

워드 임베딩과 딥러닝 기법을 이용한 SMS 문자 메시지 필터링

(SMS Text Messages Filtering using Word Embedding and Deep Learning Techniques)

이현영*, 강승식**

(Hyun Young Lee, Seung Shik Kang)

요약

딥러닝에서 자연어 처리를 위한 텍스트 분석 기법은 워드 임베딩을 통해 단어를 벡터 형태로 표현한다. 본 논문에서는 워드 임베딩 기법과 딥러닝 기법을 이용하여 SMS 문자 메시지를 문서 벡터로 구성하고 이를 스팸 문자 메시지와 정상적인 문자 메시지로 분류하는 방법을 제안하였다. 유사한 문맥을 가진 단어들은 벡터 공간에서 인접한 벡터 공간에 표현되도록 하기 위해 전처리 과정으로 자동 띄어쓰기를 적용하고 스팸 문자 메시지로 차단되는 것을 피하기 위한 목적으로 음절의 자모를 특수기호로 왜곡하여 맞춤법이 파괴된 상태로 단어 벡터와 문장 벡터를 생성하였다. 또한 문장 벡터 생성 시 CBOW와 skip gram이라는 두 가지 워드 임베딩 알고리즘을 적용하여 문장 벡터를 표현하였으며, 딥러닝을 이용한 스팸 문자 메시지 필터링의 성능 평가를 위해 SVM Light와 정확도를 비교 측정하였다.

■ 중심어 : 스팸 문자 메시지 ; 워드 임베딩 ; 문장 벡터 ; 딥러닝 ; 이진 분류

Abstract

Text analysis technique for natural language processing in deep learning represents words in vector form through word embedding. In this paper, we propose a method of constructing a document vector and classifying it into spam and normal text message, using word embedding and deep learning method.

Automatic spacing applied in the preprocessing process ensures that words with similar context are adjacently represented in vector space. Additionally, the intentional word formation errors with non-alphabetic or extraordinary characters are designed to avoid being blocked by spam message filter.

Two embedding algorithms, CBOW and skip grams, are used to produce the sentence vector and the performance and the accuracy of deep learning based spam filter model are measured by comparing to those of SVM Light.

■ keywords : spam text message ; word embedding ; text vector ; deep learning ; binary classification

I. 서론

스마트폰이 대중화되면서 SMS(short message service) 문자 메시지 전송을 통한 의사소통 방식이 보편화됨에 따라 텍스트 형태의 데이터의 양도 대량으로 증가하고 있다[1]. 그 부작용으로 성인광고, 대출광고, 게임광고 등의 스팸 문자 텍스트 데이터도 폭발적으로 증가하여 이를 필터링 하기 위한 다양한 기법들이 개발되고 있다[2]. 이러한 스팸 문자 메시지 데이터를 효율적으로 필터링하는 방법 중 널리 보편적으로 사용되는 방법론으로는 통계적 확률 방법론과 지지 벡터 기계(SVM) 등의 기계학습 방법이 있다[3, 4].

기계학습 방법론 중, 지지 벡터 기계는 스팸 문자 메시지 데

이터를 분류하는 것에 있어서 다른 기법들보다 더 나은 성능을 보인다고 알려져 있다[5, 6]. 지지 벡터 기계는 자질 벡터(feature vector)를 이용하는 기계학습 방법으로 자질 벡터의 구성이 분류 성능에 큰 영향을 미친다. 이에 따라 스팸 문자 메시지 필터링을 위한 벡터 표현은 성능에 영향을 끼치는 중요한 요소이다[7]. 현재 텍스트에 대한 벡터의 표현 방식으로 TF-IDF가 가장 널리 사용되고 있지만, TF-IDF의 경우에는 단어 수가 증가함에 따라 차원의 수도 같이 증가하여 대용량 단어들을 처리하는데 어려움 있다. 이러한 차원수 증가 문제를 해결하기 위해 TF-IDF는 고차원의 벡터를 저차원의 벡터로 차원 축소 작업을 추가로 수행해야 한다. 하지만 신경망 언어 모델의 워드 임베딩 통한 단어 벡터는 기존의 TF-IDF보다 차원 축소에 자유롭고, 벡터를 통한 단어 의미 유사성 추론에서도 더

* 학생회원, 국민대학교 컴퓨터공학과

** 정회원, 국민대학교 소프트웨어학부

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.NRF-2017M3C4A7068186).

