단어 임베딩과 음성적 유사도를 이용한 트위터 '서치 방지 단어'의 자동 예측

이상아○
서울대학교 언어학과
visualjan@snu.ac.kr

Automatic Prediction of 'Anti-Search Variants' of Twitter based on Word Embeddings and Phonetic Similarity

Sangah Lee○
Seoul National University, Dept. of Linguistics

요 약

'서치 방지 단어'는 SNS 상에서 사용자들이 작성한 문서의 검색 및 수집을 피하기 위하여 사용하는 변이형을 뜻한다. 하나의 검색 키워드가 있다면 그와 같은 대상을 나타내는 변이형이 여러 형태로 존재할 수 있으며, 이들 변이형에 대한 검색 결과를 함께 수집할 수 있다면 데이터 확보가 중요하게 작용하는 다양한 연구에 큰 도움이 될 것이다. 본 연구에서는 특정 단어가 주어진 키워드로부터 의미 벡터 상의 거리가 가까울수록, 그리고 주어진 키워드와 비슷한 음성적 형태 즉 발음을 가질수록, 해당 키워드의 변이형일 가능성이 높을 것이라고 가정하였다. 이에 따라 단어 임베딩을 이용한 의미 유사도와 최소 편집 거리를 응용한 음성적 유사도를 이용하여 주어진 검색 키워드와 유사한 변이형들을 제안하고자 하였다. 그 결과 구성된 변이형 후보의 목록에는 다양한 형태의 단어들이 포함되었으며, 이들 중 다수가 실제 SNS 상에서 같은 의미로 사용되고 있음이 확인되었다.

주제어: 서치 방지 단어, 단어 임베딩, 음성적 유사도, 최소 편집 거리

1. 서론

최근 트위터나 블로그와 같은 SNS의 데이터를 이용하여 특정 주제에 대한 대중의 의견을 수집하고 참고하려는 움직임이 커지고 있다. 한편 개인이 SNS에서 자유롭게 피력한 의견들이 수집을 위한 검색에 노출될 것을 우려하여, 이에 반하는 전략들도 나타나기 시작하였다. 그 중 하나는 '서치 방지 단어'의 사용인데, '서치 방지 단어'란 SNS 상에서 검색을 피하기 위해 사용되는, 특정 단어의 변이형들을 말한다.

[표1] 서치 방지 트윗 예시

헬씽ㅋi 스벆은 휘핑크림을 재밌게 주는군요!

예를 들면 [표1]의 트윗에서 '헬씽ㅋi'와 '스벆'은 각각 '헬싱키'와 '스벅(스타벅스)' 대신에 사용된 서치 방지 단어라고 할 수 있다. SNS의 특성상 작성자의 신분이나 위치 등이 드러날 수 있으므로 외부 검색에 의한 접근은 원치 않으나, 해당 단어들을 언급하거나 그에 관한 글을 작성하고자 할 때 서치 방지 단어를 사용하는 것이 일반적이다. 보통은 한 가지 형태로 약속되기보다는 다양하게, 산발적으로 나타난다.

이러한 검색 방지의 유형은 크게 단어의 시각적 형태에 따른 것, 청각적 형태에 따른 것의 두 가지로 나누어 볼 수 있다. 먼저 시각적 형태에 주목한 경우에는 한글 자모를 조합한 모양이 비슷한 것들끼리 대치하는 소위 '야민정음'[1]과, 영문자나 숫자, 특수기호 등을 섞어 한글처럼 보이게 하는 경우[2] 등이 포함된다. 청각적 형태에 따른 변이형은 소리내어 읽었을 때 유사한 발음으로 실현되는 것[3]이며, 본 연구는 이 경우에 중점을 두어 진행되었다. 한편 형태 측면과는 다소 거리가 있으나, 집단 내에서 암묵적으로 약속된 대체어를 이용하는 경우[4]나 해당 단어의 초성만을 기재하는 경우[5] 역시 존재한다.

본 연구에서는 하나의 키워드를 기준으로 발생 가능한 변이형들을 자동으로 예측하여 얻고자 하였다. 이를 위하여 단어 임베딩을 이용한 단어 간의 의미적 유사도와 최소 편집 거리(Minimum Edit Distance)[1]를 응용하여 얻은 단어 간의 음성적 유사도를 함께 고려하였다. 최종적으로는 의미적 유사도와 음성적 유사도를 가중합하여 주어진 단어와 변이형 후보 사이의 유사도 점수를 계산하고, 이 점수가 높은 순서대로 변이형을 제안한다.

