논문 19-45-01-12 The Journal of Korean Institute of Communications and Information Sciences '20-01 Vol.45 No.01
https://doi.org/10.7840/kics.2020.45.1.96

트위터와 단어 임베딩을 사용한 인플루엔자 감지

김 인 환◆, 장 백 철°

Influenza Detection Using Twitter and Word Embeddings

Inhwan Kim◆, Beakcheol Jang°

요 약

유행성 독감은 매년 전 세계적으로 300~500만 명의 중증 질환을 야기하며, 29~65만 명의 사망자를 발생시키는 질병이다. 유행성 독감의 피해를 최소화하기 위하여 질병관리본부(KCDC)는 독감에 대한 표본 감시 자료를 제공하고 있지만 실제 질병의 발생과 표본 감시 자료 사이에는 1~2주의 시간 차이가 발생한다. 따라서, 검색 엔진의 검색어 및 소셜 네트워크 서비스와 같은 실시간 웹 데이터를 사용하여 자료 제공과 실제 질병 발생 사이의 시간 차이를 줄임으로써 초기에 독감 발생 징후를 감지하는 것은 중요한 일이다. 트위터는 소셜 네트워크 서비스 중 하나로서 실시간으로 독감의 발생 징후를 예측하는데 적합한 데이터이며 단어 임베딩은 트윗들을 학습하여 독감과 연관성이 높은 단어들만을 추출해 예측 모델의 정확성을 향상시킬 수 있다. 본 연구는 단어 임베딩을 사용하여 트윗을 학습시키고 독감과 연관성이 높은 단어들을 추출하여 추출된 단어들이 포함된 트윗들이 제공하는 정보를 통해 실시간으로 독감의 발생 징후를 감지하는 회귀 모델을 제안한다. 최신 단어 임베딩 기술인 Word2vec, GloVe, Fasttext을 통해 추출된 단어를 사용한 회귀 모델의 정확도를 비교하였고 Word2vec에서 추출된 단어를 사용한 회귀 모델이 실제 표본 감시 자료와 가장 높은 0.9718의 상관 비율을 갖는 것을 보였다.

Key Words : Influenza Detection, Regression Model, Twitter, Word Embeddings, Natural Language Processing, Word2vec, GloVe, Fasttext, Skip-gram, CBOW

ABSTRACT

Influenza is a disease that causes between 3 and 5 millions serious illnesses worldwide and produces between 290,000 and 650,000 deaths each year. To minimize the impact of influenza, the KCDC provides surveillance data for influenza, but there is a reporting delay of 1~2 weeks between actual outbreaks and surveillance data provided. It is therefore important to detect influenza early by using real-time web data such as search queries and social network services to reduce reporting delay. Twitter is one of the social network services that is suitable for predicting the outbreaks of influenza in real time, and word embeddings can improve the accuracy of predictive models by learning tweets and extracting words that are highly related to influenza. This study proposes a regression model that learns tweets using word embeddings, extract words that are highly related to influenza, and detects the signs of influenza in real time through the information provided by tweets that contain extracted words. We compared the accuracy of regression models using words extracted from Word2vec, GloVe and Fasttext, which are the states-of-arts word embeddings, and found that regression models using words extracted from Word2vec have the highest correlation ratio of 0.9718 with the surveillance data.

◆ First Author : Sangmyung University Department of Computer Science, moreih29@gmail.com, 학생회원
° Corresponding Author : Sangmyung University Department of Computer Science bjang@smu.ac.kr, 정회원
논문번호 : 201906-112-C-RN, Received June 26, 2019; Revised September 9, 2019; Accepted October 2, 2019

Ⅰ. 서 론

유행성 독감(Influenza)은 매년 전 세계적으로 300만~500만 명에게 중증 질환을 발생시키며, 29만~65만 명의 사망자를 발생시키는 위험한 질병이다[1]. 국내에서도 많은 유행성 독감 환자가 발생하고 있으며 질병관리본부(Korean Centers for Disease Control and Prevention: KCDC)는 인플루엔자에 대한 표본 감시 체계를 구축하여 인플루엔자의 감시를 위해 인플루엔자 의사환자(Influenza-Like illness: ILI)에 대한 임상감시와 병원체 감시를 수행하고 있다. 질병관리본부의 인플루엔자 표본 감시 체계는 일주일 단위로 수집된 자료를 제공하고 있으나 제공된 자료는 실제 질병의 유행과 약 1~2주의 시간 차이가 발생한다. 유행성 질병의 징후를 신속하게 식별하는 것은 질병으로 인한 피해를 최소화 하는데 필수적이기 때문에 웹에서 생산되는 비 임상 데이터를 활용하여 유행성 질병의 징후를 감지함으로써 제공된 정보와의 시간 차이를 줄이기 위한 많은 연구들이 진행되고 있다.

