모바일 기기의 입력 문자열 추천 및 오타수정 모델을 위한 주요 기술

이성욱*

*충주대학교

Related Works for an Input String Recommendation and Modification on Mobile Environment

Songwook Lee*

*Chungju National University

E-mail : leesw@cjnu.ac.kr

요 약

스마트폰과 모바일 인터넷의 활발한 보급으로 문자 메시지 작성, 정보검색, 소셜 네트워크 참여 등 다양한 분야에 모바일 기기를 활용하는 사용자가 증가하고 있다. 모바일 기기의 특성상 키패드는 비교적 작은 크기로 구성되어 있어, 사용자가 원하는 문장을 정확하고 신속하게 입력하는데 어려움이 있다. 본 연구에서는 모바일 기기에 적용하여 키패드 입력에 도움을 줄 수 있는 입력 문자열 추천 및 오타수정 기술을 살펴보고자 한다. 기존의 온라인 검색엔진의 검색어 추천 모델에 적용되는 주요 기술인 트라이(TRIE) 사전과 n-그램 언어모델을 이용한 관련 연구를 살펴본다.

ABSTRACT

Due to wide usage of smartphones and mobile internet, mobile devices are used in various fields such as sending SMS, participating SNS, retrieving information and the number of users taking advantage of them are growing. The keypads of a mobile device are relatively smaller than those of desktop computers. Thus, the user has a difficulty in input sentences quickly and correctly. In this study, we introduce some string recommendation and modification techniques which can be used for helping a user input in mobile devices quickly and correctly. We describe a TRIE dictionary and n-gram language model which are the main technologies of the keyword recommendation applied to the online search engines.

키워드

Input string recommendation, modification, TRIE dictionary, N-gram language model

Ⅰ. 서 론

스마트폰과 모바일 인터넷의 활발한 보급으로 SMS 작성, 정보검색, 소셜 네트워크 참여 등 다양한 분야에 모바일 기기를 활용하는 사용자가 증가하고 있다. 모바일 기기의 특성상 키패드는 비교적 작은 크기로 구성되어 있어, 사용자가 원하는 문장을 정확하고 신속하게 입력하는데 어려움이 있다. 본 연구에서는 모바일 기기에 적용하여 키패드 입력에 도움을 줄 수 있는 입력 문자열 추천 및 오타수정 모델을 위한 주요 기술 중 대표적으로 많이 사용되는 기술인 트라이(TRIE) 사전과 n-그램 언어모델을 살펴보고자 한다.

일반적으로 트라이 사전과 n-그램 언어모델은

대용량 원시 말뭉치로부터 구축할 수 있다. 그림 1 과 같이 구글 등의 인터넷 검색 서비스의 경우에 사용자 검색어 DB를 활용하여 실시간으로 사용자 검색어에 대한 추천 기능을 제공하고 있다.

그림 1. 구글 입력의 예

그러나 오프라인 환경에서 입력한 후에 온라인으로 전송하는 카카오톡, 트위터 등과 같은 어플리케이션에는 아직 문자열 추천 기능이 제공 되지 않고 있다. 본 논문에서는 이를 가능하게 하는 기술을 소개하고자 한다.

II. 트라이(TRIE) 사전

사용자 입력 문자열을 예측하여 검색어를 추천하는 기본적인 방법은 트라이 사전을 이용하는 방법이며 자동차의 네비게이션 단말기 등에 적용되어 목적지를 검색하는데 구현되어 있다.

트라이 사전은 트라이[1] 자료구조를 이용한 사전 구성 방법이며, 각 문자열의 알파벳을 비단말 노드로 구성함으로써 문자열의 알파벳 순서대로 트리를 탐색하면 단말 노드에 도달하여 해당 문자열을 탐색할 수 있는 자료구조이다. '안녕'과 '안나수이' 두 단어를 자소별로 분해하여 구성한 트라이 사전의 예는 그림 2와 같다.