이러한 흐름에 따라 한 가지 키워드를 가지고 SNS 상에 나타난 대중의 의견을 검색할 때, 형태는 조금 다르

1) (예) 파이널판타지(파이널판타지)
2) (예) ㅇ1디야(이디야커피), ㅂ6탄쇼넨단(방탄소년단)
3) (예) 스따벆스(스타벅스), 탬인(태민), 정핝(정한)
4) (예) 넬(강다니엘)
5) (예) ㅂㅌ(방탄, 방탄소년단)

지만 같은 대상을 나타내는 변이형들을 언급한 문서들도 함께 검색되도록 한다. 이를 통해 보다 풍부한 검색 결과를 얻고, 이러한 데이터의 확장은 다양한 자연어처리 연구에 도움이 될 것이다.

2. 관련 연구

검색 키워드의 가능한 변이형을 제안하는 연구는 아직 충분히 이루어지지 않은 실정이다. 주어진 단어(문자열) 사이의 유사도를 측정하는 방법은 다양하게 제시되었으나, 이를 응용하여 발생 가능한 문자열의 변이를 예측하는 연구는 드문 편이다.

생물정보학 분야의 서열 정렬(sequence alignment) 방식과 편집 거리 방식을 함께 이용하여 오자, 탈자 등의 입력 오류를 허용하는 근사 한글 검색 시스템이 앞서 제안된 바 있다. 해당 연구에서는 같은 조음 위치에서 발음되는 평음, 경음, 격음을 한 가지 평음으로, 이중모음을 단모음으로 표준화하여, 다양한 형태로 변형된 욕설을 필터링하는 시스템을 구축하였다[2]. 또한 다수의 문자열에 대하여 서열 정렬을 수행하는 다중 서열 정렬(multiple sequence alignment) 방식을 채택한 연구도 존재한다[3]. [3]의 연구에서는 하나의 문자열로부터 파생된 변이형들이 포함된 문자열 집합이 주어졌을 때 그 중 대표 문자열을 정의하고, 문자열과 문자열 집합 간의 유사도 계산 방법을 제안하여 문자열 집합 내에 특정 문자열이 포함되어 있는지 아닌지를 출력하는 문제를 다루었다.

한편 스마트폰 가상 키패드 상에서 발생할 수 있는 입력 오류에 의한 유사 단어를 검색하는 문제도 다루어졌다[4]. 이 연구에서는 스마트폰 키패드의 종류에 따라 편집 거리 알고리즘에 사용되는 편집 비용을 수정하는 방식을 택하였다.

본 연구는 단어로부터 발생 가능한 변이형을 예측한다는 점에서는 이전 연구들과 공통적이나, 단어 사이의 편집 거리를 이용해 정의한 음성적 거리와 단어 임베딩에 기반한 의미의 유사성을 함께 고려하고자 하였다.

3. 서치 방지 단어의 자동 예측

3.1 어휘 사이의 의미 유사도

먼저 문맥에서 얻어지는 단어와 단어 사이 의미의 유사성을 이용하여 변이형을 예측한다. 형태가 다를지라도 같은 것을 뜻하는 단어들의 경우 공기하는 단어들이나 문맥과 관련하여 제공하는 정보가 서로 유사할 가능성이 높기 때문이다.

이러한 의미의 유사성을 수치화하고 연산하기 위하여 트위터에서 수집한 훈련 데이터에서 Word2Vec의 C-BOW 모델을 이용하여 단어 임베딩을 구축하였다[5]. 훈련 데이터는 각각 140자 이내로 작성된 10486개의 트윗으로 이루어져 있으며, 단어마다 갖는 벡터는 100차원의 자질로 구성되어 있다.

검색에 주로 사용되는 키워드는 일반명사나 고유명사인 경우가 많으므로, 시험 데이터에서 형태소 분석을 통해 품사가 명사인 요소들만을 필터링하여 유사도 계산의 대상으로 한다. 이 때 형태소 분석에는 KoNLPy 패키지 내 Twitter 모듈을 이용하였다[6]. 이렇게 걸러낸 명사들은 넓은 의미에서 변이형의 후보가 된다. 이 때 각각의 후보 단어들과 주어진 기본 키워드 사이의 의미 유사도는 100차원의 단어 임베딩끼리의 코사인 유사도를 이용한다. 시험 데이터에서, 훈련 데이터에는 존재하지 않았던 새로운 어휘가 나타난 경우, 똑같이 100차원의 자질을 가지되 균등 분포를 따르는 임의의 벡터를 생성하여 유사도를 계산할 수 있도록 하였다. 이렇게 계산된 의미 유사도는 0에서 1사이의 값을 갖게 된다.