Ginsberg 등[7]은 구글의 검색어 정보를 활용하여 미국의 9개 지역에 대한 인플루엔자 발생을 신속하게 감지하기 위한 연구를 진행하였으며 Lampos 등[5]은 Ginsberg 등[7]이 제안한 구글의 검색어 정보 기반 인플루엔자 감시 시스템의 성능을 향상시키는 연구를 진행했다. 국내에서는 Kwon 등[3]이 네이버 검색 엔진의 검색어를 사용하여 유행성 독감을 감지하는 연구를 진행하였다. 또한 검색어 정보를 활용한 연구뿐만 아니라 소셜 네트워크 서비스(SNS)를 활용한 연구들도 진행되고 있다. 트위터는 많이 사용되는 SNS중 하나로서, 2018년을 기준으로 약 3억 명의 사용자를 보유하고 매일 5억 개의 트윗이 생성된다[2]. Paul 등[14]은 트위터를 사용하여 인플루엔자 예측의 정확도를 향상시킬 수 있다는 것을 밝혔으며, Culotta[13]는 수집된 트위터에서 연관성이 없는 트윗을 제거했을 때 인플루엔자 예측의 정확성이 향상된다는 연구 결과를 발표했고, Kim 등[16]은 한글을 사용한 국내 트위터 사용자의 트윗을 수집하여 인플루엔자의 발생을 예측하는 모델을 제안했다. 이러한 위 연구들을 통해 비 임상 데이터를 활용하여 인플루엔자 발생을 예측함으로써 표본 감시 시스템을 활용한 실제 질병의 유행 정보 제공과의 시간 차이는 줄어들고 있다. 그러나 비 임상 데이터를 사용하여 인플루엔자의 징후를 감지할 때 어려운 점들은 아직 남아있으며 예측 모델에 사용할 단어들의 선택은 그것들 중 하나이다.

인플루엔자의 감시를 위해 검색 엔진이나 트위터를 활용할 때 사용하는 검색 단어와 단어의 개수는 예측 모델의 정확성에 영향을 미치기 때문에 사용할 단어를 선택하는 것은 중요하다[7,13,16]. 그럼에도 불구하고 모든 단어에 대해 임상 감시 자료와 상관관계를 분석하는 것은 많은 시간을 소모하기 때문에 기존의 연구들은 연관성이 있다고 판단되는 일반적인 단어들만을 사용하거나[13], 다수의 단어를 선택 후 상관관계가 높은 단어들을 사용하거나[16], 일반적인 단어들을 조합하여 다수의 단어를 추출하는[3] 방법을 사용했다. 그러나 기존의 방법은 성능의 향상을 위해 단어를 추가하는 것이 어렵거나, 초기 다수의 연관 단어를 선택함에 있어 시간이 많이 소모되기 때문에 임의의 개수만큼 연관된 단어를 추출하는 방법이 필요하다.

단어 임베딩 기술은 문서 내에 포함된 각 단어들의 의미론적 유사성을 반영하여 각 단어들이 벡터 값을 가질 수 있도록 학습하는 자연어 처리 기술이다. 단어 임베딩은 뉴스 추천[9], 트위터 감정분석[15], 주제 추론[12] 등 자연어를 처리하는 다양한 분야에서 함께 사용되지만 입력되는 문서의 유형에 따라 좋은 성능을 보이는 단어 임베딩 기술이 다르기 때문에 데이터 타입에 따라 단어 임베딩 기술의 성능을 비교하는 연구가 최근 활발히 진행되고 있다[6].

따라서 본 연구는 트위터와 최신 단어 임베딩 기술인 Mikolove 등[8]의 Word2vec, Pennington 등[11]의 GloVe, Joulin 등[10]의 Fasttext를 활용하여 각각 독감과 연관된 단어들을 추출하고 추출된 단어들을 사용하여 예측 모델을 구현한다. 모델의 정확성을 통해 세 단어 임베딩 기술의 성능을 비교하고 가장 높은 정확성을 가진 모델을 사용하여 실제 유행성 독감 발생 징후를 초기에 감지하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 설명하고, 3장에서는 제안하는 모델을 소개한다. 4장에서는 제안된 방법을 사용한 실험 결과를 제시하며, 5장에서는 결과에 대한 토의와 결론, 향후 연구 목표를 제시한다.

Ⅱ. 관련 연구

2.1 인플루엔자 발생 예측 모델

최근 인플루엔자의 발생 징후를 초기에 감지하기 위한 다양한 접근법이 제안되고 있다.

Ginsberg 등[7]은 최근 인플루엔자를 예측하기 위해 구글의 검색어 정보와 미국의 CDC(Centers for Disease Control and Prevention)에서 제공하는 ILI 데이터를 사용한 선형 회귀 모델을 제안하였다. 회귀

모델에 사용된 검색어 정보는 각 단일 지역에 연관성이 높은 단어들로 45개의 검색어가 선택되었다. 제안된 모델이 예측한 ILI 값은 미국의 9개 지역에서 실제 ILI 값과 평균적으로 0.97의 상관 계수를 갖는 높은 정확성을 보였다(최소 0.92, 최대 0.99).

Kwon 등[3]은 국내의 검색 엔진인 네이버의 PC 및 모바일 기반 검색어 정보를 사용하여 인플루엔자의 발생을 예측하는 연구를 진행하였으며 로지스틱 회귀 모델과 다중 회귀 모델을 사용했을 때 정확도를 비교하여 모바일 기반 다중 회귀 모델이 0.93의 결정 계수로 다른 모델에 비해 더 높은 정확도를 갖는다는 것을 발견했다. 또한 핵심어와 부가어를 조합하는 방법을 통해 적은 단어의 수로 다수의 검색 키워드를 생성하는 방식을 사용했다.

