문장 벡터와 전방향 신경망을 이용한 스팸 문자 필터링

이현영○, 강승식
국민대학교 소프트웨어학부
le32146@gmail.com, sskang@kookmin.ac.kr

Spam Text Filtering by Using Sen2Vec and Feedforward Neural Network

Hyun-Young Lee○, Seung-Shik Kang
School of Software, Kookmin University

요 약

스팸 문자 메시지를 표현하는 한국어의 단어 구성이나 패턴은 점점 더 지능화되고 다양해지고 있다. 본 논문에서는 이러한 한국어 문자 메시지에 대해 단어 임베딩 기법으로 문장 벡터를 구성하여 인공신경망의 일종인 전방향 신경망(Feedforward Neural Network)을 이용한 스팸 문자 메시지 필터링 방법을 제안한다. 전방향 신경망을 이용한 방법의 성능을 평가하기 위하여 기존의 스팸 문자 메시지 필터링에 보편적으로 사용되고 있는 SVM light를 이용한 스팸 문자 메시지 필터링의 정확도를 비교하였다. 학습 및 성능 평가를 위하여 약 10만 개의 SMS 문자 데이터로 학습을 진행하였고, 약 1만 개의 실험 데이터에 대하여 스팸 문자 필터링의 정확도를 평가하였다.

주제어: 스팸 문자 필터링, 단어 임베딩, 문장 벡터, 전방향 신경망

1. 서론

스마트폰은 사용자의 생활을 편리하고 윤택하게 만들어주는 만큼 빠른 대중화를 이루었다. 이에 반해 스마트폰 대중화에 의한 스팸 문자 메시지 양도 폭발적으로 증가하는 추세이다. 그러한 스팸 문자 메시지의 내용은 성인광고, 대출광고, 게임광고 등이 주를 이루며, 수신자로 하여금 불쾌감을 유발하고 불편을 가중시킨다[1,2]. 스팸 문자 메시지를 접하는 사용자가 늘어나는 만큼 사용자들은 스팸 문자 메시지를 통한 소액결제 및 개인 정보 유출 등에 쉽게 노출이 된다. 이에 따라 스팸 문자 메시지 필터링의 중요성은 점점 커지고 있다.

기존의 스팸 문자 메시지를 자동 차단하는 방식은 크게 '단어 기반 사전을 통한 키워드 매칭' 방식과 '나이브 베이지안(Naive Bayesian), SVM(Support Vector Machine)' 등을 이용한 기계학습 방식으로 구분할 수 있다. 단어 기반 사전을 통한 키워드 매칭 방식은 구현하기가 쉽고 컴퓨터 자원 소모가 적지만 이를 통한 스팸 문자 메시지 필터링은 사용자가 스팸 번호, 스팸 단어 등을 직접 입력해야 하므로 사용자의 편의성이 낮다[3].

스팸 단어의 경우에는 인위적으로 조작되는 경우도 존재한다. 예를 들어, "경마"라는 단어는 "경ㅁr", "야동"이란 단어는 "ㅇk동", "0ㅑ동"으로 표현할 수 있다. 이렇게 의도적으로 조작할 수 있는 단어의 가지 수는 수없이 많기 때문에, 단어 기반 사전을 통한 스팸 문자 메시지 필터링 방식은 효율성이 떨어진다[4,5].

또 다른 스팸 문자 메시지 필터링 방식 중 하나인 기계학습 방법으로는 나이브 베이지안, SVM(Support Vector Machine) 등이 있으며, 이를 사용하기 위해서는 스팸 문자 메시지와 햄 문자 메시지(스팸이 아닌 문자 메시지)를 샘플 데이터로 하여 해당 문자 메시지를 구분할 수 있는 특징 벡터(Feature Vector)를 생성해야 한다. 예를 들어, 특징 벡터를 생성하고자 할 때, 일반적으로 추출하는 특징으로는 특수문자의 빈도, 반복되는 단어의 빈도, 명사의 빈도, 품사 태깅(POS-tagging) 등의 특징을 바탕으로 해당 문자 메시지를 표현하는 특징 벡터를 만든 후 기계학습 방법을 이용하여 스팸 문자 메시지를 필터링한다.

