사용자 입력오류를 고려한 사전 검색 방법

정형일○, 선충녕, 서정연*
서강대학교 컴퓨터공학과, *서강대학교 컴퓨터공학과/바이오융합기술협동과정
hijeong@sogang.ac.kr, wilowisp@sogang.ac.kr, seojy@sogang.ac.kr

A Method of Dictionary Search for Typographical Error

Hyoung-Il Jeong○, Choong-Nyoung Seon, Jung-Yun Seo*
Department of Computer Science and Engineering, Sogang University
*Department of Computer Science and Engineering, and Interdisciplinary Program of Integrated Biotechnology, Sogang University

요 약

디지털 기기들의 발전은 사전 검색 수요의 증가와 함께 강건한 검색 기법의 필요성도 증가시키고 있다. 기존의 사전 검색 기법들은 사용자의 입력 오류를 고려하지 않고, 검색 최적화만을 위해 설계되었다. 본 논문에서는 언어 모델 키워드와 자소 범주 키워드를 이용하여 오타에 강건한 사전 검색 방법을 제안한다. 제안된 방법은 오류가 포함된 사용자의 입력 단어에 대하여 활용 가능한 수준의 높은 성능과 검색 속도를 보여주었다.

주제어: 사전 검색, Extended Boolean 검색, N-gram 언어모델, Soundex 알고리즘

1. 서론

최근의 개인용 컴퓨터, 자동차용 네비게이션, 개인휴대용 스마트폰 등의 많은 기기의 발달로 자동화된 디지털 사전의 사용이 점차 빈번해지고 있다. 이러한 디지털 사전들은 대부분 빠른 검색이 가능한 TRIE[1]나 Hash등을 이용하여 구축된다. 그러나 TRIE와 Hash 기반 사전은 사용자의 실제 사용에서 빈번하게 발생하는 입력 오류에 대하여 사전 검색이 불가능한 문제가 있었다. 철자 오류 검색을 위한 방법으로 Levenshtein Distance를 이용한 방법이 있었으나[2], 이 방법은 문자열이 길어지거나 비교 대상이 많아질수록 계산 비용이 증가하는 문제점이 있었다[3]. 한국어를 위한 철자 오류 검색 방법으로 K-Phone이 제안되었다[3]. 이 알고리즘은 어휘 범주화 방법이므로 정답 후보를 상위에 제시하지 못하기 때문에 사전 검색의 방법으로는 사용이 불가능하다. 이러한 문제를 해결하기 위해, 본 논문에서는 사용자 입력 오류에 강건한 사전 검색 방법을 제안한다.

2. 제안 방법

제안하는 시스템의 전체적인 구성은 다음의 그림 1과 같다. 제안 시스템은 일반적인 역파일 색인/검색 시스템의 형태를 갖는데 크게 세 부분으로 분류할 수 있다. 이들은 사용자 입력 단어 또는 검색 대상 후보 단어를 색인단위로 변환시키는 언어 분석기, 후보 대상 단어들을 각 색인 단위에 대한 역파일로 구성해 색인 파일을 생성해 주는 색인기, 그리고 입력 단어를 색인파일에서 검색해 주는 검색기이다. 이를 위하여 Extended Boolean Model[4] 기반의 색인/검색기를 구축하였다.

제안 시스템의 언어 분석기는 크게 두 부류의 결과를 출력하는데, 이들은 언어모델 키워드와 자소범주 키워드이다. 언어모델 키워드는 동일한 음절이나 자소의 포함 정도에 따른 가중치를 부여하기 위한 키워드이며, 자소범주 키워드는 발음 등의 조건의 유사성에 따른 가중치를 부여하기 위한 키워드이다. 이 키워드들을 역파일 색인 및 검색의 기본단위로 한다. 언어모델 키워드는 2.1절에, 자소범주 키워드는 2.2절에서 설명한다.

