트위터에서의 사례 기반 이벤트 지명 검출 기법

하현수, 황병연
가톨릭대학교 컴퓨터공학과
{hss0924, byhwang}@catholic.ac.kr

A Method for Detecting Event-location based on Example in Tweet

HyunSoo Ha, Byung-Yeon Hwang
Dept. of Computer Science and Engineering, The Catholic University of Korea

요 약

본 논문에서는 트위터 내용을 통해 이벤트를 탐지하는 시스템에서 지명 검출 정확도를 개선하는 방법을 제안한다. SNS를 이용한 개인 정보 유출 사례들이 늘어감에 따라 자신의 위치 정보를 공개하기 꺼려하기 때문에 이벤트가 발생한 지역을 검출하기 위해서는 텍스트 내용을 직접 분석해야한다. 그러나 오타나 줄임말, 동형이의어의 사용으로 정확한 지명 검출에 어려움이 발생하였다. 따라서 정확도를 향상시키기 위해 본 논문에서는 두 가지 지명 검출 기법을 제안한다. 지명 단어에서 발생되는 노이즈를 제거하는 지명 노이즈 제거 기법과 랜드 마크를 이용하여 지명 단어를 확정하는 지명 확정 기법이다. 실험 결과 기존 시스템의 정확도 49%에서 지명 노이즈 제거기법은 56%, 지명 확정 기법은 73%로 각각 향상되었다.

1. 서론

최근 SNS를 통한 빠른 정보 확산의 영향력은 확대되고 있다. SNS 중에 트위터는 최대 140자의 단문 텍스트로 작성하기 때문에 정보의 빠른 확산이 가능하다. 또한 팔로잉-팔로워 구조로 이루어져 있어서 개방적인 네트워크를 형성하고 있다. 이와 같은 특성을 활용해 트위터 이용자들을 각각의 센서로 판단하였고, 이용자가 작성하는 트윗 내용을 분석하여 이벤트를 탐지하는 TRED 시스템이 제안되었다[1].

그러나 TRED 시스템은 이벤트가 발생된 지역을 검출하는 과정에서 낮은 정확도를 보이는 문제점이 발견되었다. SNS 이용자들이 게시글을 작성할 때 맞춤법과 띄어쓰기를 제대로 지키지 않기 때문이다. 또한 동형이의어와 줄임말을 실제 지명이 아닌 단어들을 지명으로 판단하여 검출하게 되었고, 결국 이벤트 탐지에서도 오류가 발생했다. 따라서 본 논문에서는 이벤트를 탐지하는 TRED 시스템의 지명 검출 과정에 추가적인 알고리즘을 제안한다. 두 가지 지명 검출 기법을 제시하여 각각의 측면에서 결과를 비교해보았다. 두 가지 기법을 적용하였을 때, 기존 방식보다 이벤트 탐지 범위는 감소하였으나 정확도 향상에 기여하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 설명한다. 3장에서는 지명 검출 기법에 관해서 살펴본다. 이후 4장에서 실험 데이터를 통해 기존 시스템과의 비교되는 실험결과를 보이고 5장에서 결론과 향후 계획을 설명한다.

2. 관련연구

[1] 에서 제안하는 TRED 시스템은 트위터 사용자를 이벤트 탐지의 센서로 이용한 실시간 이벤트 탐지시스템이다. 실시간으로 트윗을 수집하여 정제하는 과정을 거친다. TF, VT, DA 수식을 이용하여 평소보다 자주 언급되는 지명에서 이벤트가 발생했다고 판단하는 알고리즘을 적용했다. 그러나 지명에 관한 노이즈를 제거하는 과정이 없어 이벤트 탐지 정확도가 낮다.

[2]와 [3]은 사전의 뜻풀이 말에서 추출한 통계적 의미정보를 이용해 동형이의어 중의성을 해결하는 시스템을 제안하였다. 동형이의어를 포함한 사전의 뜻풀이 말에서 정확한 의미정보를 추출한다. 이후 용언과 체언이 같이 사용되는 경우들을 파악하여 통계적 방법을 이용해 동형이의어의 정확한 의미를 분류한다.

[4]에서는 트위터 내용에서 지명 검출 정확도를 개선하는 방법을 제안하였다. 그러나 지명 노이즈 필터링 처리를 할 수 있는 데이터양과 일반화 시킬 수 있는 규칙에 한계가 있어 정확도 향상 정도가 미미하다.