이러한 특징 벡터는 해당 문자 메시지를 표현하는 특징의 종류가 다양할수록 차원 수는 증가하고, 벡터는 희소한(sparse) 형태가 된다[6]. 그림 1을 통해 희소한 형태의 특징 벡터를 확인할 수 있다.

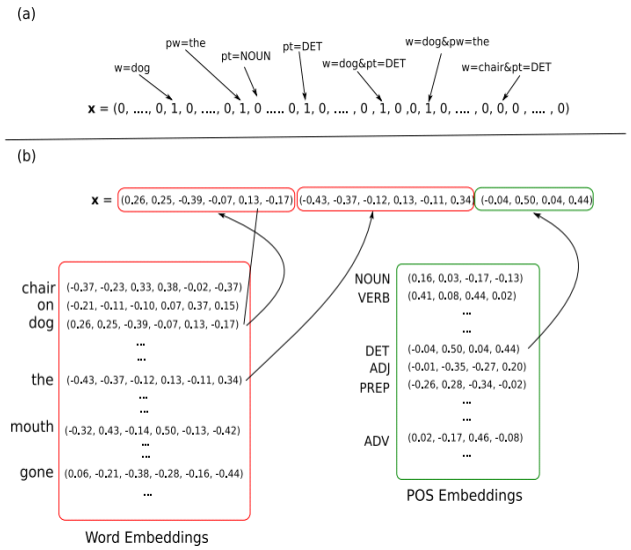

그림 1. 희소한 특징 벡터(위)와 밀집한 특징 벡터(아래)

본 논문에서는 밀집한 특징 벡터 생성을 위해 신경망

언어 모델의 단어 임베딩(Word Embedding)을 이용하여 단어 벡터를 생성하고 이를 기반으로 문자 메시지의 문장 벡터(Sentence Vector)를 생성한다. 그리고 그 문장 벡터와 인공신경망(Artificial Neural Network)의 일종인 전방향 신경망을 이용하여 스팸 문자 메시지를 필터링하는 방법을 제안한다.

2. 단어 벡터와 단어 임베딩

인공신경망의 일종인 딥러닝(Deep Learning)은 컴퓨터 비전, 패턴 인식 그리고 음성 인식 분야에서 뛰어난 성과를 보여주고 있다. 그리고 자연어 처리에서도 그림 2와 같이 딥러닝을 활용한 연구는 빠르게 증가하는 추세이다[7].

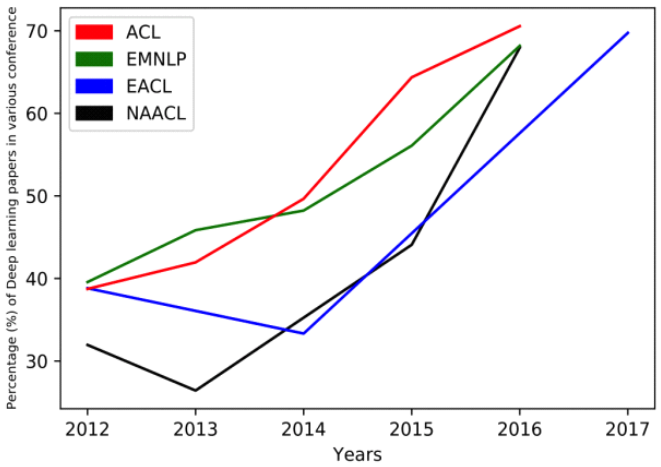

그림 2. 딥러닝 관련 자연어 처리 논문 증가 추세

그림 1과 같이 통계적 기법에서 사용하는 단어 임베딩은 희소한 형태의 단어 벡터를 생성한다. 이는 자연어 처리에서 말하는 차원의 저주인 차원 수 문제와 연산속도 및 메모리 부분에서 효율적이지 못하다.

하지만 인공신경망을 이용한 단어 임베딩은 단어 벡터의 차원 축소 및 확대에 있어서 자유롭고, 연산속도 및 메모리 공간 활용에서도 효율적이다. 이뿐만 아니라 단어 간의 유사도 분석과 단어 사이의 의미적 관계 분석에서도 우수한 효과를 보여 준다.