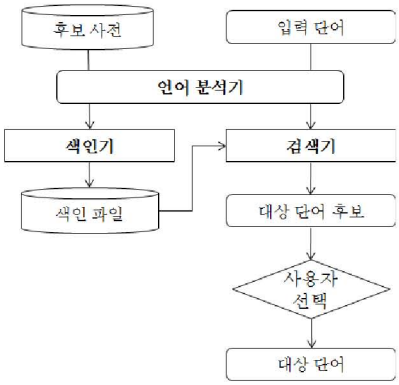

그림 1. 제안 시스템 구조

2.1. 언어모델 키워드

언어모델 키워드는 키워드 단위 역파일 색인/검색 시스템의 필수요소인 N-gram 언어모델이다. 우리는 사용자 입력 오류에 적합한 언어모델을 구성하기 위해 잘 알려진 음절 N-gram, 자소 N-gram 이외에 추가로 음절 uni-gram과 자소 bi-gram의 특성을 결합한 음절+좌자소,

음절+우자소, 음절+좌우자소 언어모델이라는 새로운 방법의 언어모델을 제안하였다.

아래의 그림 2에서는 대표적인 각 언어모델에 따라 입력 단어에 대하여 다양한 형태의 키워드를 추출하는 것을 확인 할 수 있다. 그림2의 언어모델 키워드 추출 예에서 볼 수 있듯이, 서로 다른 문자들이지만 유사한 키워드들이 추출됨을 알 수 있다. 그림2의 $는 단어의 시작과 끝을 의미하는 기호이며, (nil)은 종성자음이 없는 경우를 의미하는 기호이다.

```
예1) 음절 bi-gram 언어모델 키워드
(신수동)=($신), (신수), (수동), (동$)
(신ㅅ동)=($신), (신ㅅ), (ㅅ동), (동$)
(신ㄴ수동)=($신), (신ㄴ), (ㄴ수), (수동), (동$)

예2) 자소 bi-gram 언어모델 키워드
(신수동)=($ㅅ), (ㅅㅣ), (ㅣㄴ), (ㄴㅅ), (ㅅㅜ), (ㅜㄷ), (ㄷㅗ), (ㅗㅇ), (ㅇ$)
(신ㅅ동)=($ㅅ), (ㅅㅣ), (ㅣㄴ), (ㄴㅅ), (ㅅㄷ), (ㄷㅗ), (ㅗㅇ), (ㅇ$)
(신ㄴ수동)=($ㅅ), (ㅅㅣ), (ㅣㄴ), (ㄴㄴ), (ㄴㅅ), (ㅅㅜ), (ㅜㄷ), (ㄷㅗ), (ㅗㅇ), (ㅇ$)

예3) 음절+좌우자소 언어모델 키워드
(신수동)=($신ㅅ), (ㄴ수ㄷ), ((nil)동$)
(신ㅅ동)=($신ㅅ), (ㄴㅅㄷ), ((nil)동$)
(신ㄴ수동)=($신ㄴ), (ㄴㄴㅅ), (ㄴ수ㄷ), ((nil)동$)
```

그림 2. 언어모델 키워드 추출 예

각 언어모델 선택에 따른 검색 성능은 3.2절에서 다룬다.

2.2. 자소범주 키워드

본 절에서 다룰 자소범주 키워드는 기존의 Soundex 알고리즘[5,6]을 사용자 입력 오류 문제에 적합하도록 개량하여 적용한 것이다. 언어모델 키워드는 사용자의 단순 오타를 고려한 키워드인 반면에, 자소범주 키워드는 음가대로 읽는 등의 맞춤법 오류를 고려한 키워드이다. Soundex를 한글 및 한국어에 적용한 Kodex[7], K-Phone 등의 알고리즘이 존재하지만, 이들은 단어에 포함된 모음을 무시하는 문제 등으로 제안하는 시스템에 바로 적용하는 것은 적절치 않다.

제안하는 자소범주 키워드는 두 가지이다. 하나는 자소 N-gram을 자소범주 할당테이블을 이용하여 자소범주 코드로 단순 변환하여 키워드로 사용하는 것이다. 다른 하나는 단어에 포함된 모든 자소를 변환하여 초중종성 각 한 개씩 가장 높은 빈도의 자소범주코드를 조합하여 키워드로 사용하는 것이다. 만약 자소범주의 출현 빈도가 같다면, 단어의 초두에 가까운 코드를 선택한다. 제안하는 자소범주 할당테이블과 변환 예는 다음의 그림 3과 그림 4에 나타내었다.