3.2 어휘 사이의 음성적 유사도

다음으로는 단어와 단어가 음성적으로 유사한 정도를 척도의 하나로 이용하고자 하였다. 음성적 유사도는 최소 편집 거리 알고리즘에 한국어 자소의 음성적 특성[7]을 일부 적용하여 구현하였다.

먼저 최소 편집 거리는 두 문자열이 서로 얼마나 비슷한지를 나타내는 척도 중 하나로, 한 단어의 철자가 다른 단어와 같아지도록 수정하는 과정(철자의 삽입, 삭제, 대체) 각각에 비용을 부여하는 방식이다. 본 연구에서는 먼저 각각의 단어를 자소 단위로 분해하고, 최소 편집 거리 알고리즘을 적용하되 자소들의 특성에 따라 자소의 대체 비용에 각각 다른 값을 부여하는 규칙을 정의하였다. 기본 대체 비용을 1로 설정하고, 아래 [표2]의 기준들에 해당하는 자음, 모음의 그룹 내에서 발생하는 대체에는 0과 1 사이의 값을 정의하여 이용하였다.

[표2] 편집 대체 비용의 판단 기준

기준	자소 그룹	대체 비용
자음의 조음 위치	{ㄱ, ㄲ, ㅋ}, {ㄷ, ㄸ, ㅌ}, {ㅂ, ㅃ, ㅍ}, {ㅅ, ㅆ}, {ㅈ, ㅉ, ㅊ}	0.5
모음 발음상의 유사성	{ㅏ, ㅐ}, {ㅗ, ㅜ}, {ㅓ, ㅗ} 등	0.5
	{ㅔ, ㅐ}, {ㅖ, ㅒ} 등	0.2

자음의 조음 위치에 따른 유사성은 같은 조음 위치에서 발음되는 자음들 중 평음, 격음, 경음의 대립이 존재하는 경우를 상정하여 정의하였다. ‘ㄱ, ㄲ, ㅋ’, ‘ㄷ, ㅌ, ㄸ’ 등의 자음들은 임의로 대체하더라도 발음상의 차이가 덜하여 원래의 형태를 알기 쉬운 편이다.

모음 발음 시 혀의 위치, 입의 개폐 정도의 유사성은 주로 모음사각도 상의 거리에 기반하여 정의된다(‘ㅏ, ㅐ’, ‘ㅔ, ㅐ’ 등). 또한, 이중모음의 경우에도 발음의 유사성과 서치 방지의 일반적인 전략을 고려하여 대체하기 쉬운 모음의 목록을 구성하였다(‘ㅏ, ㅑ’, ‘ㅐ, ㅒ’ 등).

한편 종성에 쓰이는 겹받침 역시 서치 방지 단어의 생

성 전략이 될 수 있다. 발음이 같거나 비슷한 것을 이용하여 홑받침을 겹받침으로, 겹받침을 홑받침으로 대체하는 것이다. 따라서 음성적 유사도를 계산하기 이전에 이들을 각각의 자소로 분리하였다. 예를 들면 'ㅈ'를 'ㄴㅈ'로, 'ㅛ'를 'ㄹㅁ'로 변환하는 것이다.

이러한 방법으로 얻은, 주어진 키워드 w_i와 시험 데이터 내 명사 w_j 사이의 최소 편집 거리를 아래의 식 (1)을 통해 가공하여, 음성적 유사도는 0에서 1 사이의 값을 가지도록 하였다.

$$w_i, w_j\text{의 음성적 유사도} = 1/(w_i, w_j\text{의 최소 편집 거리} + 1) \cdots (1)$$

4. 실험 및 결과

본 연구에서는 각각 140자 이내로 작성된 9547개의 트윗으로 구성된 시험 데이터에 출현한 명사들을 대상으로, 주어진 키워드와의 의미 유사도와 음성적 유사도에 기반한 유사도 점수를 계산하여 실제로 변이형이 될 만한 단어들을 얻고자 하였다. 이 때 유사도 점수는 의미 유사도와 음성적 유사도에 똑같이 0.5의 가중치를 부여한 평균값이 된다.

