㅈ	ㅈ
ㅊ	ㅊ, ㅉ
ㅋ	ㅋ, ㄲ
ㅌ	ㅌ, ㄸ
ㅍ	ㅍ
ㅎ	ㅎ

<표 2> 중성 대표 자모

대표 중성	중성
ㅏ	ㅏ, ㅓ, ㅗ
ㅐ	ㅐ, ㅔ
ㅡ	ㅡ, ㅜ, ㅠ
ㅣ	ㅣ, ㅟ, ㅢ
ㅑ	ㅑ, ㅕ, ㅛ
ㅝ	ㅝ, ㅘ
ㅞ	ㅞ, ㅖ, ㅒ, ㅙ, ㅚ

<표 3> 종성 대표 자모

대표 종성	종성
ㄱ	ㄱ, ㄳ, ㄲ, ㅋ, ㄺ
ㄴ	ㄴ, ㄵ, ㄶ
ㄹ	ㄹ, ㄻ, ㄽ, ㄾ, ㄿ
ㅁ	ㅁ, ㄻ
ㅂ	ㅂ, ㅄ, ㅍ, ㅀ
ㅅ	ㅅ, ㅆ, ㄷ, ㅈ, ㅊ, ㅌ, ㅎ
ㅇ	ㅇ
- (종성 없음)	- (종성 없음)

그림 1과 마찬가지로 수식(1)을 이용하여 대표 자모 혼동행렬을 구성한다. 그림 2는 대표 모음 혼동행렬 구조이다.

$$C = \begin{matrix} P(ㅏ|ㅏ) & P(ㅏ|ㅑ) & P(ㅏ|ㅐ) & P(ㅏ|ㅓ) & P(ㅏ|ㅔ) & P(ㅏ|ㅡ) & P(ㅏ|ㅣ) \\ P(ㅑ|ㅏ) & P(ㅑ|ㅑ) & P(ㅑ|ㅐ) & P(ㅑ|ㅓ) & P(ㅑ|ㅔ) & P(ㅑ|ㅡ) & P(ㅑ|ㅣ) \\ P(ㅐ|ㅏ) & P(ㅐ|ㅑ) & P(ㅐ|ㅐ) & P(ㅐ|ㅓ) & P(ㅐ|ㅔ) & P(ㅐ|ㅡ) & P(ㅐ|ㅣ) \\ P(ㅓ|ㅏ) & P(ㅓ|ㅑ) & P(ㅓ|ㅐ) & P(ㅓ|ㅓ) & P(ㅓ|ㅔ) & P(ㅓ|ㅡ) & P(ㅓ|ㅣ) \\ P(ㅔ|ㅏ) & P(ㅔ|ㅑ) & P(ㅔ|ㅐ) & P(ㅔ|ㅓ) & P(ㅔ|ㅔ) & P(ㅔ|ㅡ) & P(ㅔ|ㅣ) \\ P(ㅡ|ㅏ) & P(ㅡ|ㅑ) & P(ㅡ|ㅐ) & P(ㅡ|ㅓ) & P(ㅡ|ㅔ) & P(ㅡ|ㅡ) & P(ㅡ|ㅣ) \\ P(ㅣ|ㅏ) & P(ㅣ|ㅑ) & P(ㅣ|ㅐ) & P(ㅣ|ㅓ) & P(ㅣ|ㅔ) & P(ㅣ|ㅡ) & P(ㅣ|ㅣ) \end{matrix}$$

(그림 2) 대표 모음 혼동행렬

3.2 검색

검색 방법으로 N-gram 방법을 이용한다. N-gram 방법은 간단하게 적용할 수 있는 문자열 비교 방법이다. 기본적인 아이디어는 두 문자열의 N-gram들이 얼마나 일치하는지 비교하는 것이다. 두 문자열 S_1과 S_2사이의 유사도는 다음의 식으로 계산된다.

$$sim(S_1,S_2)=\frac{|N_1\cap N_2|}{|N_1\cup N_2|} \quad (2)$$

수식(2)에서 N_1, N_2는 각각 S_1, S_2의 N-gram들의 집합을 나타낸다[1]. 본 논문에서는 두 문자열의 유사도를 다음의 식으로 계산한다.

$$sim(S_1,S_2)=\frac{\alpha|N_1\cap N_2|+\beta|N_1^{'}\cap N_2|}{\alpha|N_1\cup N_2|} \quad (3)$$
$$(\text{단, } N_1\cap N_1^{'}=\varnothing,\ \alpha\geq\beta)$$

수식(3)의 α, β는 상수이며, $N_1^{'}$는 S_1을 대표 자모 혼동행렬로 확장한 N-gram들의 집합이다. 이때, 자모의 확장은 다음의 식을 이용한다. 다음의 식은 실험적으로 가장 적합한 조건을 찾아낸 것이다.

