BPE를 활용한 한국어 감정사전 제작

박호민○, 천민아, 남궁영, 최민석, 윤호, 김재균, 김재훈
한국해양대학교, 컴퓨터정보공학과
homin2006@hanmail.net, minah0218@kmou.ac.kr, young_ng@kmou.ac.kr,
ehdgus5136@naver.com, 4168615@naver.com, jgk20000@naver.com, jhoon@kmou.ac.kr

Developing a Korean sentiment lexicon through BPE

Ho-Min Park○, Min-Ah Cheon, Young Nam-Goong,
Min-Seok Choi, Ho Yoon, Jae-Kyun Kim, Jae-Hoon Kim
Department of Computer Engineering, Korea Maritime and Ocean University

요 약

감정분석은 텍스트에서 나타난 저자 혹은 발화자의 태도, 의견 등과 같은 주관적인 정보를 추출하는 기술이며, 여론 분석, 시장 동향 분석 등 다양한 분야에 두루 사용된다. 감정분석 방법은 사전 기반 방법, 기계학습 기반 방법 등이 있다. 본 논문은 사전 기반 감정분석에 필요한 한국어 감정사전 자동 구축 방법을 제안한다. 본 논문은 영어 감정사전으로부터 한국어 감정사전을 자동으로 구축하는 방법이며, 크게 세 단계로 구성된다. 첫 번째는 한영 병렬 말뭉치를 이용한 한영 이중언어 사전을 구축하는 단계이고, 두 번째는 한영 이중언어 사전을 통한 한영 이중언어 그래프를 생성하는 단계이며, 세 번째는 영어 단어의 감정값을 한국어 BPE의 감정값으로 전파하는 단계이다. 본 논문에서는 제안된 방법의 유효성을 보이기 위해 사전 기반 한국어 감정분석 시스템을 구축하여 평가하였으며, 그 결과 제안된 방법이 합리적인 방법임을 확인할 수 있었으며 향후 연구를 통해 개선한다면 질 좋은 한국어 감정사전을 효과적인 방법으로 구축할 수 있을 것이다.

주제어: 감정분석, 감정사전, 이중언어 사전, 레이블 전파, BPE

1. 서론

감정분석(sentiment analysis)이란 텍스트 상의 의견, 태도 등과 같은 주관적이라고 판단되는 정보를 추출하는 분야로써, 여론 또는 시장 동향을 분석하는데 주로 활용된다. 사용자들의 요구나 특정 제품에 대한 이미지를 정확히 파악하기 위함이다. 사용자의 감정을 제대로 파악할 수만 있다면 그에대한 알맞은 대응이 가능하며 이는 곧 매출 증대로 이어질 수 있기 때문이다.

그러기 위하여 영어의 감정분석에는 크게 두 가지로, 사전 기반 방법과 기계학습 기반 방법이 있다[1]. 사전 기반 방법은 감정사전(sentiment lexicon)을 이용하여 극성(polarity) 혹은 점수를 감정어에 부여하는 방식이다. 단순하게는 긍정, 부정의 극성 또는 점수를 각 감정어에 부여하거나[1-4] 각성(arousal)과 지배(dominance)의 심리학적인 부분을 추가하여 다각적인 관점에서 감정을 분석하기도 한다[5-6].

이러한 감정사전을 제작하는 방법은 집단지성을 통한 사전 제작 방법과 자동 생성 방법이 있다[1]. 자동 생성하는 방식으로는 초기 씨앗(seed) 사전을 사용해 동의어, 반의어와 같은 관계기반으로 새로운 감정어를 추출하여 사전을 확장해 나가거나, 이중언어 사전이나 말뭉치 등을 활용해 기계학습으로 생성하는 것이 있다[7]. 집단지성을 통해 제작한 사전은 수록된 감정어의 극성과 강도가 어느정도 검증되었다는 장점이 있으나 제작하는데 노력과 시간이 많이 든다는 단점이 있다. 반대로 자동 생성 방법은 비교적 제작하기가 쉽고 빠르지만 수록된 감정어들의 검증이 부족하다.