K-pop 아이돌 그룹명인 '엑소'와 그 멤버 이름인 '세훈'을 기본 키워드로 부여하고, 시험 데이터에 쓰인 명사들 중 이들 키워드와 가장 유사한 20개의 명사를 추출하였다.

[표3] 유사도 기반 변이형 제안 결과

키워드='엑소' (유사도)	키워드='세훈' (유사도)
엑소 (1.0)	**세훈 (1.0)**
액소 (0.748)	**새훈 (0.799)**
엑스 (0.616)	**세혼 (0.745)**
예고 (0.573)	**셰훈 (0.725)**
엔시 (0.571)	**새혼 (0.711)**
앢소 (0.563)	**섀훈 (0.706)**
악수 (0.553)	**세후 (0.705)**
엓소 (0.564)	세운 (0.694)
엔씨 (0.552)	**세홍 (0.685)**
애교 (0.551)	**새후 (0.659)**
육수 (0.550)	**세후니 (0.652)**
백시 (0.549)	**새홍 (0.641)**
윾소 (0.547)	세준 (0.632)
야기 (0.531)	**새휸 (0.628)**
해고 (0.530)	쳰 (0.615)
엥셔 (0.530)	백현 (0.614)
악어 (0.530)	**쎄홍 (0.613)**
섹시 (0.528)	수호 (0.613)
센스 (0.527)	**오세훈 (0.611)**
에스 (0.526)	세호 (0.610)

위 [표3]에서 키워드인 '엑소'와 '세훈'의 실제 변이형인 명사들은 굵은 글씨로 표시하였다. 키워드가 '엑소'인 경우 제안된 상위 10개의 변이형 중에서는 4개의 명사, 상위 20개의 변이형 중에서는 5개의 명사가 실제로 사용되고 있음을 확인하였다. 또한 키워드가 '세훈'인 경우 제안된 상위 10개의 명사 중에서는 9개, 상위 20개 중에서는 14개의 변이형이 정답에 해당하는 것으로 확인되었다. 이러한 결과는 [표4]에 정리되어 있으며, 두 경우 모두 정답률은 상위 10개까지의 변이형을 채택했을 때 더 높게 나타났다.

또한 같은 키워드에 대하여 의미 유사도와 음성적 유사도 각각만을 가지고 단어 사이의 유사도를 계산하고, 이를 두 가지 유사도를 모두 사용한 경우와 비교하였다. 이 때 [표4]의 결과에 따르면 두 가지 유사도를 함께 적용했을 때의 정답률이 가장 높았으므로, 적절한 변이형을 제안하는 데 두 척도가 모두 기여한다는 것을 확인하였다.

[표4] 키워드 변이형 정답률

이용한 유사도	키워드	정답 개수 (%)	
		상위 10개	상위 20개
의미 유사도	엑소	1 (10%)	1 (5%)
	세훈	1 (10%)	1 (5%)
음성적 유사도	엑소	6 (70%)	7 (35%)
	세훈	9 (90%)	13 (65%)
의미 유사도 + 음성적 유사도	엑소	**7 (70%)**	**11 (55%)**
	세훈	**9 (90%)**	**15 (75%)**

음성적 유사도는 고려하지 않고 단어 임베딩에 의한 의미 유사도만을 이용한 경우에는 관련된 단어들을 효과적으로 추출하였으나, 정확히 같은 대상을 가리키는 변이형은 제안하지 못하였다. '엑소'와 '세훈'을 키워드로 설정했을 때, 두 경우 모두 해당 아이돌 그룹에 소속된 다른 멤버의 이름('준면', '찬열', '종인', '백현' 등)이 0.99 이상의 높은 코사인 유사도 값을 보이며 변이형으로 예측되었으나, 동일한 대상을 가리키는 변이형을 보다 정확하게 얻기 위해서는 음성적 유사도를 반영할 필요가 있다. [표3]에 제시된 '액소', '앢소', '엓소' 등의 형태는 의미 유사도만으로는 예측할 수 없기 때문이다.

그러나 단어의 의미를 고려하지 않은 채 청각적 형태의 유사성만을 변이형 예측의 단서로 삼는 것은 효과적인 방법이 아니다. 음성적 유사도만을 이용하여 변이형의 후보를 추출할 경우, '악수', '에고', '엑스' 등의 단어들이 0.5의 유사도를 가지고 높은 순위로 예측된다. 이들 단어는 자소들의 음성적 특징에 따라서는 제시된 키워드와 유사하지만 의미 면에서는 그 차원이 전혀 다르다. 따라서 특정 키워드의 변이형을 예측할 때 그 단어가 문맥 속에서 가지는 의미를 반영하여야 할 것이다.

