음절 n-gram 기반의 미등록 어휘 추정기 구현

신준수, 홍초희
샤인웨어소프트
jsshin@shineware.co.kr, chhong@shineware.co.kr

Out of Vocabulary Word Extractor based on a Syllable n-gram

Junsoo Shin, Chohee Hong
SHINEWARE SOFT

요 약

다양한 콘텐츠가 생성됨에 따라 신조어 및 미등록어도 다양한 형태로 나타나고 있다. 이러한 신조어 및 미등록어는 텍스트 처리 단계에서 오분석 되어 성능 저하의 원인이 된다. 본 논문은 이러한 문제를 해결하기 위해서 대량의 문서로부터 신조어 및 미등록 어휘를 추정하는 방법에 대해서 제안한다. 제안 방법은 대량의 문서로부터 음절 n-gram을 추출한 뒤, 각 n-gram에서 n을 한음절 축소 및 확장 시켜, (n+1)gram, (n-1)gram을 추가적으로 추출한다. 추출된 음절 n-gram을 기준으로 (n+1)gram, (n-1)gram과의 빈도 차이를 계산하여 빈도차가 급격하게 발생하는 구간을 신조어 및 미등록 어휘로 추정한다. 실험 결과 신조어 뿐만 아니라 트위터, 미투데이 등과 같은 도메인에 종속적인 미등록 어휘도 추출되는 것을 확인할 수 있었다.

주제어: 미등록어 추정, 음절 n-gram, 키워드 추출

1. 서론

최근 다양한 분야에서 새로운 콘텐츠(Content)가 생성됨에 따라 신조어 및 미등록어도 다양한 형태로 나타나고 있다. 이러한 신조어 및 미등록어는 텍스트 전처리 단계에서 오분석되어 상위 시스템의 전반적인 성능 저하로 이어진다. 본 논문에서는 이러한 문제를 해결하기 위해 대량의 말뭉치로부터 신조어 및 미등록어를 추출하기 위한 방법을 제안한다. 본 논문의 2장에서 신조어 및 미등록어와 관련된 연구에 대해서 소개하고 3장에서 제안 방법에 대해서 설명한다. 4장에서는 제안 방법에 대한 성능 실험과 결과를 평가하고 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

신조어 및 미등록어 처리를 위한 연구는 전문 용어 추출과 형태소 분석기의 미등록어 추정 연구로 분류 할 수 있다. 전문 용어 추출 연구는 제한된 영역에 대하여 CRFs, SVM 등과 같은 기계 학습 방법 등을 이용하여 용어를 추출하는 방법이다[1][2]. 기존 미등록어 추정 연구는 형태소 분석기 등과 같은 사전 정보를 활용한 시스템에서 사전에 등재되지 않은 어휘가 출현할 때 이를 추정하는 방법이다[3]. 또한 successor variety를 이용하여 대량의 코퍼스로부터 어근을 추출할 수 있는 방법이 있다. 그러나 기계 학습을 이용한 방법은 영역에 따라 서로 다른 학습 데이터가 필요하다는 단점이 있으며, 미등록어 추정은 사용자가 직접 사전 정보를 관리해야한다는 단점이 있다. 또한 successor variety는 신조어와 같이 어근이 명확하지 않은 분야에 적합하지 않다는 단점이 있다.

본 논문에서는 기존 연구의 단점을 보완하기 위해 대량의 말뭉치로부터 다양한 형태의 어휘 후보를 자동으로 추출하는 방법을 제안한다.

3. 미등록 어휘 추정 시스템 구현

문서 내에서 자주 표현되는 어휘는 앞, 뒤에 다양한 활용이 가능하다. 예를 들어 “아이폰”과 같은 어휘는 “아이폰은”, “아이폰을”, “아이폰에서”등과 같은 활용형 어휘가 나타난다. 이 때 전체 말뭉치에서 빈도수를 살펴보면 “아이폰”이 “아이폰*”보다 훨씬 더 높은 빈도를 갖는다. 따라서 이 빈도 차이를 이용하여 “아이폰”을 하나의 신조어 및 미등록어로 추정 할 수 있다. 본 논문에서는 이러한 현상을 통하여 미등록 어휘를 추정하는 시스템을 구현한다.