두 방법의 장점만 계승하기 위해, 본 논문은 검증된 감정사전으로부터 자동으로 생성하는 방법으로 제안한다. 대표적인 검증된 영어 감정사전으로 VADER[1]가 있다. VADER는 집단지성으로 제작한 감정사전과 더불어 통사적인 언어적 규칙을 추가적으로 활용하여 감정을 분석한다. 이를 통해 다른 감정사전들보다 높은 정확도를 보이며 심지어 SNS 분야에서는 사람보다 높은 정확도로 감정을 인식하고 분류한다.

본 논문에서는 한영 병렬 말뭉치에 BPE를 적용시켜 이중언어 그래프를 제작하고, VADER 감정어 점수를 레이블 전파 알고리즘으로 전파하여 한국어 감정사전을 제작하는 방법을 제안한다. 제안하는 방법은 두 자료 사이의 연관정도를 계산하는 PMI(Point-wise Mutual Information) 알고리즘[8]을 사용하여 한영 병렬 말뭉치에서 한영 이중언어 사전(bilingual lexicon)을 생성한다. 그렇게 제작된 이중언어 사전 자료와 한국어 BPE 사이의 코사인 유사도로 가중치를 측정하여 그래프를 생성한다. 생성된 그래프 상에서 VADER 감정어 점수들을 점수가 매겨지지 않은 한국어 BPE에 레이블 전파(label propagation) 알고리즘[9]으로 전파하여 한국어 감정사전을 제작한다.

본 논문의 구성은 다음과 같다. 2장에서 제안하는 병렬 말뭉치 기반의 이중언어 그래프 상에서 레이블 전파를 통한 감정사전 제작 방법에 대하여 소개하고, 3장에서는 제작한 감정사전을 감정분석 말뭉치에 적용하여 평가한다. 4장에서는 결론 및 향후 연구에 대해 기술한다.

2. 레이블 전파를 통한 감정사전 제작

본 논문에서 제안하는 레이블 전파를 통한 감정사전의 제작 방법은 크게 세 단계로 구성된다. 첫 번째는 한영 병렬 말뭉치를 이용한 한영 이중언어 사전을 구축하는 것, 두 번째는 한영 이중언어 사전을 통한 이중언어 그래프를 생성하는 것, 세 번째는 VADER 감정어의 감정값을 한국어 BPE의 감정값으로 전파하는 단계이다. 이 장에서는 이들 각 단계를 차례대로 설명한다.

2.1 한영 이중언어 사전 구축

한영 이중언어 사전은 한영 병렬 말뭉치에서 정보를 추출해내어 제작한다[10]. VADER 감정어 쌍의 행렬과 VADER 감정어 - 한국어 BPE 행렬을 제작하여 각 감정어와 한국어 BPE의 벡터화를 수행한다. VADER 감정어 쌍 행렬에서는 각 항목(index)마다 매칭되는 두 감정어의 상호정보(mutual information)값을 계산하여 저장하고, VADER 감정어 - 한국어 BPE 행렬의 항목에는 병렬 말뭉치 상의 서로의 동시 발생 빈도값을 저장한다. 서로 정렬(alignment)되어있는 문장 상에서 VADER의 감정어들은 한국어 문장 내 BPE들의 동시 발생 수치로 계산한다. 그렇게 제작된 VADER 감정어 쌍 행렬과 VADER 감정어 - 한국어 BPE 행렬로 감정어와 한국어 BPE들의 벡터화를 진행한다. 그리고 감정어들 사이의 상호정보(mutual information)와 한영 단어 간 동시 발생 빈도의 코사인 유사도를 계산하여 높은 순서대로 후보군을 생성한다. 말뭉치는 자체 제작한 한영 간 병렬로 번역된 뉴스 말뭉치를 사용한다. 해당 병렬 말뭉치의 상세 자료는 표 1의 내용과 같다.