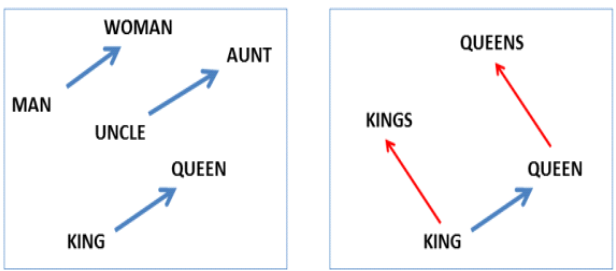

그림 3. 단어 공간에서 단어 벡터 간의 상관관계

그림 3에서 <man-woman>은 man과 woman이라는 단어의 차이 벡터인

word_vector(“man”) - word_vector(“woman”)

을 말한다. <man-woman>의 차이 벡터와 유사한 벡터는 다음과 같다.

word_vector(“uncle”) - word_vector(“aunt”)
word_vector(“king”) - word_vector(“queen”)

즉, 단어 임베딩을 통한 단어 벡터 공간에 각 단어 벡터는 성별이라는 특성을 포함하고 있어, <man-woman>의 차이 벡터는 두 단어의 성별 차이를 나타낸다. 또한 <uncle-aunt>, <king-queen>라는 두 개의 차이 벡터는 <man-woman>의 성별 차이와 유사함을 보여준다. 이러한 인공신경망을 통한 단어 임베딩은 성별, 복수 명사, 단수 명사, 시제 등과 같은 문법적, 의미적 특성을 내포하는 단어 벡터를 생성한다. 이러한 단어 벡터 효과는 자연어 처리 연구에서 우수한 효과를 보여준다[8,9,10].

인공신경망의 구조 중 하나인 전방향 신경망은 선형적 특성과 비선형적 특성으로 이루어진다. 식 (1)은 전방향 신경망의 선형적 특성을 포함하는 식이고, 식 (2)는 식 (1)을 활성화하는 식으로 비선형적 특성을 포함한다. 전방향 신경망 구조는 이 두 식의 조합을 통해 비선형적 특성을 활용한 분류를 가능하게 한다.

F(x) = Wx + b (1)
G(x) = activation(F(x)) (2)

본 논문에서는 단어 임베딩을 통해 생성한 단어 벡터를 이용하여 문장 벡터를 생성하고, 전방향 신경망을 통해 스팸 문자 메시지 필터링 문제에 비선형 분류 기법을 적용하였다.

3. 전방향 신경망을 이용한 스팸 문자 필터링

본 논문에서 제안하는 스팸 문자 메시지 필터링 방법은 전처리 과정으로 자동 띄어쓰기를 한 후, 단어 임베딩을 통해 단어 벡터를 생성하고 문장들을 구성하는 단어의 벡터 합으로 문장 벡터를 생성한다. 최종적으로는 문장 벡터와 전방향 신경망을 이용하여 그 문장이 스팸 문자 메시지인지 아닌지를 필터링하는 방법을 제안한다.

단어 벡터 토큰을 생성할 때 구분자는 공백 문자로 한다. 즉, “01야ㄱ1 사과” 라는 문장에서 단어 벡터 토큰은 각각 “01야ㄱ1”, “사과” 라는 두 개의 토큰으로 나누어진다. 단어 벡터 토큰 생성 시에는 공백 문자만을 구분자로 하여 특수기호, 숫자 등을 이용하여 의도적으로 변환한 “사♥랑”, “♥축하”, “경★ㅁr”, “ㅇk동”, “0ㅏ동” 과 같은 단어 패턴 변화에도 단어 임베딩을 적용하여 단어 벡터를 생성하였다.

전체적인 스팸 문자 메시지 필터링 과정은 그림 4와 같다.