접수일자 : 2018년 03월 05일

게재확정일 : 2018년 04월 28일

수정일자 : 2018년 04월 11일

교신저자 : 강승식 e-mail : sskang@kookmin.ac.kr

나은 성능을 보여준다[8, 9, 10].

문서나 문장 등의 데이터를 이미 정해진 범주에 따라 분류하는 방법으로는 K-NN 방법, SVM, Naive Bayes 등의 기계 학습 방법들을 이용하고 있다[11]. 하지만 스팸 문자 메시지 필터링은 정해진 범주를 2개로 고정하여 여러 문장 및 구 형태의 텍스트 데이터를 분류하는 이진 분류(binary classification) 문제이므로 분류 기법에 적합한 딥러닝 모델을 이용하여 스팸 문자 메시지를 필터링하고자 하였다.

딥러닝 모델 중 분류 문제에 가장 널리 사용되는 모델은 연속성의 데이터를 분류하는 RNN(recurrent neural network), 특징 벡터를 추출하여 분류하는 CNN(convolution neural network), 고정 길이의 특징 벡터를 분류하는 전방향 신경망(feed-forward neural network) 등이 있다[12, 13, 14].

본 연구에서는 신경망을 이용한 워드 임베딩(word embedding)으로 문자 메시지를 벡터로 표현하고 전방향 신경망을 이용하는 스팸 문자 메시지 필터링 기법을 제안한다.

II. 본 론

1. 워드 임베딩과 문장 벡터 구성

SMS 문자 메시지 데이터를 살펴보면, 하나의 문장 및 구는 단어들의 집합이므로 각 단어들의 빈도수를 이용하여 하나의 벡터로 표현될 수 있다. 이러한 빈도수 기반의 벡터는 단어의 수가 많아짐에 따라 차원의 수도 비례하게 커지는 차원의 저주라는 문제점이 있다. 예를 들어, 아래 5개의 문장을 살펴보면, 띄어쓰기를 기준으로 벡터로 표현하기 위한 총 차원 수는 13이다. 1번 문장을 구성하는 “회의”, “중이니”, “나중에”, “하겠습니다.”라는 단어들을 가지고 문장 벡터로 표현한다면, 문장을 구성하지 않는 단어들은 0으로, 문장을 구성하는 단어들은 각 단어의 빈도수로 채우는 형태의 [0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0]와 같은 문장 벡터를 생성하게 된다. 하지만 이러한 문장 벡터는 단어 수가 증가함에 따라 차원 수도 증가하는 문제점을 내포하고 있다.

1. 회의 중이니 나중에 전화하겠습니다.
2. 즐거운 주말 보내세요
3. 지금 전화를 받을 수 없습니다.
4. 저녁 약속있나요?
5. 저녁먹자^^~~

이러한 문제점을 해결하기 위해, 차원 수 조절에 자유롭고, 단어의 의미를 벡터로 표현하는데도 효율적인 신경망을 이용한 CBOW와 skip gram으로 단어 벡터를 생성하여 전체 단어

수에 비례하게 커지는 차원 수 증가 문제를 해결하였다[13].

그림 1은 단어 벡터를 표현하는 방법인 CBOW와 skip gram이다. 이 두 개의 워드 임베딩 방법은 서로 다른 방법으로 단어를 벡터로 표현한다. CBOW의 경우는 중심 단어의 주변 단어들로부터 중심단어를 예측하고, skip gram은 중심단어에서 주변 단어들을 예측하여 중심 단어를 벡터로 표현한다. 단어 벡터를 생성하는 CBOW와 skip gram으로 문장을 구성하는 단어 벡터를 생성하여 문장 벡터로 표현하였다.