표 1. 한영 병렬 말뭉치의 상세 정보

항목	개수
문장 수	424,985
영어 어절 수	15,002,423
한국어 어절 수	13,778,428
영어 단어 수	187,554
한국어 형태소 수	179,695

표 2. 제작된 한영 이중언어 사전 예시

영어 단어	최상 유사도 후보
attraction	_유혹
greedy	_탐욕
hatred	_증오
love	_사랑

2.2 한영 이중언어 그래프 생성

레이블 전파 알고리즘을 적용하기 위해선 그래프가 필요하다. 한영 감정어 - BPE 간 그래프를 생성할 때, 전체 감정어와 한국어 BPE를 정점(vertex)으로 한다. BPE 생성엔 sentencepiece 모듈[11]을 사용하여 5,000개로 크기를 고정하여 생성했다. 한국어 BPE들 간의 연결은 격자(mesh) 네트워크 형태로 연결한다. 감정어와 한국어 BPE 간의 연결은 앞서 제작한 한영 이중언어 사전을 통하여 연결하였다. BPE들의 동시 발생 빈도로 코사인 유사도를 계산한다. 해당 유사도를 정점 간 가장자리(edge)의 가중치(weight)로 활용하도록 구성한다. 이중언어 사전을 통해 연결한 가장자리의 가중치는 1로 적용한다. 원시언어와 목표언어가 1대1로 대응되는 사전이므로 가중치를 1로 고정한다.

2.3 감정값 전파

VADER는 영단어 7048개와 이모티콘 469 개, 총 7517개의 감정어로 이루어져있다. 감정어의 감정값은 각 단어마다 10명의 전문가에 의해서 부여된 감정값의 평균이다.

```
love   3.2     0.4     [3, 3, 3, 3, 3, 3, 3, 4, 4, 3]
loved  2.9     0.7     [3, 3, 4, 2, 2, 4, 3, 2, 3, 3]
lovelies       2.2     0.74833 [3, 3, 3, 1, 2, 2, 3, 2, 1, 2]
lovely 2.8     0.6     [2, 3, 3, 3, 2, 3, 4, 3, 2, 3]
lover  2.8     0.87178 [3, 1, 2, 3, 4, 3, 2, 3, 4, 3]
loverly        2.8     0.74833 [3, 2, 4, 3, 3, 2, 3, 2, 2, 4]
lovers 2.4     1.11355 [2, 3, 2, 4, 4, 1, 1, 3, 3, 1]
loves  2.7     0.9     [3, 3, 3, 2, 2, 4, 4, 2, 1, 3]
loving 2.9     0.53852 [3, 2, 3, 3, 3, 2, 4, 3, 3, 3]
```

그림 1. VADER 감정사전 예시

이러한 감정값을 한국어 BPE로 옮기기 위해 VADER 감정어들의 점수를 하나의 레이블(label)로 활용하여 레이블이 없는(unlabeled) 한국어 BPE에게 전파해 주는 레이블 전파 알고리즘(label propagation algorithm)을 제작한 그래프 상에 적용시킨다.

이중언어 사전의 감정어들과 그 점수를 씨앗 사전(seed dictionary)으로 레이블 전파를 실시한다. 레이블 전파를 통해 점수가 부여되지 않은 한국어 BPE에 점수를 전달하여 한국어 감정사전을 제작한다. 레이블 전파 알고리즘의 수행 과정은 다음과 같다. 첫 번째, 한영 이중언어 그래프를 생성하고 초기화한다. 이 때 한국어 BPE들의 레이블은 0으로 초기화된다. 두 번째, 각 정점간 가장자리 연결 및 가중치를 초기화한다. 한영 이중언어 사전을 통한 감정어와 한국어 BPE의 연결에는 가중치 1을 부여하고 한국어 BPE 간의 연결에는 동시 발생 빈도의 코사인 유사도 값을 사용한다. 세 번째, 그래프 내 각 정점 별 점수를 반복하여 계산하며 갱신한다. 각 정점은 주변에 연결된 정점과 가장자리에 따라 계산을 반복할 때 마다 점수가 갱신된다. 네 번째, 세 번째의 작업이 목표한 수렴치에 도달할 때까지 반복한다. 이러한 과정을 통해 한국어 감정사전이 최종으로 생성된다.

3. 실험 및 평가

본 논문에서는 제작된 감정사전을 평가하기 위해 기 구축된 한국어 감정분석 시스템[12]을 사용하여 외부 평

가를 수행한다. 이 장에서는 한국어 감정분석 시스템과 평가 방법을 구체적으로 기술한다.

