음소 단위 임베딩 모형을 이용한 감성 분석

현경석*, 최우성*, 정순영*, 정재화**
*고려대학교 컴퓨터학과
**한국방송통신대학교 컴퓨터과학과
e-mail: jaehwachung@knou.ac.kr

Sentimental Analysis using the Phoneme-level Embedding Model

Kyeongseok Hyun*, Woosung Choi**, Soon-young Jung*, Jaehwa Chung**†
*Dept. of Computer Science, Korea University
**Dept. of Computer Science, Korea National Open University

요 약

형태소 분석을 통하여 한국어 문장을 형태소 단위의 임베딩 및 학습 관련 연구가 되었으나 최근 비정형적인 텍스트 데이터의 증가에 따라 음소 단위의 임베딩을 통한 신경망 학습에 대한 요구가 높아지고 있다. 본 논문은 비정형적인 텍스트 감성 분석 성능 향상을 위해 음소 단위의 토큰을 생성하고 이를 CNN 모형을 기반으로 다차원 임베딩을 수행하고 감성분석을 위하여 양방향 순환신경망 모델을 사용하여 유튜브의 비정형 텍스트를 학습시켰다. 그 결과 텍스트의 긍정 부정 판별에 있어 90%의 정확도를 보였다.

1. 서론

딥러닝은 2012년 시신경의 동작 원리에 착안하여 이미지를 인식하는 신경망인 CNN의 성과를 시작으로 급속도로 응용되는 분야가 확산되고 있다. 최근 BERT, ELmo 등의 신경망 기반의 언어 모델과 대량의 텍스트를 활용하여 우수한 성능을 나타내고 있는 자연어처리(NLP) 분야에서도 딥러닝 기술이 활발하게 활용되고 있다. 예를 들어, 텍스트를 분석하여 필자의 감성을 분석하는 감성분석(sentiment analysis) 기술은 마케팅, 선거전략 등의 분야에서 활용될 수 있다.

신경망 기반의 텍스트 감성분석을 위해서는 일반적으로 텍스트를 단어 등의 작은 단위로 나누어주는 토크나이징(tokenizing) 기술이 필요하다. 한국어의 경우 형태소 분석을 통해 한국어 문장을 형태소 단위로 토크나이징하는 방법이 사용되거나 데이터 기반으로 학습된 sub-word를 이용하는 방법(wordpiece[3], sentencepiece[4])이 주로 사용되어 왔다.

토크나이징 기술은 신경망 기반 텍스트 감성 분석의 성능에 직접적 영향을 끼친다. 한편, 최근 제안된 [5]에 따르면 문자 단위의 토큰을 생성하는 것이 형태소(morphology)가 풍부한 언어에 더욱 적합한 것으로 밝혀졌다. 본 논문은 이러한 원리에 착안하여, 음소 단위의 토크나이징을 활용하는 텍스트 감성 분석 기법을 제안한다.

† 교신저자: 정재화 한국방송통신대학교 jaehwachung@knou.ac.kr
이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2016R1D1A1B03930907). 또한 이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2019R1F1A1062719)

2. 관련연구

신경망을 통하여 감성분석과 관련된 연구들이 활발히 이루어지고 있으며, 영어의 경우 언어적 특성에 따라 문자 단위분석이 용이하여 문자 단위의 연구가 이루어졌으나, 한글의 경우 형태소 단위의 연구가 대다수로 주를 이루었다.

한글 문서의 형태소 단위의 감성분석 연구 중, 초성어의 정규화를 통해 감성 분석의 정확도를 향상시키는 것을 목적으로 [1]이 제안 되었다. 영상매체의 리뷰나, 트위터에서 자주 사용되는 초성 'ㅋㅋ', 'ㅋㅋㅋㅋ' 와 같은 초성이지만 연속표기의 차이로 문맥상으로 감성의 표현의 강도는 다를 수 있지만 같은 의미를 내포하는 초성을 전처리 작업을 통하여 정규화 시킨 후, 이를 기반으로 감성분석을 수행하였다. 사용 빈도가 가장 높은 초성어를 기준으로 정규화 전처리를 적용한 문서를 학습할 시 감성분석의 정확도의 차이가 1.9%정도 향상되는 것으로 나타났다.