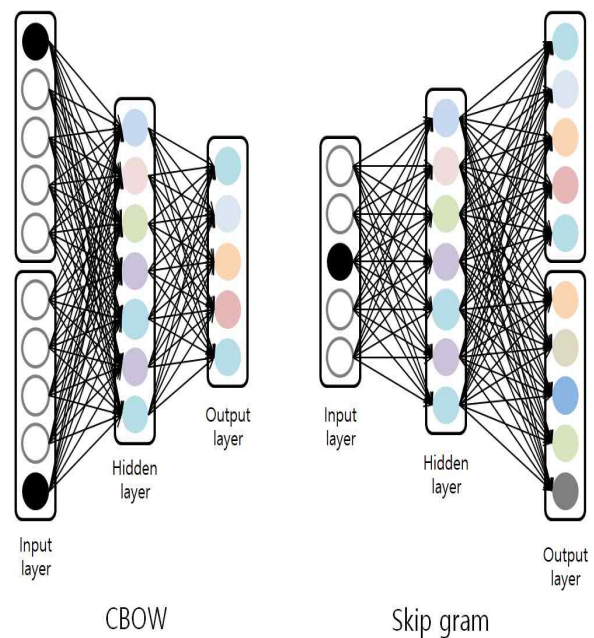


그림 1. CBOW와 skip gram

2. SMS 문자 메시지의 띄어쓰기 오류 문제

사용자들이 문자 메시지를 전송할 때 단어들 사이에 공백 문자를 의도적으로 삽입하거나 제거하여 문자 메시지를 전송한다. 아래 예제는 스팸 문자 메시지에 자동 띄어쓰기를 적용한 예제의 일부이다.

- 자동 띄어쓰기 전:
1억을진짜먹네남일같았던1억강원에뺏긴돈이곳에선가능신년엄청혜택QQ2480.com
 - 자동 띄어쓰기 후:
1억을 진짜먹네 남일 같았던 1억 강원에 뺏긴돈 이곳에선 가능신년 엄청혜택 QQ2480.com
- 위 예제를 살펴보면, “1억을진짜먹네남일같았던1억강원에

뻗긴돈이곳에선가능신년엄청해택QQ2480.com”의 문장은 띄어쓰기를 기준으로 문장 자체가 하나의 단어인 희소한 형태로 이루어진다. 이러한 희소한 형태의 단어는 워드 임베딩에 의한 문장 및 구를 벡터로 표현하는 질을 저하 시키는 요인이 될 수 있다[13].

- 공백이 삽입된 단어의 자동 띄어쓰기 예

띄어쓰기 전	:	띄어쓰기 후
애플	:	애플
N H 금융	:	NH금융
대리운전	:	대리운전

위 예제는 문자 메시지의 한 어절마다 공백문자를 삽입하여 단어를 변형시킨 스팸 문자메시지의 예제 일부이다. 위 예제를 살펴보면 하나의 어절마다 띄어쓰기를 적용하여 “N H 금융” 등과 같이 고의로 공백문자를 삽입하여 “NH금융”의 단어를 변형시킨다. 이러한 변형된 형태는 의미가 유사한 단어는 그 단어의 주변의 단어들도 유사한 분포를 가진다는 분산 가설(distribution hypothesis)에 의한 워드 임베딩의 단어를 표현하는 벡터의 질을 저하시킨다. 즉, “애플”이라는 단어 벡터 생성 시 자동 띄어쓰기를 적용하지 않는 경우 “애플”과 “플”이라는 각각 음절이 서로를 의미를 나타내는 문맥(context)으로 사용되어 저 빈도 음절 벡터를 생성한다. 그리하여 “애플”이라는 단어 벡터 생성 시 필요한 문맥에 영향을 주어 올바른 단어 벡터 생성에 영향을 미친다[15, 16].