그림 1은 본 논문에서 제안하는 미등록 어휘 추정 시스템의 구조도이다. 다량의 문서로부터 음절 n-gram과 wildcard를 적용한 음절 n-gram을 추출하고 각 빈도를 계산한다. 계산된 빈도를 기반으로 음절 n-gram 빈도와 wildcard 음절 n-gram 빈도를 비교 후 임계치 이상의 빈도차를 보이는 어휘를 최종 미등록 어휘로 판별한다.

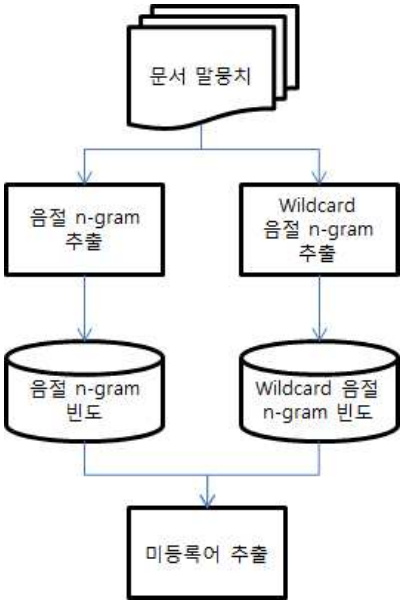

그림 1. 미등록 어휘 추정 시스템 구조도

3.1. 음절 n-gram 추출

문서 내 출현하는 어절을 기준으로 음절 n-gram을 추출하여 빈도수를 계산한다. 추출된 음절 n-gram의 예는 아래 표 1과 같다.

표 1. 음절 n-gram 및 빈도수

음절 n-gram	빈도수
아이	3098
아이폰	2017
아이폰은	28
아이폰을	29
아이폰에	20
아이폰에서	17
아이폰용	19
트윗	2072
트윗은	175
트윗을	158
트윗이	154

3.2. Wildcard 음절 n-gram 추출

Wildcard 음절 n-gram은 검색 엔진에서 사용되는 wildcard 질의처리 방식을 이용하여 추출한다. 예를 들어 “아이폰은”, “아이폰을”, “아이폰에”는 “아이폰*”로 표현이 가능하다. 이때 “아이폰*”의 빈도수를 측정한다. 추출된 wildcard 음절 n-gram의 예는 아래 표 2와 같다.

표 2. Wildcard 음절 n-gram 및 빈도수

n-gram	빈도수
아*	3098
아이*	2017
아이폰*	96
아이폰에*	17
트*	2072
트윗*	487

3.3. 미등록 어휘 추출

추출된 음절 n-gram 중 wildcard 음절 n-gram과 비교하여 빈도수의 차이가 급격하게 나타나는 n-gram을 최종 미등록 어휘로 추출한다. 이 때 wildcard 음절 n-gram에서 출현하지 않는 어휘는 추출하지 않는다. 예를 들어 표 1의 음절 n-gram “아이폰”의 빈도수는 2017이며, 표 2의 wildcard 음절 n-gram “아이폰*”의 빈도수는 96로 급격한 차이를 보이므로 미등록 어휘로 추출된다. 그러나 음절 n-gram “아이폰에”와 wildcard 음절 n-gram “아이폰에*”의 빈도는 각각 20, 17으로 빈도수 차이가 상대적으로 낮으므로 미등록 어휘로 추출되지 않는다. 표 1과 표 2를 통해 추출된 미등록 어휘는 아래 표 3과 같다. 표 3에서는 빈도수의 차이가 음절 n-gram 빈도수의 70% 이상이 되는 어휘만을 최종 후보로 추출하였다.