$$E_{ij}=\begin{cases} 1 & if\ (C_{jj}>0.94 \text{ and } C_{ij}>0.006) \\ & \text{or } (0.94\geq C_{jj}>0.8 \text{ and } C_{ij}>0.033) \\ & \text{or } (C_{ij}>0.09) \\ 0 & otherwise \end{cases} \quad (4)$$

수식(4)에서 E_{ij}는 자모 i를 자모 j로 혼동한 수치로 E_{ij}가 1이면 자모 j를 자모 i로 확장한다. 본 논문에서는 자모 단위의 3-gram을 사용하여 검색한다. 표 4는 대표 자모 혼동행렬을 이용하여 확장된 “팬더”의 3-gram들을 보여준다.

<표 4> 확장된 “팬더”의 3-gram

	3-gram
N_1	$ㅍㅐ$1 ㅍㅐㄴ2 ㅐㄴㄷ3 ㄴㄷㅏ4 ㄷㅏ-5 ㅏ-$6
$N_1^{'}$	$ㅂㅐ$1 ㅂㅐㄴ2 $ㅍㅏ1 ㅍㅏㄴ2 ㅏㄴㄷ3 ㅍㅐㅁ2 ㅐㅁㄷ3 ㅁㄷㅏ4 ㅍㅐㅂ2 ㅐㅂㄷ3 ㅂㄷㅏ4 ㅍㅐ-2 ㅐ-ㄷ3 -ㄷㅏ4 ㅐㄴㅌ3 ㄴㅌㅏ4 ㅌㅏ-5 ㄴㄷㅐ4 ㄷㅐ-5 ㅐ-$6 ㄷㅏㄹ5 ㅏㄹ$6

<표 5> “판다”의 3-gram

	3-gram
N_2	$ㅍㅏ1 ㅍㅏㄴ2 ㅏㄴㄷ3 ㄴㄷㅏ4 ㄷㅏ-5 ㅏ-$6

표 4와 표 5에서 아래 첨자로 쓰인 숫자는 3-gram의 위치를 나타낸다. 수식(3)의 α가 5, β가 4라고 할 때, “팬더”과 “판다”의 유사도는 3/5이다.

“윈도우” - “윈도”, “로보트” - “로봇”, “네트웍” - “네트워크”와 같이 길이가 다른 외래어의 검색을 위해 학습 데이터에서 길이가 다른 외래어 쌍을 비교하여 표 6과 같은 규칙을 추가하였다.

<표 6> ‘ㅡ’, ‘ㅇ’ 제거 규칙

	규칙 (‘*’는 종성, ‘-’는 종성 없음)
‘ㅡ’ 제거	-*ㅋ -> ㄱ*, -*트 -> ㅌ*, ㅅ*트 -> ㅌ* -*르 -> ㄹ*, -*프 -> ㅂ*
‘ㅇ’ 제거	ㅗ우 -> ㅗ, ㅗ오 -> ㅗ, ㅖ이 -> ㅖ ㅠ우 -> ㅠ, ㅣ유 -> ㅠ, ㅖ인 -> ㅖㄴ* ㅣ엔 -> ㅖㄴ*, ㅗ올 -> ㅗㄹ*

3.3 유사도 재검증

대표 자모를 이용하여 유사도를 계산하면 유사하지 않은 단어가 그룹화되는 문제점이 발생한다. 이를 해결하고자, 그림 3의 거리 계산 알고리즘을 이용하여 검색 결과로 생성된 후보 외래어들의 유사도를 재검증한다.

예를 들어, 수식(3)의 α가 5, β가 4라고 할 때, “개스”와 “가스”의 유사도는 3/5이고 “코스”와 “가스”의 유사도는 7/10이므로 “개스”와 “가스”의 유사도보다 “코스”와 “가스”의 유사도가 더 높다. 그러나 그림 3의 알고리즘에 따라 유사도를 재검증하면 “개스”와 “가스”의 거리는 1이 되고, “코스”와 “가스”의 거리는 4가 되어, “개스”와 “가스”가 더 유사함을 알 수 있다.

```
Distance(S1, S2)
1  d = 0
2  while (EndOfString(S1) and EndOfString(S2))
3  do i ← nextchar(S1), j ← nextchar(S2)
4     if (i = j)
5        then continue
6     elseif (Cjj > 0.94 and Cij > 0.006)
          or (0.94 ≥ Cjj > 0.8 and Cij > 0.033) or (Cij > 0.09)
7        then d := d+1
8     elseif (Cjj > 0.94 and 0.006 ≥ Cij > 0.004)
          or (0.94 ≥ Cjj > 0.8 and 0.033 ≥ Cij > 0.022) or (0.09 ≥ Cij > 0.06)
9        then d := d+2
10       else d := d+3
11 return d
```

(그림 3) 자모 혼동행렬을 이용한 거리 계산

4. 실험 및 성능평가

제안 기법의 성능 평가를 위해 표 7에 나타난 세 가지 종류의 외래어 목록을 실험 데이터로 사용하였다. 실험 데이터는 총 834개의 외래어 표기와 355개의 유사 외래어 표기 쌍으로 구성된다.