본 연구에서는 분산 가설에 의해 자동 띄어쓰기를 적용하여 어절마다 공백이 삽입된 단어는 공백을 제거하고, 공백을 제거한 문장은 공백을 삽입하여 올바른 형태의 단어와 문장으로 바꾸는 과정을 수행하여 단어의 의미를 잘 표현하는 단어 벡터를 생성하였다[17].

3. 문장 벡터를 이용한 스팸 문자 메시지 필터링

분산 가설에 의한 워드 임베딩은 유사한 의미를 가진 단어들은 벡터 공간상에 유사한 위치에 위치시킨다. 이를 위하여 전처리 과정으로서 자동 띄어쓰기를 적용하고 문장을 구성하는 단어의 분포를 올바르게 유지하여 최대한 유사한 문맥들과 위치한 단어들은 벡터 공간상에서 유사한 위치에 위치하도록 하였다. 아래 예제를 분산 가설로 살펴보면 “호텔에서”와 “숙소에서”의 단어 의미는 문맥인 “보험”, “가입”, “후”, “나는”, “늦은”, “밤에”, “저녁밥을”, “먹었다”에 의해 표현되어 벡터 공간 상에 유사한 위치에 나타난다[17].

- 보험 가입 후 나는 늦은 밤에 호텔에서 저녁밥을 먹었다.

- 보험 가입 후 나는 늦은 밤에 숙소에서 저녁밥을 먹었다.

스팸 문자 메시지에서는 “첫!가입!”, “㉠㉠N0”, “ㅇㅈ ㅁㅈ 토가”, “ㄱ ㅏ입”, “ㄷㅏ입즉ㅏ!”와 같은 한글단어들의 초성, 중성, 종성을 유사한 형태의 영어나 다른 나라의 언어, 특수기호, 숫자 등으로 대체하여 단어를 다양한 형태에 단어로 왜곡시킨다. 그리하여 보편적인 스팸 문자 메시지 필터링 시스템에서는 이러한 왜곡된 단어들을 “첫가입”, “가지노”, “가입즉시”와 같은 정규화 과정을 거쳐 스팸 문자 메시지 필터링을 위한 어휘 사전과의 비교를 통해 스팸 문자 메시지를 차단하고 있다[18, 19]. 분산 가설에 의한 워드 임베딩은 왜곡된 단어의 문맥과 왜곡되지 않은 단어의 문맥이 유사하다면 같은 문맥을 가진 단어는 벡터 공간상에서 유사한 위치에 나타나도록 단어 벡터를 위치시킨다. CBOW와 skip gram을 통한 워드 임베딩에서는 “가입즉시”와 “ㄷㅏ입즉ㅏ!”의 각 단어 벡터는 같은 문맥과 함께 사용된다면 벡터 공간상에 유사한 위치에 분포한다. 이러한 분포를 통해 두 단어는 유사한 의미를 지닌 단어라고 유추가 가능하고, 또한 왜곡된 단어의 문맥과 왜곡 되지 않은 단어의 문맥이 다르다면 각 단어는 벡터 공간상에서 유사한 위치가 아닌 먼 거리에 위치하여 각 단어를 구분하는데 용이하다. 이를 통해 자모를 특수한 기호로 대체한 왜곡된 형태의 단어의 정규화를 거친 단어들이 포함된 정상적인 문자 메시지를 스팸 문자 메시지로 잘 못 필터링하는 경우를 예방하고자 하였고, 왜곡된 단어를 정규화하는 불필요한 과정을 생략하였다. 이러한 왜곡된 단어들도 하나의 단어 토큰(token)으로 하는 방법을 통해 왜곡되지 않은 단어와 왜곡된 단어의 구분하기 좋은 단어 벡터들을 생성하였다.