Paul 등[14]은 특정 검색 엔진의 검색어 정보를 기반으로 한 접근법들의 데이터의 신뢰성과 데이터 간의 의존성에 대한 문제[17]를 해결하기 위해 데이터의 투명성, 사용 편의성에 있어 검색어 정보보다 이점이 있는 트위터 데이터를 사용한 회귀 모델을 제안하였고 트위터를 사용한 모델이 예측 오류를 감소시킬 수 있음을 밝혔다. 실험 결과 과거 데이터만 사용하는 경우보다 약 17~30%의 예측 오류를 감소시켰으며 Google Flu Trends를 사용하는 모델보다 정확성이 높다는 것을 확인했다.

Kim 등[16]은 한글을 사용한 국내 트위터 사용자의 트윗을 사용하여 예측 모델을 구현하기 위해 인플루엔자와 연관된 단어들이 포함된 트윗들을 수집하고 LASSO 방법을 활용하여 500개의 연관된 단어들에 계수를 할당하여 ILI와 상관성이 높은 40개의 단어를 선택하였고 소량의 데이터에서도 데이터 복제를 통해 인플루엔자의 초기 발생을 효과적으로 탐지할 수 있는 방법을 제안했다.

2.2 단어 임베딩

단어 임베딩은 문서 내 단어들의 의미론적 유사성을 고려하여 벡터 값을 지정하기 위한 자연어 처리 기술이다.

Mikolov 등[8]은 문맥에서 유사한 의미를 갖는 단어들은 가까운 거리를 갖는다는 가정[19]을 기반으로 특정 단어와 일정 거리 내에 근접한 단어들이 해당 단어와 동시에 등장할 확률을 학습하여 단어의 벡터 값을 조정하는 Word2vec 모델을 제안하였다. Word2vec의 학습 방법은 Skip-gram과 CBOW(Continuous Bag-of-Words)가 있으며 Skip-gram은 목적 단어로부터 주변 단어의 출현 확률을 통해 벡터 값을 조정하고 CBOW는 목적 단어의 주변 단어로부터 목적 단어의 출현 확률을 통해 벡터 값을 조정한다.

Pennington 등[11]은 Word2vec의 학습 방법이 문맥의 단어들 간의 지역적인 정보를 사용하지만 문서 전체의 전역적인 정보를 사용하지 않는 다는 점을 개선하기 위해 특정 단어가 문서 전체에서 다른 단어와 함께 등장한 비율을 활용하여 각 단어의 벡터 값을 조정하는 GloVe 모델을 제안하였다.

Joulin 등[10]은 Word2vec을 사용한 자연어 처리의 성능은 우수하지만 학습 및 검증 시간이 많이 소모되어 큰 데이터에서의 사용이 제한되는 점과 학습하지 않은 단어에 대응할 수 없는 점을 개선한 Fasttext 모델을 제안했다. Fasttext는 Word2vec과 동일한 Skip-gram, CBOW 학습 방식을 사용하며 단어를 n개의 글자 단위로 분리하여(n-gram) 단어에 포함된 내부 단어의 의미를 학습하기 때문에 학습하지 않은 단어나 오타가 포함된 단어의 경우에도 학습된 내부 단어를 사용하여 벡터 값을 생성한다.

Ⅲ. 실험 방법

3.1 데이터 수집

3.1.1 국내 인플루엔자 환자 데이터

질병관리본부는 국내의 인플루엔자 환자 발생 통계를 일주일 단위로 제공하고 있다. 제공되는 정보는 일주일 동안의 인플루엔자 의사환자 분율이며 이것은 기간 내 인플루엔자 의사환자 수를 기간 내 총 진료환자 수로 나눈 것을 의미한다. 집계된 각 주의 통계 자료는 약 일주일 후 제공된다. 본 실험은 질병관리본부에서 제공한 2017년 36주차부터 2019년 8주차까지, 총 77주 동안의 ILI데이터를 사용한다.

3.1.2 트위터 데이터

트위터는 사람들이 주로 사용하는 SNS 중 하나로서 선택한 기간 동안 특정 단어가 포함된 트윗만을 추출할 수 있는 API를 제공하고 있으며 제공된 트윗을 분석하여 인플루엔자의 발생을 탐지하고 예측하는 연구들에 사용되고 있다[7, 16]. 본 연구는 실험을 위해 질병관리본부의 ILI데이터와 동일한 기간인 2017년 9월 4일부터 2019년 2월 24일까지 트윗을 추출하였으며 질병과 연관된 트윗만을 수집하기 위해 일반 단어(질병, 전염, 바이러스 등) 14개, 증상 관련 단어(발열, 기침, 두통 등) 13개, 법정 감염병명(A형간염, 메르스, 수두 등) 21개로 총 48개 단어가 포함된 트윗만을 수

집하였다.

3.2 선형 회귀 모델

본 실험은 Ginberg 등[7]이 검색어 정보에 적용시킨 선형 모델과 유사한 선형 모델을 사용하여 수행하였으며 적용한 선형 모델은 다음과 같다.

$$logit(I(t)) = \alpha \times logit(Q(W,t)) + \beta + \epsilon \quad (1)$$

I(t)는 특정된 일주일 동안의 ILI를 백분율로 나타낸 것을 의미한다. W는 실험을 위해 선택한 단어들의 집합을 의미하며 $W = \{w_1, w_2, w_3, ..., w_n\}$으로 나타낼 수 있다. Q(W, t)는 특정된 일주일 간의 트윗들 중에서 W에 포함된 각 단어를 포함한 모든 트윗들의 수를 일주일 동안의 전체 트윗 수로 나눈 것을 의미한다. α, β는 회귀 계수이며, ϵ은 오차항을 뜻한다. logit(p)는 로짓 변환을 의미하며 $logit(p) = \ln(p/(1-p))$로 나타낼 수 있다. 이것은 입력 값의 범위가 0과 1사이일 때 출력 값의 범위를 -∞와 ∞사이로 변환한다. [그림 1]은 제안된 회귀 모델이 적용되는 전반적인 시스템 구조를 보여준다.