트라이 사전을 이용한 검색어 예측 추천 방법은 현재 입력된 사용자 문자열을 따라 트라이 사전의 노드들을 탐색한 후 현 노드의 자식 노드들을 이용하여 추천하는 방법이다. 그림 2에서와 같이 사용자가 'ㅇㄴ'라고 입력을 했다고 가정하자. 검색어 추천 모델은 트라이 사전에서 주어진 입력 'ㅇㄴ'까지 탐색한 후, 자손 노드들에 존재하는 모든 단말노드를 탐색하여 '안녕'과 '안나수이' 등

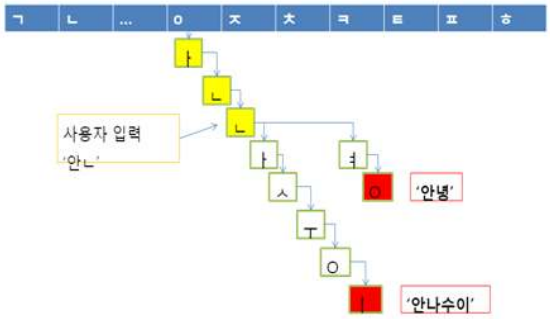

그림 2. 트라이 사전을 이용한 추천

으로 추천하게 된다.

III. N-그램 언어 모델

n-그램 언어 모델은 단어 인식 및 예측에서 사용되는 대표적인 통계적인 방법이며, 한국어의 자동 띄어쓰기 모델[2]과 휴대폰의 입력 문자열 예측[3]에 사용되기도 하였다.

m개의 단어 W로 구성된 문장이 발생할 언어 모델의 확률식은 다음 수식(1)과 같다[4]. 이를 n-그램으로 추정한 확률식은 수식(2)와 같으며 m개의 음절(또는 자소)로 이뤄진 문자열의 추정도 W를 단어 대신 음절로 간주하면 동일한 수식이 적용된다.

$$P(W_1,\ldots,W_m)=\prod_{i=1}^{m}P(W_i|W_i,\ldots,W_{i-1}) \quad (1)$$

$$\approx\prod_{i=1}^{m}P(W_i|W_{i-(n-1)},\ldots,W_{i-1}) \quad (2)$$

만약 '$W_1,\cdots,W_{k-1}$'가 이미 관찰되었다면, k번째 단어 W_k를 예측하는 수식은 다음 수식(3)과 같다.

$$W_k=\arg max_{w_k}P(W_1,\ldots,W_k) \quad (3)$$

$$\approx\arg max_{w_k}\prod_{i=1}^{k}P(W_k|W_{i-(n-1)},\ldots,W_{i-1})$$

수식 (3)에서 argmax만 계산하면 되기 때문에 이미 관찰된 '$W_1,\cdots,W_{k-1}$' 열의 확률곱은 W_k를 결정하는데 영향을 끼치지 않는다. 따라서 수식(3)은 다음 수식(4)와 같이 간소화 되며 수식(4)는 MLE(Maximum Likelihood Estimation) 방식으로 확률값을 계산할 수 있다.

$$W_k\approx\arg max_{w_k}P(W_k|W_{i-(n-1)},\ldots,W_{i-1}) \quad (4)$$

그림 2의 예제와 같이 'ㅇㄴ'까지 입력되었고 다음 자소를 예측하려고 한다고 가정하자. 자소 트라이그램(tri-gram)을 사용하였을 경우, P(ㅏ|ㄴ

ㄴ)과 P(ㅕ|ㄴㄴ)의 확률을 비교하여 'ㅏ'와 'ㅕ' 두 자소 중 큰 확률을 갖는 자소를 예측하게 된다.

Ⅳ. 적용 방안

4.1 추천 후보 문자열의 순위화

가장 단순한 순위화 방법은 자동차 네비게이션 단말기 등에서 사용되는 방법으로 트라이사전에 탐색된 순서에 따라 그대로 나열하는 방법이다. 트라이사전의 탐색 특성에 따라 각 문자열은 정렬된 형태로 나열된다. 그림 3은 구글과 네이버의 검색어 추천 예를 보여주는데, 두 업체의 검색어 추천 순위는 서로 다르며, 단순히 정렬된 형태는 아닌 것을 알 수 있다. 검색 시점에 따라 추천 순위가 조금씩 바뀌는 것으로 볼 때, 온라인에서 유입되는 검색어의 빈도 등 여러 가지 요소가 결합되어 순위화를 하고 있음을 알 수 있다.