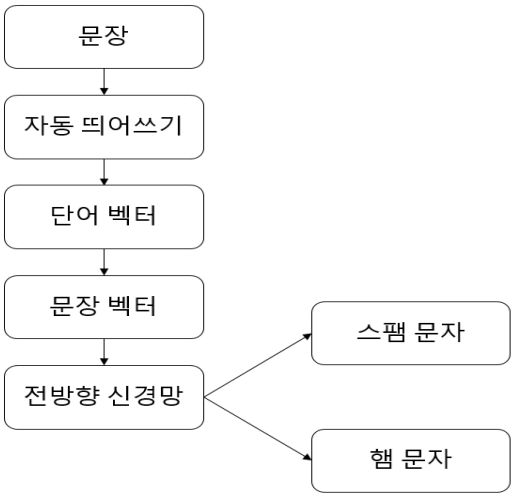

그림 4. 스팸 문자 메시지 필터링 과정

3.1 자동 띄어쓰기

문장을 구성하는 단어 패턴은 다양한 구성을 보여준다. 즉, 하나의 문장은 숫자, 특수문자, 한글, 영어 등의 복합적인 구성이다. 본 논문에서는 복합적인 단어 패턴에서 공백 문자를 구분자로 하여 단어 벡터 토큰을 생성하였다. 하지만 문자 메시지에서 일반 사용자들은 단어를 띄어쓰기하지 않고 한 줄에 연이어 문장을 작성하는 경우가 여러 존재하였다. 이러한 경우에 공백 문자를 기준으로 단어 벡터 토큰을 선택하려 할 때, 문제가 발생하였다. 즉, 띄어쓰기를 하지 않은 긴 길이의 문장도 하나의 단어로 처리하게 되고 이는 희소한 단어를 이용하여 의미 없는 단어 벡터를 생성하게 된다. 이와 같은 문제를 해결하기 위하여 의미있는 단어 벡터를 생성하기 위한 전처리 과정으로 자동 띄어쓰기를 적용하였다.

sentence = “좋은밤되세요내용없음”
word_vector(“좋은밤되세요내용없음”)

자동 띄어쓰기를 적용한 후, 문장은 다음과 같다.

sentence = “좋은 밤 되세요 내용 없음”

자동 띄어쓰기를 적용한 문장의 단어 벡터는 다음과 같다.

word_vector1(“좋은”), word_vector2(“밤”),
word_vector3(“되세요”), word_vector4(“내용”),
word_vector5(“없음”)

3.2 단어 벡터와 문장 벡터

단어 벡터는 신경망 언어 모델 중 하나인 CBOW(Continuous Bag-of-Words) 모델을 기반으로 생성하였다. 그림 5와 같이, CBOW는 단어 W(t)를 중심으로 오른쪽, 왼쪽에 있는 단어를 윈도우 크기만큼 이용하여 중심단어 W(t)의 단어 벡터를 생성한다. 그림 5는 윈도우 크기를 2로 하여, 단어 W(t)를 중심으로 오른쪽 2개의 단어, 왼쪽 2개의 단어를 이용하여 중심단어 W(t)의 단어 벡터를 생성하는 것을 보여준다.

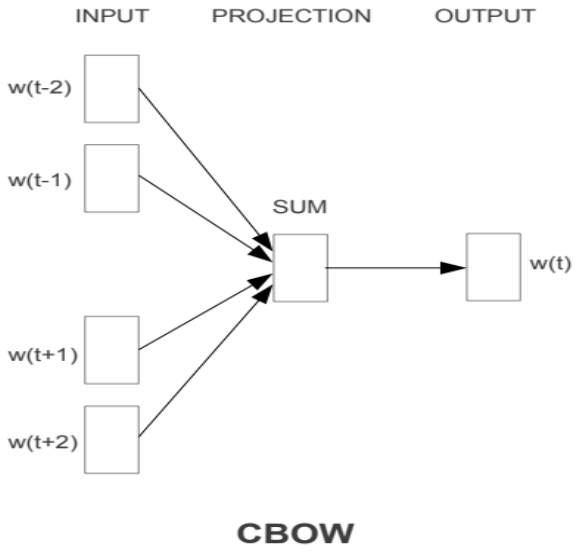

그림 5. CBOW(Continuous Bag-of-Words)

예를 들어 “한글 및 한국어 정보처리 학술대회”라는 문장에서 그림 5와 같이 윈도우 크기는 2이고 CBOW를 이용한다면 “한국어” 단어 벡터는 “한국어”의 오른쪽 2개의 단어(“한글”, “ 및”), 왼쪽 2개의 단어(“정보처리”, “학술대회”)를 이용하여 “한국어” 단어 벡터를 생성한다.