표 3. 추출된 미등록 어휘

결과	음절 n-gram 빈도	wildcard 음절 n-gram 빈도	빈도 차이
아이폰	2017	111	1906
트윗	2072	487	1585

4. 실험 및 결과

본 논문에서 제안한 미등록 어휘 추출 방법의 성능 평가를 위해 트위터 10만개, 미투데이 30만개의 글을 수집하였다. 표 4는 제안 시스템으로 추출된 어휘 중 빈도수 상위 20개를 나타낸다.

표 4. 제안 방법의 추출 결과 중 빈도수 상위 20 개의 어휘

순위	트위터 어휘	빈도	미투데이 어휘	빈도
1	;;	3001	ㅠㅠ	21692
2	아이폰	2150	ㅎㅎ	15104
3	트윗	1924	ㅜㅜ	5724
4	트위터	1642	미투데이	3886
5	한성주	1291	친신	3592
6	아..	988	ㅡㅡ	3386
7	ㄷㄷ	795	댓글	3128
8	화이팅	749	아이유	2786
9	대박	718	미친님들	2205
10	멘션	715	2011년	2046
11	:)	713	미친들	2002
12	ㅇㅇ	697	심심해	1974
13	아님	587	ㄱㄱ	1870
14	아~	581	아...	1844
15	선팔	567	격반	1571
16	음.	466	드림하이	1560
17	한국쓰리엠	425	인첵	1555
18	우와	421	>_<	1536
19	팔로우	385	아~	1535
20	네이버	368	[인첵]	1492

본 논문에서 제안한 방법으로 트위터에서 추출된 상위 20개의 결과 중 “트윗”, “트위터”, “멘션”, “선팔” 등과 같이 트위터에서 주로 사용되는 어휘가 추출된 것을 확인할 수 있다. 또한 미투데이에서 추출된 결과 중 “친신”(친구신청), “미친님들”, “미친들”(미투데이 친구들), “격반”(격하게 반긴다), “인첵”(인원체크) 등의 줄임말이 출현하는 것을 확인할 수 있다. 상위 20개의 결과 외에도 트위터에서는 “카톡”, “아이패드”, “블랙베리”, “페이스북”, “안드로이드” 등의 어휘 들이 추출 되었으며, 미투데이에서는 “소녀시대”, “비스트”, “화이트데이”, “샤이니” 등의 어휘들이 추출되는 것을 확인할 수 있다.

그러나 “;;”, “ㄱㄱ”, “우와”,“암튼”, “있네요” 등과 같이 의미가 없는 어휘들도 상당 수 추출되는 것을 확인 할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 음절 n-gram을 기반으로 한 미등록 어휘의 추정 방법에 대해서 제안하고 이를 평가하기 위하여 신조어 및 미등록어가 많이 출현하는 데이터를 사용하여 실험 하였다. 실험 결과 특정 사이트에서 빈번하게 사용되는 어휘들이 추출되는 것을 확인할 수 있었으며 연예인명, IT 용어 등과 같은 신조어들도 추출되는 것을 확인할 수 있었다. 그러나 의미가 없는 어휘들이 추출되는 것을 확인하였는데 이를 해결하기 위해 시간대 별로 카테고리를 분류하여 “ㄱㄱ”,“우와” 등과 같이 시간에 무의미한 어휘를 추출하여 적용할 수 있는 방법에 대해서 연구할 예정이다. 또한 주제 별로 카테고리를 분류하여 주제에 무의미한 일반적인 어휘를 추출하여 적용할 수 있는 방법에 대해서도 연구를 진행할 예정이다.

참고문헌

[1] 송수민, 신준수, 김학수, “음절 바이그램과 CRFs를 이용한 의학 전문 용어 추출”, 한국정보처리학회 춘계 학술발표대회, pp. 505-507, 2010.

[2] 최맹식, 김학수, “기계학습에 기반한 한국어 미등록 형태소 인식 및 품사 태깅”, 정보처리학회논문지, 18-B권 1호, pp. 45-50, 2011.

[3] 고영중, “한국어 명사 출현 특성과 후절어를 이용한 명사 추출기” 정보과학회논문지: 소프트웨어 및 응용 제 37권 12호, pp. 919-927, 2010