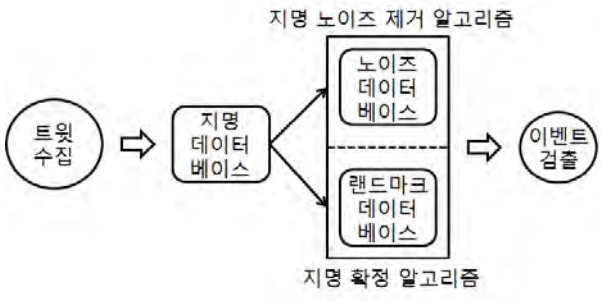
3. 지명 노이즈제거 기법

3.1 이벤트 탐지 시스템

본 논문에서 제안하는 기법을 적용할 이벤트 탐지 시스템의 구조는 (그림 1)과 같다. 이벤트 탐지 시스템은 트윗

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2011-0009407).

수집, 트윗 분석, 이벤트 탐지의 세 단계로 구성된다. 우선 트윗 수집 단계에서는 트위터에서 무료로 제공되는 API를 이용해 트윗을 수집한다[5]. 트윗을 분석하기 위해 수집된 트윗을 루씬 형태소 분석기를 통해 어절 단위로 나눈다. 나누어진 어절 단위의 트윗 내용에서 키워드를 추출하고, 키워드 중에서 이벤트가 발생한 지명을 검출한다. 지명 검출 기법은 중간 단계인 트윗 분석의 지명 검출 구간에 적용된다. 우선 지명 데이터베이스를 거쳐 1차적으로 지명을 탐지한 후 다른 두 데이터베이스를 거치는 작업을 진행한다.



(그림 1) 이벤트 탐지 시스템 흐름도

3.2 지명 노이즈 제거 기법

지명 노이즈는 크게 동형이의어와 지명을 포함한 단어 두 가지로 나뉜다. 우선 동형이의어는 형태는 같으나 의미가 다른 관계에 있는 단어를 의미한다. 예를 들어 '용인'이라는 단어가 '용인할', '용인되어'와 같이 조사가 함께 사용되면 '용인하다'라는 의미로 단어가 사용된다. 또한 지명을 포함하고 있는 단어로 인해 노이즈가 발생된다. '마이구미' 이라는 단어를 지명 '구미'라고 판단하거나, '아파주면'이라는 단어를 지명 '파주'로 판단해 검출되는 것을 예로 들 수 있다. 이러한 지명 검출 과정에서 생기는 노이즈를 제거해 나가는 기법이 '지명 노이즈 제거 기법' 이다.

3.3 지명 확정 기법

지명 확정 기법은 랜드 마크 데이터베이스를 통해 실제 지명인 단어들만 검출하여 지명으로 확정한다. 랜드 마크란 지역의 이미지를 대표하는 특이성 있는 시설이나 건물을 의미한다. 이벤트가 발생한 위치를 지명보다 더 자세하게 검출하려는 목표를 두고 고안한 기법이다. 각 지역에 있는 모든 학교, 다리, 공항, 지하철역, 항구, 공원, 산, 유적지는 랜드 마크로 저장한다. '서대문구'의 '현대 백화점', '인천시'의 '인천공항' 등을 랜드 마크의 예시로 들 수 있다.

4. 실험결과

본 논문에서 제안한 기법의 성능을 평가할 기준 데이터는 2014년 12월 이후부터 2015년 7월까지 네이버 뉴스 속보로부터 선정한 100개의 이벤트이다[6].

정확도는 시스템에서 탐지한 이벤트 수와 탐지된 이벤트 중에서 실제로 발생한 이벤트 수의 비율을 나타낸다. 정확도를 구하는 공식은 식(1)과 같으며, 실제로 발생한 이벤트를 탐지했는가에 대한 척도이다.

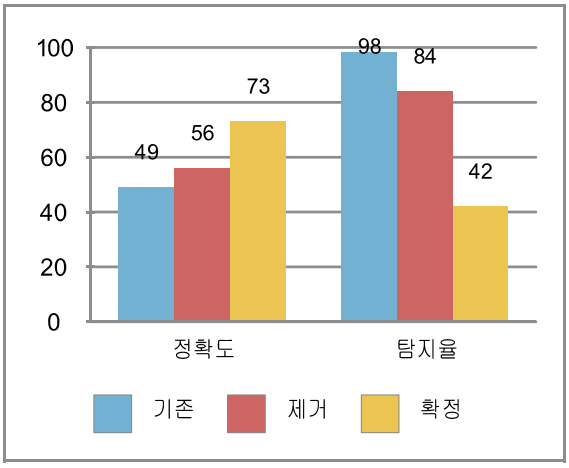
$$\text{정확도}(\%) = \frac{\text{실제로 발생한 이벤트 수}}{\text{시스템에서 탐지한 이벤트 수}} \times 100 \quad (1)$$

기존의 시스템에서 탐지한 100개의 이벤트 중 실제로 발생한 이벤트는 49개로 정확도가 49%였다. 그러나 지명 노이즈 제거 기법을 적용하였을 때 정확도는 56%, 지명 확정 기법의 정확도는 73%로 향상되었다.