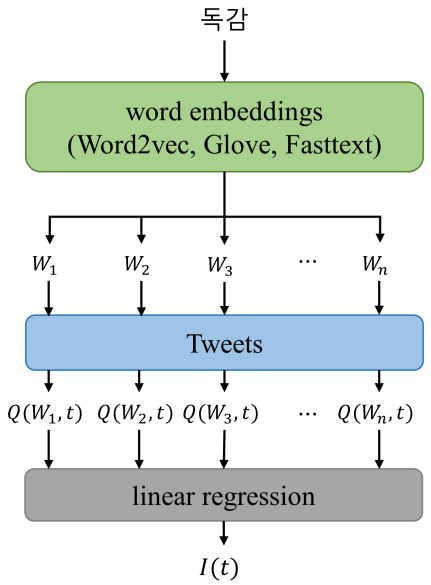


그림 1. 인플루엔자 감지 시스템 구조
Fig. 1. Influenza detection system architecture

3.3 단어 선정

'독감'과 연관된 단어들을 회귀 모델에 적용하기 위해 단어 임베딩을 사용하였다. 단어를 선정하기 위한 과정은 다음과 같다. 먼저, 수집된 트위터는 형태소 분석기인 KoNLPy[18]에서 제공하는 Twitter 품사 태그 클래스를 사용하여 품사를 기준으로 단어들로 나누어지게 된다. 나누어진 단어들은 단어 임베딩 모델에 의해 의미론적 유사성을 반영하여 벡터 값을 갖는다. 각 단어들은 코사인 유사도를 통해 유사성을 확인할 수 있기 때문에 이를 사용하여 '독감'의 벡터 값과 유사한 벡터 값을 가진 단어 100개를 추출한다.

3.4 성능 평가

본 실험은 3가지 단어 임베딩 모델(Word2vec, GloVe, Fasttext)를 사용하여 각 단어 임베딩 모델 별 '독감'과 유사한 100개의 단어를 추출한다. 추출된 단어는 Q(W, t)의 값을 계산하기 위해 사용된다. 또한, 총 77주 동안의 실험 데이터 중 50주 동안의 데이터는 선형 회귀 모델에 적용하기 위한 학습 데이터로 사용되며 남은 27주 동안의 데이터는 적용된 선형 회귀 모델을 검증하기 위한 데이터로 사용된다. 회귀 모델의 정확도는 실제 KCDC의 ILI값과 회귀 모델이 예측한 값에 대해 피어슨 상관 계수를 사용하여 측정하며 회귀 모델이 가장 높은 정확도를 보였을 때 사용된 연관 단어의 개수와 측정된 정확도를 통해 3가지 단어 임베딩의 성능을 평가한다.

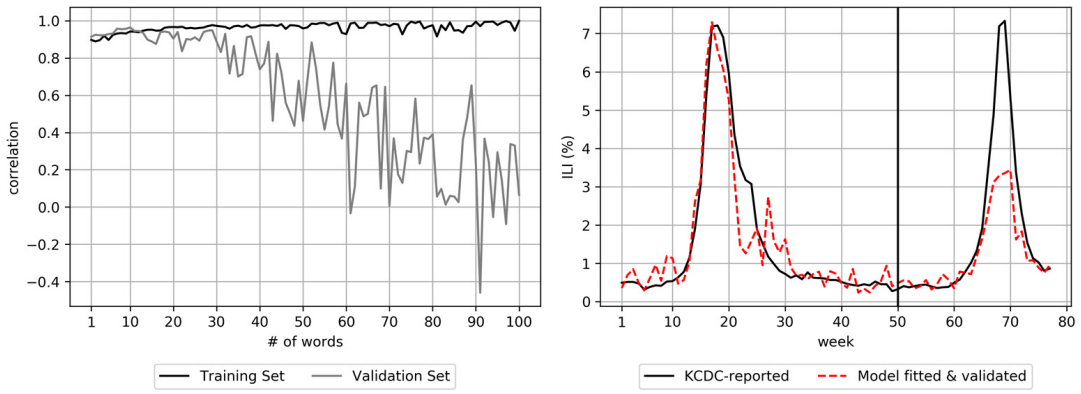
Ⅳ. 실험 결과

4.1 Word2vec

4.1.1 Skip-gram

[그림 2-a]는 Word2vec의 Skip-gram모델을 사용하고 사용된 단어의 개수를 1개부터 100개까지 증가시켜가며 정확도를 측정한 것이다. 학습을 위한 50주차까지의 데이터는 단어의 개수를 증가시킬수록 정확도가 지속적으로 증가하는 모습을 보였다. 그러나 검증을 위한 27주 동안의 데이터는 단어의 개수가 10개일 때 가장 높은 정확도를 보였고 이후로는 단어의 개수가 증가하였음에도 불구하고 정확도는 오히려 감소하는 경향을 보였다.