그림 3. 구글과 네이버의 검색어 추천 비교

온라인 검색어 추천과 같이 각 단어의 발생 빈도수를 이용할 수 있다면 순위화에 큰 도움이 될 것이다. 또한 어절 빈도수와 음절 n-그램 확률을 이용하여 순위화를 할 수도 있다. 음절수가 다르면 각 어절의 n-그램 확률을 비교하기 어렵기 때문에 후보 문자열들 중 최소길이를 기준으로 음절 n-그램 확률을 계산하면 된다. 그림2의 예제인 '안녕'과 '안나수이'의 경우, 음절 최소길이는 2음절이며 각 후보 문자열 중 2음절로 된 부분문자열 '안녕'과 '안나'에 대한 n-그램 확률을 각각 계산하여 비교할 수 있다.

4.2. 자판배열을 고려한 오타 수정

모바일 기기는 작은 크기의 키패드로 인해 오타가 나기 쉽다. 추천 문자열이 사전에 존재하지 않을 때는 사용자 입력에서 오타가 발생한 것으로 간주하고, 최종 입력된 자소가 위치한 자판 배열의 이웃 자소들로 탐색 영역을 확장하여 올바른 문자열을 추천할 수 있다. 예를 들어, 사용자가 '안년'이라고 입력했을 때, 사전에는 '안년'으로 시작하는 문자열이 존재하지 않았다. 이 때, 그림 4와 같이 마지막으로 입력된 'ㄴ'과 이웃한 자소들인 'ㅈ', 'ㄷ', 'ㅁ', 'ㅇ', 'ㅋ', 'ㅌ' 등의 자소 중 트라이 사전에 존재하는 자소들로 사전 탐색을 확장하면 된다. 그 후, 탐색된 문자열을 순위

그림 4. 쿼티 자판 배열에서의 이웃 자소

화하고 적절한 자소를 선택함으로써 오타를 수정하고 문자열을 추천한다. 이 경우에, 만약'ㅇ'이 선택되면 '안년'->'안녕'으로 입력 문자열을 추천할 수 있게 된다.

그 외, 일반적으로 온라인 환경에서의 검색어 추천에서 오타 수정은 편집 거리(Edit Distance)[5]를 이용할 수 있다. 편집 거리 방식은 오타를 포함한 문자열을 올바른 문자열로 편집하는 데 필요한 편집명령(삽입, 삭제, 치환)을 최소화 하는 문자열을 선택하는 방법이다.

Ⅴ. 결론

우리는 모바일 기기의 입력기에 적용할 수 있는 문자열 추천 및 오타수정 기술들을 살펴보았다. 대표적인 사전 구조인 트라이를 이용한 입력 문자열 추천 방법과 통계적 방법인 N그램 언어 모델을 살펴보고 이를 이용한 문자열 추천 방법을 살펴보았다. 검색어 추천 기술과 모바일 기기의 입력 문자열 추천은 유사한 문제이며 문제 해결에 이용할 수 있는 자원의 차이에 따라 구현 방법에 있어 차이를 보인다. 본 연구에서 소개한 방법을 이용한다면 스마트폰의 각종 응용 프로그램에서 이용할 수 있는 입력기의 구현에 이용할 수 있을 것이다.

참고문헌

[1] Edward Fredkin, "Trie Memory". Communications of the ACM 3 (9): 490 499, 1960.

[2] 최성자, 강미영, 허희근, 권혁철, "음절 N-Gram과 어절 통계 정보를 이용한 한국어 띄어쓰기 시스템", 제 15회 한글 및 한국어 정보처리 학술대회, pp. 47-53, 2003

[3] M. D. Dunlop, A. Crossan, "Predictive text entry methods for mobile phones", Personaland Ubiquitous Computing, Vol. 4, No. 2-3, pp. 134-143, 2000

[4] Christopher D. Manning, Hinrich Sch□tze, Foundations of Statistical Natural Language Processing, MIT Press: 1999.

[5] Levenshtein VI, "Binary codes capable of correcting deletions, insertions, and reversals". Soviet Physics Doklady 10: 707 10, 1966.