한편 당초 상정하지 않았던 변이형이 높은 점수를 보이며 예측 결과에 포함되는 경우도 존재하였다. [표3]에 굵은 글씨로 표시되어 있는 '윾소', '세혼, 새혼, 세후, 새후, 세후니, 새휸, 쎄홍, 오세훈'[6]이 그것이다. 이들 변이

6) '세후, 새후'의 경우 '세후나(세훈아)', '새후니(새훈이)' 등의 형태소 분석 오류에 의한 것을 포함한다.

형은 학습 데이터와 시험 데이터를 수집하는 과정에서 미리 설정하지 않았으므로 우연히 텍스트 안에 포함된 것들이다. 이는 곧 미리 정답셋으로 구성한 '엑소, 액소, 엑쏘, 엑오, 앸오, 엑소', '세훈, 새훈, 세훙, 새훙, 셰훈, 섀훈'과 같은 형태 외에도 다양한 변이형이 추론될 수 있고, 정답인 변이형을 포함하리라고 보장할 수 없는 임의의 데이터에서도 동일한 연산이 가능하다는 것을 뜻한다.

이러한 분석을 통해, SNS 상에서 사용자들이 특정 주제에 해당하는 키워드를 언급하되 검색 결과에 노출되는 것을 피하기 위하여 어떤 변이형을 쓰는지 예측할 수 있을 것이다.

5. 결론

본 연구는 Word2Vec 모델에 기반한 단어 임베딩을 구성하여 단어 간의 의미 유사도를 구하고, 최소 편집 거리를 응용하여 단어 사이의 음성적 유사도를 계산하여, 주어진 단어로부터 '서치 방지 단어' 즉 해당 단어의 변이형의 후보들을 이끌어내는 것을 목표로 하였다.

의미와 음성적 형태를 동시에 고려한 점수가 높은 명사일수록 변이형이 될 가능성이 높으며, 이러한 변이형을 포함한 문서 즉 트윗은 주어진 키워드의 검색 결과로 함께 제안할 만하다. 해당 트윗은 검색된 키워드와 같은 대상을 가리키는 단어를, 자소의 결합으로 이루어진 형태는 조금 다를지라도, 포함하고 있을 가능성이 높다. 이에 따라 기존의 검색 방법으로는 쉽게 얻을 수 없었던 문서들을 검색 결과에 노출시키고, 특정 주제에 대한 자연어 데이터를 확장함으로써 다양한 연구에 도움이 될 것이다.

향후에는 음성적 유사도를 위한 자소 간의 대체 비용을 정의할 때, 조음 위치나 혀의 위치 이외에 자음 동화와 같은 추가적인 음운 지식을 적용해볼 수 있을 것으로 생각된다. 또한 트위터와 같은 SNS의 특성상 문서 작성자의 ID나 최근 관심사, 거주 지역과 같은 정보, 사용자 간 네트워크 정보 등도 변이형 제안에 중요한 요소로써 고려해볼 수 있을 것이다.

참고문헌

[1] Jurafsky and Martin, Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition, Pearson Prentice Hall, 2009.

[2] 윤태진, 조환규, 정우근, "제한된 한글 입력환경을 위한 음소기반 근사 문자열 검색 시스템", 정보과학회논문지 : 소프트웨어 및 응용, 제37권, 제10호, pp. 788-801, 2010.

[3] 김성환, 조환규, "다중서열정렬을 이용한 변형 문자열 집합의 유사도 계산 기법", 정보과학회논문지 : 소프트웨어 및 응용, 제40권, 제1호, pp. 53-60, 2013.

[4] 송명길, 김학수, "다양한 스마트폰 키패드 환경에서 유사 단어 검색을 위한 수정된 편집 거리 계산 방법", 한국콘텐츠학회논문지, 제11권, 제12호, pp.12-18, 2011.

[5] Mikolov et al., Distributed representations of words and phrases and their compositionality, In Advances in neural information processing systems, pp. 3111-3119, 2013.

[6] 박은정, 조성준, "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지", 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.

[7] 이호영, "국어 음성학", 태학사, 1996.