[그림 2-b]는 [그림 2-a]에서 가장 높은 정확도를 보였던 단어 10개를 사용한 회귀 모델의 예측 값을 나타내는 그래프이다. 50주차에 표시된 검은 선을 기준으로 왼쪽은 학습에 사용된 데이터를 사용한 결과이고 오른쪽은 검증을 위해 학습에 사용되지 않은 데이터를 사용한 결과이다. 학습에 사용된 데이터는 회귀 모델에 과적합되어 높은 정확도를 보이고 있지만 검증을 위한 데이터는 검증 데이터 시기의 가장 높은



(a) (b)

그림 2. (a) Word2vec의 Skip-gram모델에서 단어의 개수 변화에 따른 정확도, (b) Word2vec의 Skip-gram모델에서 가장 높은 정확도를 보인 단어 개수를 사용한 회귀 모델
Fig. 2. (a) the accuracy according to the number of words in skip-gram model of Word2vec, (b) the regression model using word count that showed the highest accuracy in skip-gram model of Word2vec

ILI값을 가진 69주차의 값을 매우 낮게 예측하는 모습을 보였다.

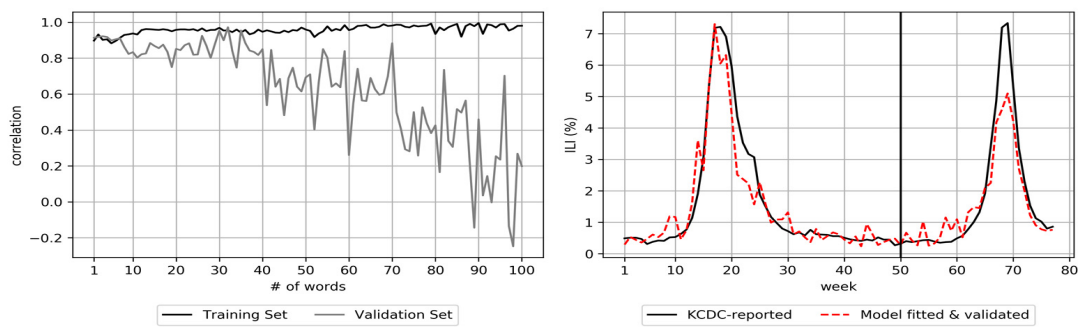
4.1.2 CBOW

[그림 3-a]는 Word2vec의 CBOW모델을 사용하고 사용된 단어의 개수를 1개부터 100개까지 증가시켜가며 정확도를 측정한 것이다. 학습 데이터는 Word2vec의 Skip-gram모델과 마찬가지로 단어가 증가할수록 정확도 또한 지속적으로 증가하는 모습을 보였다. 검증 데이터의 경우 Word2vec의 Skip-gram모델보다 많은 단어 개수인 32개 일 때 가장 높은 정확도를 보였으며 이후 정확도가 감소하는 모습을 보였다.

[그림 3-b]는 [그림 3-a]에서 가장 높은 정확도를 보였던 단어 32개를 사용한 회귀 모델의 예측 값을 나타내는 그래프이다. Word2vec의 Skip-gram모델과 유사하게 학습 데이터에서는 과적합된 모습을 보였으나 검증 데이터에서는 69주차의 높은 ILI값을 더 정확히 예측하는 모습을 보였다.

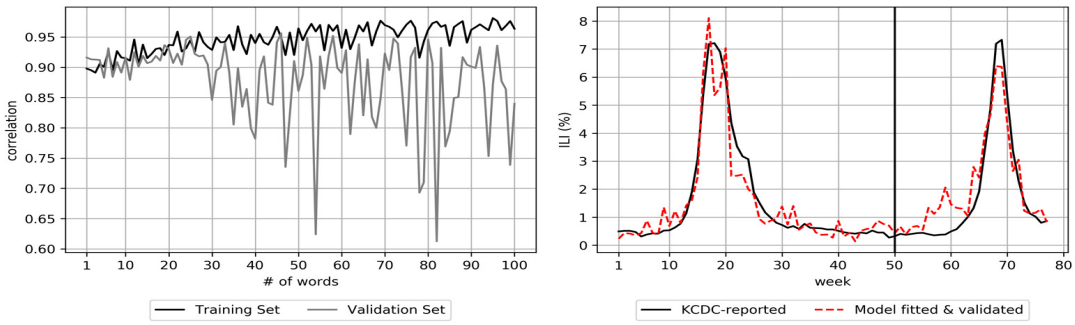
4.2 GloVe

[그림 4-a]는 GloVe모델을 사용하고 사용된 단어의 개수를 1개부터 100개까지 증가시켜가며 정확도를 측정한 것이다. 다른 모델과 마찬가지로 학습데이터는 단어의 개수가 증가할수록 정확도가 증가하는 모습을 보였지만 다른 모델과 다르게 정확도가 큰 폭으로 진동하며 불안정한 정확도를 보였다. 검증데이터 또한 정확도가 큰 폭으로 진동하는 모습을 보였지만 다른 모델과 달리 정확도가 지속적으로 감소하는 모습을



(a) (b)

그림 3. (a) Word2vec의 CBOW모델에서 단어의 개수 변화에 따른 정확도, (b) Word2vec의 CBOW모델에서 가장 높은 정확도를 보인 단어 개수를 사용한 회귀 모델
Fig. 3. (a) the accuracy according to the number of words in CBOW model of Word2vec, (b) the regression model using word count that showed the highest accuracy in CBOW model of Word2vec



(a) (b)

그림 4. (a) GloVe모델에서 단어의 개수 변화에 따른 정확도, (b) GloVe모델에서 가장 높은 정확도를 보인 단어 개수를 사용한 회귀 모델
Fig. 4. (a) the accuracy according to the number of words in GloVe model, (b) the regression model using word count that showed the highest accuracy in GloVe model

나타나지 않았다.