탐지율은 실제로 발생한 이벤트의 수와 시스템에서 탐지한 이벤트 수의 비율을 의미하며 공식은 식(2)와 같다. 탐지율이 높을수록 시스템에서 탐지할 수 있는 이벤트의 범위가 넓어지는 것이다.

$$\text{탐지율}(\%) = \frac{\text{시스템에서 제대로탐지한 이벤트 수}}{\text{실제로 발생한 이벤트 수}} \times 100 \quad (2)$$

지명 노이즈 제거 기법은 실제 발생된 속보 이벤트 100개 중에 84개를 탐지하였다. 지명 확정 기법은 42개를 탐지하였다. 기존 시스템에서 98개를 탐지한 점을 감안하면 탐지율은 두 기법을 적용하였을 때 오히려 감소하였다. 지명 노이즈 제거 기법은 실제 지명 단어를 노이즈로 판단하였기 때문이고, 지명 확정 방식은 데이터베이스에 저장되어있는 랜드 마크가 부족했기 때문이다. (그림 2)는 본 논문에서 제시한 두 기법과 기존의 시스템 탐지율 및 정확도를 비교한 그래프이다. (그림 2)에서의 '기존'은 기존 시스템을 의미하고, '제거'는 지명 노이즈 제거 기법, '확정'은 지명 확정 기법을 의미한다.



(그림 2) 정확도와 탐지율

5. 결론 및 향후 연구

본 논문에서는 트위터를 이용한 이벤트 탐지 시스템의

지명 검출 과정을 보다 정확하게 실행시키기 위한 두 기법을 제안하였다. 제안된 기법을 적용하였을 때 지명 노이즈 제거 기법의 정확도는 기존 시스템 보다 7% 향상되었다. 특히 지명 확정 기법의 정확도는 기존 시스템에 비해 24% 대폭 향상되었다. 지명 확정 방식의 정확도가 100%가 아닌 이유는 랜드 마크로 저장된 일부 단어에서도 노이즈가 발생했기 때문이다.

한편, 제안된 기법을 적용할 경우에 기존 시스템보다 탐지율이 하락되었다. 하지만 지명 노이즈 제거 기법은 실제 지명 단어를 노이즈로 판단해 지우는 사례를 줄여나감으로써 탐지율을 확장시킬 수 있다. 또한 지명 확정 기법은 랜드 마크 데이터베이스를 추가적으로 확장해나가면서 탐지율을 확장시킬 수 있다. 랜드 마크는 시시각각 변하는 것이 아니고, 다소 정적으로 변하는 요소임을 감안할 때 탐지율이 대폭 향상될 수 있을 것으로 기대된다.

향후 연구로는 이벤트 탐지 후 전파 방법과 지명 확정 기법에서의 랜드 마크 추가 방법을 찾는 것이다. 전파 방법은 웹과 스마트폰 어플리케이션을 사용하여 알림기능을 적용시킬 계획이다.

참고문헌

[1] J. Yim and B. Hwang, “Twitter Based Realtime Event-Location Detector,” KIPS Transactions on Software and Data Engineering, Vol. 4, No. 8, pp. 301-308, 2015.

[2] J. Hur and C. Ock, “A Homonym Disambiguation System based on Semantic Information Extracted from Dictionary Definitions,” Journal of KIISE : Software and Applications, Vol. 28, No. 9, pp. 688-698, 2001.

[3] J. Shin and C. Ock, “A Stage Transition Model for Korean Part-of-Speech and Homograph Tagging,” Journal of KIISE : Software and Applications, Vol. 39, No. 11, pp. 889-901, 2012.

[4] J. Yim, H. Ha, and B. Hwang, “The Method for Removing Noises from Event Detection using Twitter,” Proc. of KSII Fall Conference, pp. 105-106, 2014.

[5] Twitter Streaming API, http://dev.twitter.com/docs/streaming-apis, 2015.

[6] Naver breaking internet news, http://news.naver.com/main/list.nhn?mode=LSD&mid=sed&sid1=001