그림 4의 자소범주 키워드 변환 예에서 보듯이 “근린공원”과 “글린공원”은 서로 다른 문자열임에도 불구하고, 유사한 발음을 같기 때문에 동일한 자소범주 키워드로 변환됨을 볼 수 있다.

각 자소범주 키워드 변환방법 선택에 따른 성능은 3.3절에서 다룬다.

초성(H), 종성(T)	중성(M)
1: ㄱ ㅋ ㄲ 2: ㄴ ㄹ ㅁ ㅇ ㅎ 3: ㄷ ㅌ ㄸ 4: ㅂ ㅍ ㅃ 5: ㅅ ㅈ ㅉ ㅊ ㅆ	1: ㅏ ㅑ ㅐ ㅒ ㅣ 2: ㅜ ㅠ ㅡ 3: ㅗ ㅛ 4: ㅓ ㅕ ㅔ ㅖ 5: ㅝ ㅟ ㅞ ㅘ ㅙ ㅢ ㅚ

그림 3. 자소범주 할당 테이블

```
예1) 자소범주 bi-gram 변환
(근린공원) → ($ㄱ), (ㄱㅡ), (ㅡㄴ), (ㄴㄹ), (ㄹㅣ), (ㅣㄴ), ...
          → $H1,  H1M2, M2T2,  T2H2, H2M1,  M1T2, ...
(글린공원) → ($ㄱ), (ㄱㅡ), (ㅡㄹ), (ㄹㄹ), (ㄹㅣ), (ㅣㄴ), ...
          → $H1,  H1M2, M2T2,  T2H2, H2M1,  M1T2, ...

예2) 자소범주 조합 변환
(근린공원) → [ H1M2T2 H2M1T2 H1M3T2 H2M5T2 ] → H1M2T2
(글린공원) → [ H1M2T2 H2M1T2 H1M3T2 H2M5T2 ] → H1M2T2
```

그림 4. 자소범주 키워드 변환 예

3. 실험 및 평가

3.1 실험 구성

사전 검색 성능을 평가하기 위하여 사용된 자료는 680k여 한국 내 지명 단어이다. 그 중 비관련 전공 대학생 20여명이 25k개의 지명 단어를 직접 타이핑하여, 실제 입력오류가 발생한 5k여 개 단어를 검색질의로 선정하였다.

시스템은 각 검색질의 단어로 검색을 하여 검색순위 상위 20개의 단어를 정답으로 제시한다. 성능을 평가하기 위한 척도는 $P@20$과 MRR을 사용하였다.

$P@20$은 식 (1)로 정의하였는데, 여기서 $|Q|$는 질의 단어의 수 즉 검색횟수를 의미하고, C은 시스템이 제시한 후보에 정답 단어가 포함된 횟수를 의미한다. $P@20$이 높다는 것은 후보 단어에 정답 단어가 포함될 가능성이 높음을 말한다. 예로 $P@20$이 0.5인 경우는 정답 단어가 후보 내에 포함될 가능성이 0.5임을 말한다.

$$P@20 = \frac{C}{|Q|} \tag{1}$$

MRR은 다음 식(2)로 정의하였는데, 여기서 $rank_i$는 i번째 질의 단어에 대하여, 시스템이 제시한 후보 단어에서 정답단어의 순위를 뜻한다. MRR이 높다는 것은 제시된 후보 단어에서 정답 단어의 순위가 높음을 말한다. 예로 MRR이 0.5인 경우는 정답 단어가 후보 단어 내에서 평균 2위의 순위로 제시됨을 뜻한다.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{2}$$

3.2. 언어모델 키워드별 성능 평가

언어모델 키워드 추출 방법 변화에 따른 성능 평가 결과는 다음의 표 1에서 볼 수 있다.