본 논문에서는 단어 벡터 생성을 위해 학습 데이터와 평가 데이터의 총 11만 문장에 포함된 단어들을 기반으로 윈도우 크기가 8인 CBOW를 통해 단어 벡터를 생성하였고, 이 단어 벡터들을 문장 벡터로 합성하였다.

예를 들어, “한글 및 한국어 정보처리 학술대회”의 문장 벡터를 생성하는 과정은 다음과 같다.

x = “한글 및 한국어 정보처리 학술대회”
word1 = word_vector(“한글”),
word2 = word_vector(“및”)
word3 = word_vector(“한국어”)
word4 = word_vector(“정보처리”)
word5 = word_vector(“학술대회”)

Sen2Vec(x) = word1 + word2 + word3 + word4 + word5

3.3 스팸 문자 메시지 필터링을 위한 신경망 구조

전방향 신경망은 완전히 연결된 신경망(FNN: Fully connected Neural Network)으로 계층적 구조로 객체를 분류하는 문제를 해결하는데 사용하는 기본적인 신경망 구조이다. 본 논문에서 사용한 전방향 신경망의 구조는 그림 6과 같다.

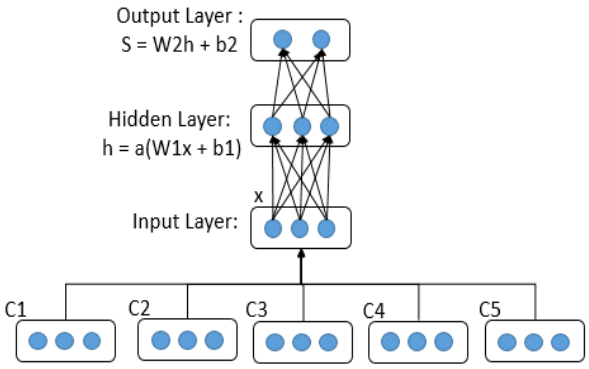

그림 6. 스팸 문자 필터링을 위한 전방향 신경망 구조

이 신경망 구조의 특징은 은닉층(Hidden Layer) 수, 은닉층의 뉴런 개수 조절이 자유롭고, 이를 조절하여 객체 분류 정확도를 개선한다. 그리고 최종 출력(Output)의 값을 이용하여 비용 함수(Cost Function)를 평가하고 객체 분류를 수행한다. 최적의 분류를 수행하기 위해 신경망 구조 기계 학습 모델은 역전파(Backpropagation) 알고리즘을 이용한다. 이 알고리즘은 비용함수의 값을 최소화하는 방향으로 신경망의 인수(W, b)를 학습한다.

본 논문에서 사용한 비용 함수는 크로스 엔트로피(Cross Entropy), 그리고 역전파 알고리즘으로는 경사하강법(Gradient Descent)을 사용하였다.

4. 실험 및 결과

본 논문은 기존의 기계학습 모델인 SVM light와 인공신경망의 일종인 전방향 신경망을 이용한 실험에서 학습 데이터 및 평가데이터, 단어 벡터, 문장 벡터의 차원 수를 표1과 같이 동일하게 사용하였다.

표 1. 데이터 기본 설정

학습데이터	spam	약 5만 문장
	ham	약 5만 문장
평가데이터	spam	약 5천 문장
	ham	약 5천 문장
단어 벡터의 차원수		300 차원
문장 벡터의 차원수		300 차원

*총 어절 수: 약 910,000어절
*한 문장 당 평균 어절 수 : 8.27

표 2는 학습 데이터와 평가 데이터의 총 합인 11만 문장을 기반으로 단어 벡터를 생성하고, 문장을 구성하는 단어의 벡터 합을 통해 생성된 문자 메시지의 문장 벡터를 SVM light와 전방향 신경망을 사용하여 스팸 문자 메시지인지 아닌지를 필터링한 정확도를 보여주고 있다.