문자 메시지는 이를 구성하는 단어들의 벡터 합으로 문장 벡터로 표현하였다. 이는 유사한 단어들의 집합으로 이루어진 문자 메시지는 단어 집합에 속한 단어들의 벡터 합으로 벡터 공간 상에 유사한 위치로 위치시킬 수 있다. 그리하여 문자 메시지를 구성하는 단어들의 벡터 합으로 문장 및 구를 벡터로 표현하였다.

본 연구에서는 분류기로 자질 벡터의 구성이 분류 성능에 영향을 큰 영향을 미치는 지지 벡터 기계보다 벡터 외에도 은닉층(hidden layer)의 수, 활성화 함수(activation function), 학습 알고리즘 등의 요인들을 통해 분류 성능에 영향을 끼치는 전방향 신경망을 분류기로 사용하여 문자 메시지를 스팸 문자 메시지 또는 정상적인 문자 메시지로 분류하였다.

그림 2의 분류(classification)과정은 전방향 신경망에 의한 이진 분류이며 전방향 신경망의 비선형 함수와 비용(cost) 함수로는 시그모이드(sigmoid) 함수, 교차 엔트로피(cross-entropy) 함수를 이용하였다.

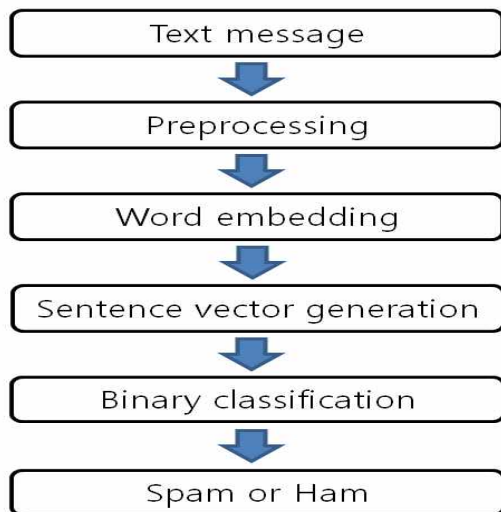


그림 2. 워드 임베딩과 딥러닝을 이용한 문자 메시지 필터링 과정

교차 엔트로피 함수의 입력 값으로는 스팸 문자 메시지인지 정상적인 문자 메시지인지를 나타내는 2차원 벡터 형태의 정답 레이블(label)과 전방향 신경망의 출력 층의 2차원 벡터 값을 소프트 맥스(softmax) 함수로 정규화한 값으로 이 두 개 2차원 벡터 값들의 확률분포 비교하여 교차 엔트로피 함수를 최소화하는 방향으로 학습을 진행하여 스팸 문자 메시지를 필터링하였다.

4. 실험 및 성능 평가

워드 임베딩과 전방향 신경망을 이용한 스팸 문자 메시지 필터링을 위한 실험 데이터는 총 109,993개의 문자 데이터 집합이다. 이 데이터는 중복 단어를 포함하여 910,143개 단어들로 구성되어 있고 중복 단어 제거하면 895,918개의 단어들로 이루어졌다. 실험 데이터를 표 1과 같이 학습 및 테스트용으로 분할하였다.

표 1. 문자 메시지의 데이터 수

		Lines	Words (with duplication)	Words (without duplication)
Test	Ham	5,000	27,785	27,732
	Spam	5,000	44,642	44,610
Train	Ham	49,993	413,441	403,520
	Spam	50,000	424,275	420,056
Total		109,993	910,143	895,918

Ham : 정상적인 문자 메시지
Spam : 스팸 문자 메시지

문자 메시지를 구성하는 문장 및 구에 대한 벡터 표현을 위해 CBOW와 skip gram이라는 두 가지 방식의 워드 임베딩 알고리즘을 이용하여 단어 벡터를 생성하는 실험을 수행하였다. 각각 워드 임베딩 방식을 통하여 하나의 문자 메시지를 구성하는 단어들의 벡터 합으로 문자 메시지를 벡터로 표현하였고, 이러한 고정 길의 문장 벡터로 딥러닝 모델인 전방향 신경망과 SVM Light를 이용하여 스팸 문자 메시지 필터링의 정확도를 비교 측정하였다. 표 2는 전방향 신경망과 SVM Light에 의한 스팸 문자 메시지 필터링의 정확도 측정 결과이다. 전방향 신경망의 학습 방법은 경사도 하강법(gradient descent)을 이용하여 학습하였다.