<표 7> 실험 데이터

1	IT 외래어 표기 용례집	한국정보통신기자협회(2002. 12)
2	깁고 더한 우리말의 바른 표기와 표준어 지도자료(2661어) 중 VI. 외래어의 새 표기법	경상남도 교육청
3	이런말실수 저런글실수 중 부록 2. 기본외래어표기	문화관광부

본 논문에서는 N-gram 방법, CKODEX 알고리즘[2]과 제안한 기법의 성능을 비교 실험하였다. 평가 방법으로는 F-measure[5] 값을 사용한다. 정확도(Precision)와 재현율(Recall)은 서로 반비례 관계에 있으므로 본 논문에서는 정확도와 재현율의 조화 평균을 이용하는 F-measure 값을 이용한다. N-gram 방법, CKODEX 알고리즘과 본 논문에서 제안한 기법의 성능은 표 8과 같다. s와 d는 문자열 유사도 계산에서 사용된 임계값을 나타내고 α와 β는 수식(3)의 상수이다. 각 임계값은 F-measure 값을 최대로 하는 값을 선택하였다.

<표 8> N-gram과 제안 기법 성능 비교

	N-gram ($s=0.5$)	CKODEX 알고리즘[2]	제안 기법 ($s=0.33, d=\|S_1\|, \alpha=5, \beta=4$)
정확도	0.894	0.865	0.914
재현율	0.553	0.689	0.786
F-measure	0.683	0.767	0.854

N-gram 방법, CKODEX 알고리즘과 본 논문에서 제안한 기법을 비교해 본 결과 본 논문에서 제안한 기법이 N-gram 방법과 CKODEX 알고리즘보다 정확도, 재현율, F-measure 값에서 모두 높은 성능을 보였다.

5. 결론 및 향후 연구

본 논문에서는 한국어 자모 혼동행렬을 기반으로 한 유사 외래어 표기 검출 기법을 제안하였다. 혼동하기 쉬운 자모를 하나의 그룹으로 묶어 대표 자모를 구성하여 유사 외래어 표기의 재현율을 높이고, 유사도 재검증 기법을 통해 정확도를 높일 수 있었다. 따라서 본 논문에서 제안한 기법으로 검색 질의어 확장, 대역어 제시 등 정보검색의 여러 분야에 응용할 수 있다는 데에 본 연구의 의의가 있다.

향후 연구해야 할 과제로는 유사 외래어 표기 검출 성능 향상을 위한 새로운 알고리즘 및 규칙에 대한 연구 및 속도 향상을 위한 검색 방법에 대한 연구가 필요할 것이다.

참고문헌

[1] 강병주, 이재성, 최기선, “외국어 음차 표기의 음성적 유사도 비교 알고리즘”, 정보과학회 논문지(B), 제26권 제10호, pp. 1237-1246, 1999.

[2] 고숙현, 이재성, “문맥을 고려한 유사 외래어 검출 알고리즘의 성능 향상”, 한국정보과학회 언어공학연구회 학술발표 논문집, pp. 114-121, 2007.

[3] 김태일, “최대 엔트로피 모델을 이용한 다국어 정보검색에서의 영-한 음차 표기 모델”, 서강대학교 컴퓨터학과 공학석사 학위 논문, 1999.

[4] 안재철, 오일석, “문자 인식을 이용한 한글 문서 검색”, 한국정보과학회 춘계학술발표논문집, 제28권 제1호, pp. 544-546, 2001.

[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, Introduction to Infomation Retrieval, Cambridge University Press, 2008.

[6] Lawrence Phillips, Hanging on the Metaphone, Computer Language, vol. 7, no. 12, pp. 39-43, 1990.

[7] Savitha Srinivasan, and Dragutin Petkovic, “Phonetic Confusion Matrix Based Spoken Document Retrieval”, Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development In information Retrieval, pp. 81-87, 2000.

[8] Ulrich Pfeifer, Thomas Poersch, and Norbert Fuhr, “Retrieval effectiveness of proper name search methods”, Information Processing and Management, vol. 32, no. 6, pp. 667-679, 1996.