[그림 4-b]는 [그림 4-a]에서 가장 높은 정확도를 보였던 단어 46개를 사용한 회귀 모델의 예측 값을 나타내는 그래프이다. 검증 데이터의 경우 Word2vec의 두 모델에 비해 69주차의 높은 ILI값을 정확하게 예측했지만 55주차부터 63주차 동안의 값을 예측하는데 있어 정확도가 낮았다.

4.3 Fasttext

4.3.1 Skip-gram

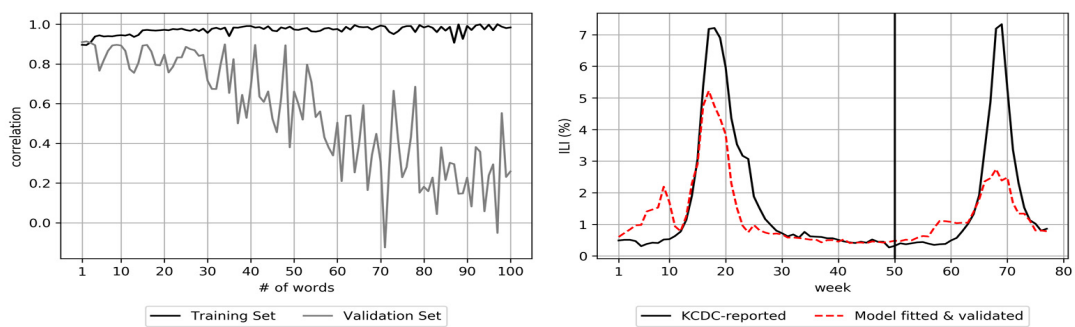
[그림 5-a]는 Fasttext의 Skip-gram모델을 사용하고 사용된 단어의 개수를 1개부터 100개까지 증가시켜가며 정확도를 측정한 것이다. 검증 데이터의 경우 단어를 2개 사용하였을 때 가장 높은 값을 가졌으나 다른 모델과 비교했을 때 가장 낮은 최댓값이었으며 이후로 단어의 개수가 증가했음에도 정확도는 지속적으로 감소하는 모습을 보였다.

[그림 5-b]는 [그림 5-a]에서 가장 높은 정확도를 보였던 단어 2개를 사용한 회귀 모델의 예측 값을 나타내는 그래프이다. 학습 데이터와 검증 데이터 모두 다른 모델에 비해 정확도가 낮은 모습을 보였다.

4.3.2 CBOW

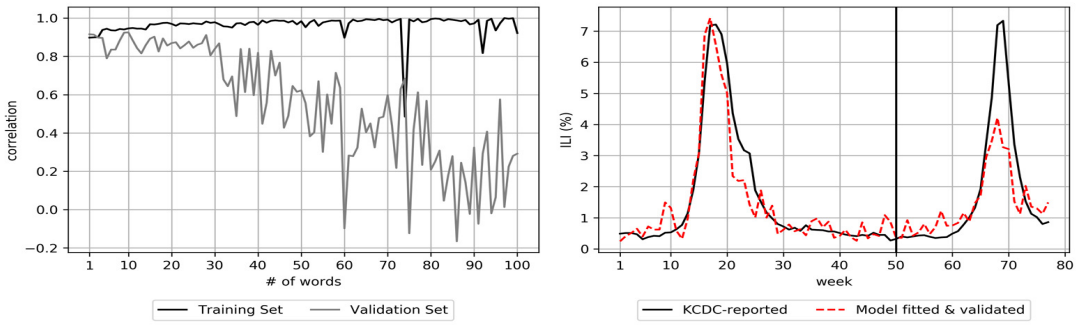
[그림 6-a]는 Fasttext의 CBOW모델을 사용하고 사용된 단어의 개수를 1개부터 100개까지 증가시켜가며 정확도를 측정한 것이다. 실험 데이터의 경우일부 구간에서 급격한 하락을 보였지만 다른 모델과 유사하게 지속적으로 증가하는 모습을 보였으며, 검증 데이터의 경우 단어를 10개 사용했을 때 가장 높은 정확도를 보였고 이후로 감소하는 모습을 보였다.

[그림 6-b]는 [그림 6-a]에서 가장 높은 정확도를



(a) (b)

그림 5. (a) Fasttext의 Skip-gram모델에서 단어의 개수 변화에 따른 정확도, (b) Fasttext의 Skip-gram모델에서 가장 높은 정확도를 보인 단어 개수를 사용한 회귀 모델
Fig. 5. (a) the accuracy according to the number of words in skip-gram model of Fasttext, (b) the regression model using word count that showed the highest accuracy in skip-gram model of Fasttext



(a) (b)

그림 6. (a) Fasttext의 CBOW모델에서 단어의 개수 변화에 따른 정확도, (b) Fasttext의 CBOW모델에서 가장 높은 정확도를 보인 단어 개수를 사용한 회귀 모델
Fig. 6. (a) the accuracy according to the number of words in CBOW model of Fasttext, (b) the regression model using word count that showed the highest accuracy in CBOW model of Fasttext

보였던 단어 10개를 사용한 회귀 모델의 예측 값을 나타내는 그래프이다. Word2vec의 Skip-gram모델과 비슷하게 실험 데이터의 경우 높은 정확도를 보였지만 검증 데이터의 69주차 ILI값을 낮게 예측하는 모습을 보였다.