표 2. SVM Light와 전방향 신경망의 정확도

		Layer	Accuracy(%)
SVM Light	CBOW	·	95.72
	Skip gram	·	93.57
Feed-forward Neural Network	CBOW	1	95.4
		2	95.19
		3	95.87
		4	95.53
	Skip gram	1	93.56
		2	93.84
		3	93.57
		4	94.06

워드 임베딩에서 윈도우 크기는 8로 하였고, 워드 벡터는 300차원으로 구성하였다. 표 2의 스팸 문자 메시지는 스팸 문자 메시지로 정상적인 문자 메시지는 정상적인 문자 메시지로 올바르게 분류하는 정확도에서 딥러닝 모델인 전방향 신경망을 이용한 경우는 95.87%가 가장 높은 정확도로 SVM light의 95.72%보다 0.15% 더 높은 성능을 보여주고 있다.

표 3. CBOW와 Skip gram의 차원 수에 따른 스팸 문자 메시지 필터링의 정확도

		Dimension	Accuracy(%)
SVM Light	CBOW	200	93.84
		250	95.65
		300	95.72
	Skip gram	200	93.78
		250	93.43
		300	93.57

표 3의 SVM light를 사용하여 CBOW와 skip gram의 벡터 차원 수에 따른 정상적인 문자 메시지는 정상적인 문자메시지로 스팸 문자메시지는 스팸 문자메시지로 올바르게 분류하는

정확도에서는 CBOW와 skip gram보다 더 높은 성능을 보여 주고 있다. 이를 통해, 단어를 벡터로 표현하는 방법에서는 skip gram보다 CBOW를 이용하는 것이 정확도가 더 높은 것을 알 수 있었다.

III. 결론

핸드폰의 SMS 문자 메시지로 전송되는 대량의 광고문자들을 필터링하는 방법으로 워드 임베딩과 딥러닝 기법을 이용한 스팸 문자 메시지 필터링 방법을 제안하였다. 스팸 문자들은 의도적으로 띄어쓰기 규칙을 지키지 않거나 한글 자모 대신에 특수기호나 기호를 이용하여 음절을 구성하는 경우가 많은데 자동 띄어쓰기를 이용하여 띄어쓰기 오류를 교정하고, 자모 해체와 특수기호로 대체된 경우는 왜곡된 단어 형태를 포함하여 워드 임베딩을 수행하였다. 문자 메시지를 구성하는 문장 벡터 표현은 CBOW와 skip gram이라는 두 가지 방식으로 워드 임베딩 실험을 수행하였고, 딥러닝 기법의 효용성을 검증하기 위해 SVM 방법의 문자 필터링과 성능을 비교하였다. 그 결과로 SVM과 딥러닝 기법 모두 CBOW를 이용한 워드 임베딩 성능이 skip gram 방식보다 정확도가 높은 것을 알 수 있었고, 또한 딥러닝 기법이 SVM 방식보다 더 성능이 좋은 것을 확인하였다.