표 1. 기본키워드별 성능평가 결과

기본키워드	*P@20*	*MRR*	평균검색시간 (초/단어)
음절 uni-gram	.7481	.5421	0.053
음절 bi-gram	.6992	.4994	0.008
음절+우자소	.8510	.6738	0.009
음절+좌자소	.8063	.6015	0.014
음절+좌우자소	.8761	.6703	0.016
자소 bi-gram	.9541	.7951	0.423
자소 tri-gram	.9462	.7702	0.096

표 1의 결과에서 보듯이 가장 성능이 높은 언어모델 키워드는 자소 bi-gram이다. 그러나 자소 tri-gram에 비해 성능차가 그리 크지 않은 반면에, 후보 단어가 될 수 있는 단어가 매우 많아 검색시간이 과도하게 커지는 것을 확인하였다. 따라서 일반적인 환경에서는 자소 bi-gram이 가장 적합한 언어모델 키워드 추출이라고 할 수 있다. 그리고, 음절+자소 언어모델의 경우 성능은 자소 N-gram보다 다소 낮지만 검색시간은 매우 빠르기 때문에, 하드웨어의 성능이 좋지 않은 휴대용 기기에서는 적합한 방법이라고 할 수 있다.

3.3. 자소범주 키워드별 성능 평가

자소범주 키워드 변환방법에 따른 성능평가 결과는 다음의 표 2에 나타내었다. 언어모델 키워드 중 성능이 우수한 자소 tri-gram에 자소범주 키워드를 추가하였을 때의 성능변화를 측정하였다.

표 2. 확장키워드별 성능평가 결과

확장키워드	*P@20*	*MRR*	평균검색시간 (초/단어)
자소 tri-gram + 자소범주 bi-gram	.9460	.7701	1.963
자소 tri-gram + 자소범주 tri-gram	.9608	.7914	0.633
자소 tri-gram + 자소범주 조합	.9531	.7908	0.120

실험결과 자소범주 tri-gram을 사용하였을 때 가장 높은 성능을 보임을 확인하였다. 그러나 확장 N-gram 키워드는 자소범주 조합 키워드에 비해 평균검색시간이 과도하게 크기 때문에, 일반적인 환경에서는 자소범주 조합 키워드를 사용하는 것이 적절할 것이다.

4. 결론 및 향후계획

본 논문에서는 오류가 포함된 사용자 입력 문제를 고려한 사전 검색 방법에 대하여 연구하였다. 이러한 문제를 해결하기 위해 언어모델 기반의 키워드 추출방법들과 Soundex 유사 알고리즘을 이용한 자소범주 키워드 변환 방법들을 제시하고 성능을 평가하였다. 실험 결과 일반적인 환경에서는 자소 tri-gram을 이용한 키워드와 자소범주 조합 키워드를 사용하는 것이 추천된다. 제한적인 하드웨어 환경에서는 음절+자소 형태의 언어모델 키워드만으로 검색하는 것을 추천한다.

향후 성능과 속도를 고려한 새로운 검색 키워드들에 대한 연구와 단어길이 등에 대한 필터, 그리고 입력 단어와 후보 단어와의 유사도 측정을 통한 검색 성능 향상에 대한 연구를 진행할 예정이다.

* 이 논문은 한국연구재단의 중견연구자 프로그램의 (No. 2009-0086194) 지원으로 이루어진 결과의 일부입니다.

참고문헌

[1] R. De La Briandais, “File Searching using Varible Length Keys”, Proceedings of the Western Joint Computer Conference, pp. 295-298, 1959.

[2] F.J. Damerau, “A Technique for computer detection and correction of spelling errors”, Communications of the ACM, vol. 7, no. 3, pp. 171-176, 1964.

[3] 김효경, 한글 철자 검사를 위한 음성적 유사도 계산 알고리즘, 성균관대학교, 석사학위논문, 2006.

[4] G. Salton, F.A. Edward and W. Harry, “Extended Boolean Information Retrieval”, Communications of the ACM, vol. 26, no. 11, 1983.

[5] R.C. Russell and M.K. Odell, “(Unknown Title)”, U.S. Patent 1,261,167, 1918.

[6] R.C. Russell and M.K. Odell, “(Unknown Title)”, U.S. Patent 1,453,663, 1922.

[7] 강병주, 이재성, 최기선, “외국어 음차 표기의 음성적 유사도 비교 알고리즘”, 정보과학회 논문지(B), 제26권, 제10호, pp. 1237-1245, 1999.