3.1 한국어 감정분석 시스템

분석할 한국어 문서의 텍스트가 입력되면 sentencepiece 모듈을 이용해 BPE를 수행한다. 그렇게 분석된 BPE 목록을 제작된 감정사전과 비교하여 감정어를 찾아내어 값을 부여한다. 그렇게 모든 감정값이 더해지고, 출력 시에는 전체 문서의 감정값을 정규화하여 -1.0 ~ 1.0 사이의 실수로 출력한다.

3.2 한국어 감정 말뭉치

한국어의 감정 말뭉치는 대표적으로 세종 말뭉치에 포함되어 있는 세종 감성분석 말뭉치[13]와 한국어 뉴스 감정 말뭉치(KMU 감정 말뭉치)[14]와 네이버 감정 영화 말뭉치[15] 등이 있다. 한국어 뉴스 감정 말뭉치는 뉴스 기사의 댓글에 감정 극성을 제공하고 있으며 네이버 감정 영화 말뭉치는 유명 포털사이트에서 제공하는 영화평을 모아 별점에 따라 분류해놓은 말뭉치이다.

3.3 감정사전의 감정값 부여

감정사전의 감정어 개수는 5,000개로 긍정적인 감정어는 2,941개, 부정적인 감정어는 2,059개가 생성되었다. 점수 구간별 분포는 표 3의 내용과 같다.

표 3. 제작된 감정사전의 점수 구간별 단어 분포

점수 범위	긍정적인 단어	부정적인 단어
0.0 ~ 1.0	0	0
1.0 ~ 2.0	2,106	1,434
2.0 ~ 3.0	788	556
3.0 ~ 4.0	47	69

긍정과 부정 모두 1.0 ~ 2.0 구간의 감정어가 사전에서 높은 비율을 차지한다. 또한 집단지성 기반의 감정사전은 평균 3할정도 부정적인 감정어 개수가 긍정적인 감정어 개수보다 많은 반면에[1] 본 논문에서 제작한 감정사전은 반대로 긍정적인 감정어가 3할정도 부정적인 감정어보다 많았다. 이것은 한영 이중언어 사전을 통한 연결에서 긍정적인 감정어가 더 많았음을 알 수 있다.

3.4 한국어 감성분석 시스템을 통한 성능평가

감정분석 말뭉치에서의 실험 결과를 통해 긍정과 부정 양쪽에 균형잡힌 성능을 보임을 알 수 있다. 말뭉치는 네이버 감정 영화 말뭉치와 KMU 감정 말뭉치를 사용했다. 네이버 감정 영화 말뭉치에는 긍정, 부정의 두 가지 종류로 분류되어 있으며, 한국어 감정 말뭉치는 긍정, 중립, 부정의 세 가지 종류로 분류되어 있다.

평가를 위해 각 문서의 감정분석 결과 수치는 정규화(normalization)한다. 평가 결과는 표 4, 5와 같다.

표 4. 네이버 감정 영화 말뭉치를 이용한 감정시스템 평가 결과

		정답	
		긍정 (Positive)	부정 (Negative)
분석	긍정 (True)	30,687	48,489
	부정 (False)	46,789	25,287

표 5. KMU 감정 말뭉치를 이용한 감정시스템 평가 결과

		정답	
		긍정 (Positive)	부정 (Negative)
분석	긍정 (True)	2,631	2,450
	부정 (False)	3,029	5,156

네이버 감정 영화 말뭉치를 기반으로 평가한 결과는 긍정과 부정의 인식에 있어서 균형잡힌 결과를 보여준다. 그에 비해 KMU 감정 말뭉치를 기반으로 평가한 결과는 부정적인 감정분석에 조금 더 높은 정확도를 보였다. 긍정적인 값이 매겨진 감정어가 약 30%정도 많은 감정사전이지만 평가 결과는 부정, 긍정 사이에 큰 차이가 없다. 이는 감정사전의 단어만 풍부해진다면 긍정적, 중립적인 문서에 있어서도 높은 정답률을 보일 수 있음을 시사한다고 할 수 있다. 게다가 단순한 긍정, 부정의 이진 분류나 긍정, 중립, 부정의 삼진 분류가 아닌 감정 점수에 따른 조금 더 복잡한 분석인 만큼 가정에 대한 확신을 부여한다.