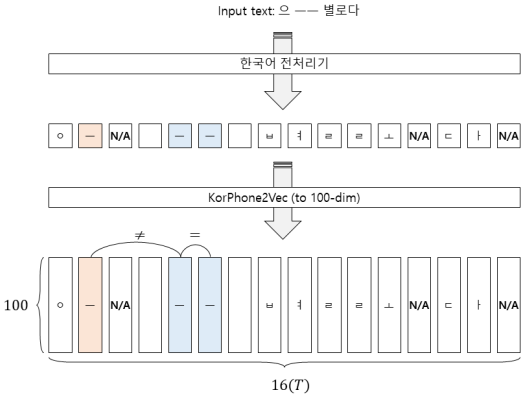
한글 문서의 음소 단위를 기반으로 신경망을 이용한 감성분석으로는 [2]가 제안 되었다. "이 영화 정말 재밌다!!" 와 같은 텍스트가 입력 되었을 때, 음소 단위로 "ㅇㅣ ㅇㅕㅇㅎㅗㅏ ㅈㅓㅇㅁㅏㄹ ㅈㅐㅁㅣㅆㄷㅏ !!" 와 같이 음소 단위로 분리 한 후 one-hot 벡터로 변환 후 학습을 진행하였다. 학습은 LSTM과 GRU로 진행하였으며, 결과적으로 LSTM이 8% 높은 정확도를 가지는 것을 보였다.

위 연구들은 형태소 단위 및 음소 단위의 처리를 위하여 전처리과정에서 신경망을 활용하지 않았다. [1]의 경우 빈번하게 사용되는 초성을 정규화 하는 과정에서 문맥의

모호성을 감소 시켰지만 결과적으로 문맥에서의 감성정보 손실이 발생한다. [2]의 경우 문서를 음소로 분리 한 후 one-hot 벡터로 변환하여 초성, 중성, 종성간의 연관관계를 반영하는데 한계를 지니게 된다. 따라서 음소가 내포하는 유의미한 정보를 최대한 표현하기 위해 음소 또한 다차원 공간의 벡터로 변환이 필요하며, 본 논문은 이를 CNN을 통하여 학습하고 이를 기반으로 GRU 모델을 통하여 감성분석을 진행한다.

3. 제안연구의 모델

본 논문에서는 음소단위로 분리 한 텍스트 데이터를 다차원 벡터로 학습시키는 전처리 과정과, 이를 기반으로 입력된 텍스트의 긍정, 부정을 예측할 수 있도록 학습하는 두 단계의 모형을 제시한다.

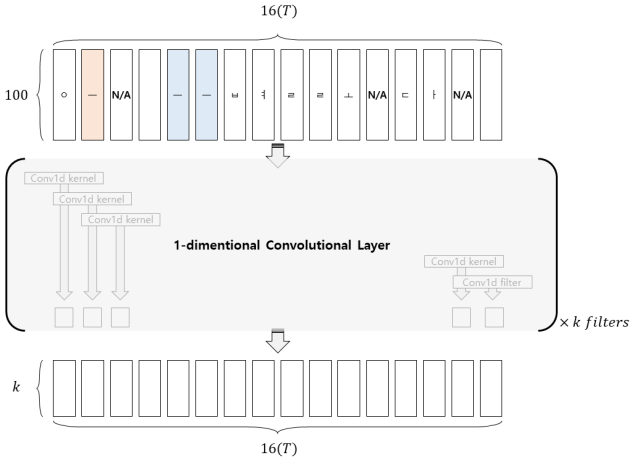


(그림 1) 음소단위 전처리 과정

(그림 1)는 음소단위 전처리의 과정을 표현한 것으로 입력된 텍스트를 음소단위(초성, 중성, 종성)으로 분리 후 음소를 각각 100차원으로 변환한다. “으 —— 별로다” 라는 텍스트가 입력되었을 때 이를 음소 단위로 모두 분리하고 종성이 없을 경우 N/A로 대체하여 표현하게 된다. 위와 같이 차원 변환을 함으로써 “으” 의 중성과 “——” 다른 차원에 벡터로 표현이 가능하며, 이는 다른 시멘틱을 갖는 것으로 해석이 가능하게 된다. 하지만 기존연구에서와 같이 이를 one-hot 벡터로 표현하게 된다면 모두 같은 벡터로 보기 때문에 감성정보의 손실이 발생하게 된다.

그 후 100차원으로 변환된 음소의 벡터들을 CNN을 통하여 (그림 2)와 같이 (그림 1)의 100차원 벡터들을 size = 3 인 k 개의 CNN 필터를 통하여 초성, 중성, 종성 간의 연관 정보 즉, 하나의 글자로 서의 의미를 표현할 수 있을 뿐만 아니라, 글자와 다음 초성간의 연관 정보를 점진적으로 내포하여 차원변환을 할 수 있게 된다.