4.4 단어 임베딩 모델 비교

[표 1]은 각 단어 임베딩 모델 별로 검증 데이터에서 가장 높은 정확도를 보였던 단어의 개수와 상관 계수 값을 표시한 것이다. Word2vec의 CBOW모델은 32개의 단어를 사용하여 약 0.9718의 정확도로 모든 모델에서 가장 높은 정확도를 보였다. 두 번째로 정확도가 높았던 모델은 Word2vec의 Skip-gram이었으며 정확도 약 0.9642로 Word2vec의 CBOW보다 조금 낮았지만 Fasttext를 제외한 나머지 모델에서 가장 적은 단어를 사용하여 높은 정확도를 보였다. Fasttext의 두 모델의 경우 사용한 단어의 개수는 Word2vec의 Skip-gram보다 적은 단어를 사용하였지만 정확도가 약 0.91, 0.92로 낮은 모습을 보였다. GloVe모델은 약 0.9561의 정확도를 보였으나 단어 개수의 증가마다 정확도의 변화 폭이 커서 불안정한 모습을 보였다. 그러나 각 모델 들은 정확도에서 차이를 보였으나 모든 모델은 11주차의 높은 ILI,값을 비교적 정확히 예측한 것과 달리 69주차의 높은 ILI값의 예측 정확도는 낮은 모습을 확인할 수 있었다. [그림 7]은 모든 모델에 사용된 '독감' 단어가 포함된 트윗 비율과 실제 KCDC의 ILI값을 나타낸 그래프이며, 이것을 통해 제안된 모델의 정확도가 연관 단어를 추출하기 위한 핵심 단어의 트윗 비율에 많은 영향을 받는다는 것을 보여준다.

표 1. 단어 임베딩 모델 별 검증 데이터에서 가장 높은 정확도와 사용한 단어 개수
Table 1. The highest accuracy and the number of words used in the validation data by word embedding models

Model	Correlation ratio (# of words)
Word2vec-Skip-gram	0.9642 (10)
Word2vec-CBOW	**0.9718 (32)**
GloVe	0.9561 (46)
Fasttext-Skip-gram	0.9135 (2)
Fasttext-CBOW	0.9251 (9)

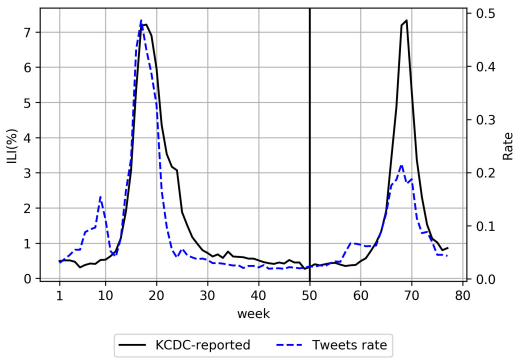


그림 7. KCDC의 ILI값과 '독감' 단어가 포함된 트윗 비율
Fig. 7. ILI of KCDC and the rate of tweets that contain the word 'influenza'

V. 결론 및 토의

본 연구는 트위터와 단어 임베딩을 활용하여 유행성 독감의 발생을 감지하는 방법을 제안하였다. 제안

된 방법은 단어 임베딩을 사용하여 예측 모델의 성능에 큰 영향을 미치는 독감과 연관된 단어들을 자동적으로 추출해 예측 모델의 성능을 향상 시켰고 최신 단어 임베딩 기술들의 성능을 비교하였으며 트위터와 같이 활용하여 유행성 독감을 감지할 때 적합한 단어 임베딩을 찾을 수 있었다. Word2vec의 CBOW모델에서 추출된 단어들을 사용한 회귀 모델의 예측 값이 가장 높은 정확도를 보였고 Word2vec의 Skip-gram은 사용한 단어의 개수 대비 높은 정확도를 보인 것으로 보았을 때 단어 임베딩 기술 중 Word2vec 기술이 트위터에 적용했을 때 높은 성능을 기대할 수 있음을 확인했다. 또한 제안된 단어 임베딩을 사용한 방법은 기존의 다수의 단어를 사용하여 회귀 모델의 정확성을 향상시키는 방법에서 다수의 단어를 빠르게 선별하는 데 있어 효과적으로 사용될 수 있다.

결과적으로 제안된 방법은 기존의 예측 모델을 구현할 때 어려운 점이었던 예측 모델에 사용될 단어의 선택을 자동적으로 할 수 있게 하였으며, 실시간으로 올라오는 트위터를 활용하여 질병관리본부가 ILI 데이터를 제공하기까지 걸리는 약 일주일간의 정보 공백을 최소화할 수 있다. 또한, 질병관리본부는 일주일 단위 데이터를 제공하는 반면 트위터는 하루 단위 데이터를 사용할 수 있다는 것을 고려하면 약 1~2주 먼저 유행성 독감의 발생을 감지할 수 있다고 판단된다.