레이블 전파 알고리즘을 사용한 허찬 외(2017)[7]에서 제작된 감정사전은 분야(domain)에 제한적이고, 검증되지 않은 발견적인(heuristic) 씨앗 사전을 사용하여 영화평, 상품평의 감정분석을 진행하였다. 본 논문에서 제안하는 방법의 감정사전과 가장 큰 차별점은 적용 분야에 제한이 없다는 점과 영어의 검증된 VADER 감정사전으로부터 한국어 BPE에 감정값을 옮겨왔다는 것이다. 그러므로 말뭉치의 추가와 감정값 전달을 해결하면 의미있는 정답률을 기록할 것으로 예측된다.

4. 결론 및 향후 연구

병렬 말뭉치 기반으로 제작한 이중언어 그래프 상의 감정 점수 전파로 한국어 감정사전을 제작하였다. 영어에 있어서 검증된 감정사전인 VADER를 사용하여, 엄청난 노력과 시간이 걸리는 집단지성 방식과 달리 자동생성 방식으로 일정 수준 이상의 감정사전을 얻을 수 있었다. 그러나 감정 말뭉치로 제작된 감정사전을 평가한 결과, 전체적인 정확도의 보완이 필요함을 알 수 있었다.

이를 보완하기 위해 향후 연구로써 '한영 이중언어 사전의 크기 증가 방법 모색', '다양한 극성의 감정어가 다량 포함된 새로운 병렬 말뭉치 또는 비교가능한 말뭉치로 사전 제작'을 수행 할 예정이다. 그렇게 '긍정 감정어의 추가'와 본 논문의 평가에서 중립 점수는 약한 부정과 약간 긍정 사이를 뜻하므로 '긍정, 부정 점수의 다양한 분포'를 구성할 것이다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017M3C4A7068187, 한국어 정보처리 원천 기술 연구 개발)

참고문헌

[1] C. J. Hutto and E. Gilbert, "VADER : A parsimonious rule-based model for sentiment analysis of social media text", Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[2] F. A. Nielsen, "A new ANEW : Evolution of a word list for sentiment analysis in microblogs", Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts', pp 93-38, 2011.

[3] L. K. Hansen, A. Arvidsson, F. A. Nielsen, E. Colleoni and M. Etter, "Good friends, bad news affect and virality in twitter", Proceedings of the 2011 International Workshop on Social Computing, Network, and Services, 2011.

[4] M. Thelwall, K. Buckley, G. Paltoglou, and D. Cai, "Sentiment strength detection in short informal text", Journal of the American Society for Information Science and Technology, vol. 61, no.12, pp 2544-2558, 2010.

[5] M. M. Bradley and P. J. Lang, Affective Norms for English Words(ANEW) : Instruction Manual and Affective Ratings, Technical Report C-1, 1999.

[6] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas", Behavior Research Methods, vol. 45, no. 4, pp 1191-1207, 2013.

[7] 허찬, 온승엽, "Word2vec와 Label Propagation을 이용한 감성사전 구축 방법", 한국차세대컴퓨팅학회 논문지, vol. 13, no. 2, pp. 93-101, 2017년.

[8] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography", Computational Linguistics, vol. 16, no. 1, pp. 22-29, 1990.

[9] Z. Xiaojin, and G. Zoubin, Learning from Labeled and Unlabeled Data with Label Propagation, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[10] Jae-Hoon Kim, Hong-Seok Kwon, and Hyeong-Won Seo, "Evaluating a pivot-based approach for bilingual lexicon extraction", Computational Intelligence and Neuroscience, vol. 2015, pp. 1-13. 2015.

[11] [online] https://github.com/google/sentencepiece, 2018.

[12] 박호민, "영어 감정사전의 감정 점수 전파를 통한 한국어 감정사전 제작", 한국해양대학교 컴퓨터공학과 석사학위 논문, 2019.

[13] 김흥규, 강범모, 홍정하, "21세기 세종계획 현대국어 기초말뭉치 : 성과와 전망", 한국정보과학회 언어공학연구회 학술발표 논문집, pp. 311-316, 2007.

[14] 이공주, 김재훈, 서형원, 류길수, "뉴스 댓글의 감정 분류를 위한 자질 가중치 설정", 한국마린엔지니어링학회지, vol. 34, no. 6, pp. 871-879, 2010.

[15] [online] https://github.com/e9t/nsmc, 2015.