기존의 SVM light를 이용한 이진 분류(Binary Classification)의 정확도 보다 전방향 신경망을 통한 이진 분류가 높은 정확도를 보여준다.

또한, 신경망을 이용한 분류 정확도는 은닉층의 수를 증가함으로써 정확도가 개선되는 것을 확인하였다. 하지만 은닉층의 수가 증가하는 만큼 정확도의 증가 폭도 비례하게 증가하기 보다는 감소하였다. 즉 다시 말해, 은닉층의 수가 1에서 2로 증가함에 따라, 정확도의 증가 폭은 1.2로 나오는 반면에 은닉층 수가 2에서 3으로 증가 시, 정확도의 증가 폭은 오히려 0.58로 감소하였다.

표 2. 실험 결과

		정확도			
SVM light		95.25 %			
	활성화 함수	비용 함수	최적화 알고리즘	층수	정확도
전방향 신경망	Sig	크로스 엔트로피	경사 하강법	2	93.72%
				3	94.92%
				4	95.50%

*Sig : Sigmoid Function
*층수: 은닉층 + 출력층

5. 결론

본 논문은 신경망 언어 모델 중 하나인 CBOW를 한국어에 적용하여 생성된 단어 벡터를 문장 벡터로 합성하고, 전방향 신경망을 이용한 스팸 문자 필터링 방법을 제안하였다. 실험 결과를 통해 이진 분류 기능에서 신경망을 활용한 방법이 기존의 SVM light보다 우수할 수 있음을 보여주었다.

신경망 구조에서는 은닉층 수에 따라 정확도를 향상할 수 있음을 확인하였다. 하지만, 은닉층의 수가 증가하는 폭만큼 정확도가 비례하여 증가하지 않았음을 보여주었다. 이를 통해, 전방향 신경망을 이용한 스팸 문자 필터링에 효율적인 은닉 층의 수를 계산하는 방안이 필요할 것으로 생각된다.

본 논문에서는 단어 임베딩을 통해 문장 벡터를 생성하고 전방향 신경망으로 스팸 문자 필터링 방법을 제안했지만, 정확도 향상을 위해 다양한 단어 임베딩(skip-gram, GloVe)을 통한 단어 벡터 생성과 CNN(Convolution Neural Network)을 이용한 문장 벡터 생성 등 다양한 시도를 통해 스팸 문자 필터링의 정확도를 분석하는 시도가 필요할 것으로 생각된다[10].

참고문헌

[1] 강승식, “메일 주소 유효성과 제목-내용 가중치 기법에 의한 스팸 메일 필터링”, 멀티미디어학회 논문지, Vol.9, No.2, pp.255-263, 2006.
[2] 손대능, 이정태, 이승욱, 신중휘, 임해창, “문자 메시지의 특성을 고려한 한국어 모바일 스팸필터링 시스템”, 한국산학기술학회논문지, 제11권, 제7호, pp.2595-2602, 2010.
[3] 이승재, 최덕재, “사용자 맞춤형 스팸 문자 필터링 시스템”, Vol.11, No12, 2011.

[4] 강승식, 장두성, “SMS 변형된 문자열의 자동 오류 교정 시스템”, 정보과학회논문지, 35권, 6호, pp.386-391, 2008.
[5] 김성윤, 차태수, 박제원, 최재현, 이남용, “통계적 기법을 이용한 스팸메시지 필터링 기법”, 한국IT서비스학회지, 제 13권, 제3호, pp. 299-308, 2014.
[6] Goldberg, Yoav. “A Primer on Neural Network Models for Natural Language Processing." Journal of Artificial Intelligence Research(JAIR) 57, pp.345-420, 2016.
[7] Young, T., Hazarika, D., Poria, S., & Cambria, E., “Recent Trends in Deep Learning Based Natural Language Processing,” arXiv preprint arXiv:1708.02709, 2017.
[8] Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. “Linguistic Regularities in Continuous Space Word Representations,” Hlt-Naacl. Vol.13. 2013.
[9] Mikolov, Tomas, et al. “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013.
[10] Pennington, Jeffrey, Richard Socher, and Christopher Manning, “Glove: Global Vectors for Word Representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1532-1543 2014.