본 논문에서는 3개의 1차원 convolutional 레이어를 사용하였으며, 각 레이어 에서 다음 레이어로 $k \times T$ 벡터를 넘겨준다. 매번 convolutional 레이어를 거치면서 초성의 앞뒤만 고려했던 연관 정보의 영향력이 확장되어 벡터로써 표현되게 된다. 해당 모델링을 통해 ‘그래’, ‘그러자’, ‘그럽시다’ 와 같이 변형이 많은 한국어의 특성상 단순 초성간의 관계로 파악하기 어려운 단어들을 유사한 벡터로 표현이 가능하다.



(그림 2) CNN 기반 임베딩 수행 과정

음소단위 전처리기를 통한 $k \times T$ 벡터가 산출되면 양방향 순환신경망 (bidirectional GRU) 모델을 통하여 벡터를 순차 방향으로 학습한 벡터와 역순으로 학습한 결과를 도출한 뒤 긍정과 부정이 될 확률을 계산한다. 해당 접근 방법으로 주어, 목적어, 서술어 어순으로 구성되는 한국어의 통사적 특징 상, 정보 전달에 있어서 중요한 주어 와 서술어를 양 방향으로 고려함으로 보다 문맥의 모호성을 감소하고 긍정, 부정 판별에 용이하다고 볼 수 있다.

4. 실험

본 연구에서 음소단위 임베딩을 적용한 데이터는 유튜브 영상의 댓글을 크롤링하여 10,000 개의 댓글을 긍정, 부정으로 분류하였다. <표 1>은 수집한 데이터의 예시로, 데이터의 10%를 무작위로 테스트 데이터로 선정하였다.

<표 1> 데이터 샘플

텍스트	감성
저도 오늘 보쌈 먹었는뎅ㅋㅋㅋㅋㅋㅋㅋ	긍정
안녕하세요! 맛있게 잘 드시네용ㅎㅎ	긍정
다이어트 하는데 ㅜㅜ	부정

해당 텍스트를 음소단위로 분리 후 다차원의 벡터 값으로 변환 후 CNN 기반으로 초성, 중성, 종성을 고려할 수 있도록 학습을 진행하였다. 그 후 양방향 순환신경망을 통하여 감성 분석을 <표 2>와 같은 환경에서 진행하였다. 파라미터 최적화 Adam을 사용하였으며, 목적함수는 BCE(Binary Cross Entropy)를 사용하였다.

<표 2> 사용 장비

구분	품목
CPU	Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz
GPU	NVIDIA GeForce RTX 2080Ti
RAM	DDR4 16G x 8

또한 양방향 순환신경망이 한국의 통사적 구조를 반영하여 감성분석의 정확도를 향상시키는지 확인하기 위하여 동일한 하이퍼 파라미터로 단방향 순환신경망(Unidirectional GRU)로 학습을 진행하였다. <표 3>은 hyper parameter 별로 20 epoch를 수행했을 때의 정확도를 보여준다. E는 음소를 임베딩한 차원의 수, H는 hidden 차원의 수, L 은 GRU 레이어의 수를 의미한다. 결과 적으로 감성 분석의 정확도는 89% ~ 90% 보이며 임베딩 차원이 200, hidden 차원이 512 일 때 가장 정확도가 높았다.

<표 3> hyper parameter 별 정확도

GRU	모델 (E, H, L)	정확도
Bidirectional	(100, 128, 3)	0.8962
	(120, 256, 3)	0.8959
	(150, 256, 3)	0.8940
	(200, 512, 3)	0.9039
Unidirectional	(100, 128, 3)	0.8971
	(120, 256, 3)	0.8960
	(150, 256, 3)	0.8996
	(200, 512, 3)	0.8967

5. 결론

본 논문에서는 초성 및 비언어적 감성이 담긴 데이터를 포함하여 음소 단위 임베딩을 하고 이를 기반으로 감성분석을 진행하였다. 또한 한국어의 통사적 구조를 반영하기 위하여 양방향 순환신경망을 사용 하였으나, 예상과 달리 단방향 순환신경망의 경우 양방향 신경망과 유사한 정확도를 보였다. 이는, 유튜브 댓글의 비정형적인 특징으로 인하여 한국의 통사적 구조가 잘 들어나지 않는 것으로 예상된다.

참고문헌

[1] Junseo Park. (2019), A Study on the Sentimental Analysis through the Pre-processing of Sentimental First-Phoneme Korean words, The International Conference on Big data, IoT, and Cloud Computing

[2] 이재준, 안성만. (2017). 한글 음소 단위를 적용한 딥러닝 모형의 감성분석. 한국지능정보시스템학회 학술대회논문집, 113-114.

[3] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).

[4] Kudo, Taku, and John Richardson. "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing." arXiv preprint arXiv:1808.06226 (2018).

[5] Kim, Yoon, et al. "Character-aware neural language models." Thirtieth AAAI Conference on Artificial Intelligence. 2016.