그러나, 본 연구에서 단어 임베딩을 사용해 트위터에서 추출한 '독감'과 연관된 단어들은 독감과 연관성이 없어 회귀 모델의 정확성을 크게 감소시키는 단어나 '독감에', '독감도' 같이 '독감'에 조사가 붙어 형태소를 완전히 분리하지 못한 단어도 포함되어있다. 트위터는 형식이 없고 문법을 정확히 지키지 않는 문장이 많기 때문에 형태소 분석의 정확도가 떨어진 것으로 판단된다. 또한 '독감'과 연관된 단어들을 사용한 회귀 모델은 각 단어 임베딩 모델 별로 다른 단어를 추출했지만 공통으로 포함된 '독감' 단어의 영향을 받아 모든 모델에서 동일하게 예측 정확도의 하락을 보이는 구간이 존재했다. 추가적으로 트위터의 광고성 트윗과 핵심 단어를 포함하지만 실제 연관성이 없는 트윗은 단어 임베딩의 학습을 저해한다. 따라서 향후 연구로 형태소 분석의 정확도를 향상시키기 위해 트위터 문장에서 불필요한 부분을 제거하고 '독감'과 같은 핵심 단어에 영향을 과도하게 받지 않기 위해 각 단어의 가중치를 조정하여 예측 회귀모델을 구현하는 것과 광고성 혹은 연관성이 적은 트윗을 제거하는 전처리 과정을 추가하는 연구를 진행하고자 한다.

References

[1] World Health Organiztion, *Influenza fact sheet*(2018), Jun. 14, 2019, from https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)

[2] S. Aslam, *Twitter by the numbers: Stats, demographics & fun facts*(2018), Jun. 15, 2019, from https://www.omnicoreagency.com/twitter-statistics/

[3] C. M. Kwon, S.-W. Hwang, and J. U. Jung, "Monitoring seasonal influenza epidemics in korea through query search," *J. Korea Soc. Simulation*, vol. 23, no. 4, pp. 31-39, Dec. 2014.

[4] Y. Wang, et al., "A comparison of word embeddings for the biomedical natural language processing," *J. Biomed. Informatics*, vol. 87, pp. 12-20, Jul. 2018.

[5] V. Lampos, et al., "Advances in nowcasting influenza-like illness rates using search query logs," *Scientific Reports*, vol. 5, p. 12760, Aug. 2015.

[6] M. R. Seok, et al., "Comparison of NER performance using word embeddings," in *The 4th Int. Conf. Artificial Intell. and Appl.,* pp. 754-88, Melbourne, Australia, Feb. 2018.

[7] J. Ginsberg, et al., "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, p. 1012, Feb. 2009.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Wrkshps. 2013,* Scottsdale, Arizona, May 2013.

[9] S. Okura, et al., "Embedding-based news recommendation for millions of users," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Halifax, NS, Canada, 2017.

[10] A. Joulin, et al., "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter of the Assoc. for Computat. Linguistics,* vol. 2, Short Papers, Valencia, Spain, 2017.

[11] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word

representation," in *Proc. EMNLP*, pp. 1532-1543, Doha, Qatar, Oct. 2014.

[12] C. Li, et al., "Topic modeling for short texts with auxiliary word embeddings," in *Proc. 39th Int. ACM SIGIR Conf. Res. and Development in Inf. Retrieval.*, Pisa, Italy, 2016.

[13] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proc. First Wrkshps. Soc. Media Analytics,* Washington DC, USA, 2010.

[14] M. J. Paul, M. Dredze, and D. Broniatowski, "Twitter improves influenza forecasting," *PLoS Curr.*, vol. 6, Oct. 2014.

[15] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proc. 38th Int. ACM SIGIR Conf. Res. and Development in Inf. Retrieval*, Santiago, Chile, 2015.

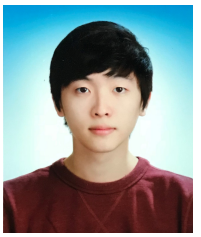
[16] E.-K. Kim, et al., "Use of hangeul twitter to track and predict human influenza infection," *PloS one*, vol. 8, no. 7, p. e69305, Jul. 2013.

[17] D. Lazer, et al., "The parable of Google Flu: traps in big data analysis," *Sci.*, vol. 343, no. 6176, pp. 1203-1205, Mar. 2014.

[18] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python," in *Proc. 26th Annu. Conf. Human & Cognitive Language Technol.*, pp. 133-136, Chun Cheon, Korea, 2014.

[19] M. Sahlgren, "The distributional hypothesis," *Italian J. Disability Stud.*, vol. 20, pp. 33-53, 2008.

김 인 환 (Inhwan Kim)



2016년 3월~현재 : 상명대학교 컴퓨터과학과 학석사 연계과정

<관심분야> 인공지능, 빅데이터, 자연어 처리

[ORCID:0000-0002-2621-386X]

장 백 철 (Beakcheol Jang)

2001년 2월 : 연세대학교 컴퓨터과학과 학사

2002년 8월 : 한국과학기술원 컴퓨터과학과 석사

2009년 8월 : North Carolina State University, 컴퓨터과학과 박사

2012년~현재 : 상명대학교 컴퓨터과학과 교수

<관심분야> 무선 네트워크, 사물 인터넷, 빅데이터, 인공지능

[ORCID:0000-0002-3911-5935]