REFERENCES

- [1] 박경민, 최훈, 이창건, 황인태, 이칠우, "휴대 단말을 위한 지능형 사용자 인터페이스 플랫폼," *스마트미디어저널*, 제1권, 제4호, 44-51쪽, 2012년 12월
- [2] 손대능, 이정태, 이승욱, 신중휘, 임해창, "문자 메시지의 특성을 고려한 한국어 모바일 스팸 필터링 시스템," *한국산학기술학회논문지*, 제11권, 제7호, 2595-2602쪽, 2010년 7월
- [3] M. Salib, "MeatSlicer: Spam Classification with Naive Bayes and Smart Heuristics," *Proceedings of the Spam Conference*, Dec. 2002.
- [4] K. Schneider, "A Comparison of Event Models for Naive Bayes Anti-Spam E-Mail Filtering," *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics(EACL 2003)*, Budapest, Hungary, vol. 1, pp. 307-314, April. 2003.
- [5] 강승식, "메일 주소 유효성과 제목-내용 가중치 기법에 의한 스팸 메일 필터링," *멀티미디어학회 논문지*, 제9권, 제2호, 255-263쪽, 2006년 2월
- [6] Drucker, H., Wu, D., & Vapnik, V. N., "Support Vector Machines for Spam Categorization," *IEEE Transactions on Neural networks*, vol. 10, Issue 5, pp. 1048-1054, Sep. 1999
- [7] 허기수, 정현태, 박아론, 백성준, "양자 유전 알고리즘을 이용한 특징 선택 및 성능 분석," *스마트미디어저널*, 제1권, 제1호, 40-45쪽, 2012년 3월
- [8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J., "Distributed Representations of Words and Phrases and their Compositionality," *In Advances in neural information processing systems, Lake Tahoe, the United States*, pp. 3111-3119, Dec. 2013.
- [9] Mikolov, Tomáš, et al., "Recurrent neural network based language model," *Eleventh Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, pp. 1045-1048, Sep. 2010.
- [10] Mikolov, T., Yih, W. T., & Zweig, G., "Linguistic Regularities in Continuous Space Word Representations," *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, the United States, pp. 746-751, Jun. 2013.
- [11] Manevitz, L. M., & Yousef, M., "One-class SVMs for Document Classification," *Journal of machine Learning research*, vol. 2, pp. 139-154, Dec. 2001.
- [12] Socher, R., Lin, C. C., Manning, C., & Ng, A. Y., "Parsing Natural Scenes and Natural Language with Recursive Neural Networks," *In Proceedings of the 28th international conference on machine learning (ICML-11)*, Bellevue, Washington, USA, pp. 129-136, Jul. 2011.
- [13] Chen, D., & Manning, C., "A Fast and Accurate Dependency Parser using Neural Networks," *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar, pp. 740-750, Oct. 2014.
- [14] Simard, P. Y., Steinkraus, D., & Platt, J. C., "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," *In Proceedings of the 7th International Conference on Document Analysis and Recognition(ICDAR 2003)*, Edinburgh, Scotland, UK, vol. 2, pp. 958-962, Aug. 2003.
- [15] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [16] Sahlgren, M., "The distributional hypothesis," *Italian Journal of Disability Studies*, vol.20, pp.

33-53, 2008.

- [17] 강승식, "음절 bigram을 이용한 띄어쓰기 오류의 자동 교정," *음성과학*, 제8권, 제2호, 83-90쪽, 2001년 6월
- [18] 강승식, 장두성, "SMS 변형된 문자열의 자동 오류 교정 시스템," *정보과학회논문지*, 제35권, 제6호, 386-391쪽, 2008년 6월
- [19] 강승식, "스팸 문자 필터링을 위한 변형된 한글 SMS 문장의 정규화 기법," *정보처리학회논문지*, 제3권, 제7호, 271-276쪽, 2014년 7월

저 자 소 개



이현영(학생회원)

2016년 국민대학교 컴퓨터공학 학과 학사 졸업(공학사).
 2016년 ~ 2017년 SK hynix memory solutions inc. Intern
 2017년 ~ 현재 국민대학교 컴퓨터공학 학과 석사과정.

<주관심분야 : 자연어처리, 머신 러닝, 딥러닝, 빅데이터 분석>



강승식(정회원)

1986년 서울대학교 전자계산기공학과 학사 졸업.
 1988년 서울대학교 전자계산기공학과 학과 석사 졸업.
 1993년 서울대학교 전자계산기공학과 학과 박사 졸업.

<주관심분야 : 자연어처리, 텍스트 마이닝, 빅데이터 분석, 상황인지 컴퓨팅>