KIPS Trans. Softw. and Data Eng.
Vol.9, No.2 pp.53~60 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2020.9.2.53

Generating a Korean Sentiment Lexicon Through Sentiment Score Propagation

Ho-Min Park† · Chang-Hyun Kim†† · Jae-Hoon Kim†††

ABSTRACT

Sentiment analysis is the automated process of understanding attitudes and opinions about a given topic from written or spoken text. One of the sentiment analysis approaches is a dictionary-based approach, in which a sentiment dictionary plays an much important role. In this paper, we propose a method to automatically generate Korean sentiment lexicon from the well-known English sentiment lexicon called VADER (Valence Aware Dictionary and sEntiment Reasoner). The proposed method consists of three steps. The first step is to build a Korean-English bilingual lexicon using a Korean-English parallel corpus. The bilingual lexicon is a set of pairs between VADER sentiment words and Korean morphemes as candidates of Korean sentiment words. The second step is to construct a bilingual words graph using the bilingual lexicon. The third step is to run the label propagation algorithm throughout the bilingual graph. Finally a new Korean sentiment lexicon is generated by repeatedly applying the propagation algorithm until the values of all vertices converge. Empirically, the dictionary-based sentiment classifier using the Korean sentiment lexicon outperforms machine learning-based approaches on the KMU sentiment corpus and the Naver sentiment corpus. In the future, we will apply the proposed approach to generate multilingual sentiment lexica.

Keywords : Sentiment Lexicon, Sentiment Analysis, Word Embedding, Label Propagation, Word Graph

감정점수의 전파를 통한 한국어 감정사전 생성

박 호 민† · 김 창 현†† · 김 재 훈†††

요 약

감정분석은 문서 또는 대화상에서 주어진 주제에 대한 태도와 의견을 이해하는 과정이다. 감정분석에는 다양한 접근법이 있다. 그 중 하나는 감정사전을 이용하는 사전 기반 접근법이다. 본 논문에서는 널리 알려진 영어 감정사전인 VADER를 활용하여 한국어 감정사전을 자동으로 생성하는 방법을 제안한다. 제안된 방법은 세 단계로 구성된다. 첫 번째 단계는 한영 병렬 말뭉치를 사용하여 한영 이중언어 사전을 제작한다. 제작된 이중언어 사전은 VADER 감정어와 한국어 형태소 쌍들의 집합이다. 두 번째 단계는 그 이중언어 사전을 사용하여 한영 단어 그래프를 생성한다. 세 번째 단계는 생성된 단어 그래프 상에서 레이블 전파 알고리즘을 실행하여 새로운 감정사전을 구축한다. 이와 같은 과정으로 생성된 한국어 감정사전을 유용성을 보이려고 몇 가지 실험을 수행하였다. 본 논문에서 생성된 감정사전을 이용한 감정 분류기가 기존의 기계학습 기반 감정분류기보다 좋은 성능을 보였다. 앞으로 본 논문에서 제안된 방법을 적용하여 여러 언어의 감정사전을 생성하려고 한다.

키워드 : 감정사전, 감정분석, 단어표상, 레이블 전파, 단어 그래프

1. 서 론

정보통신기술(information communication technology)의 발전과 널리 보급된 통신기기로 사람들은 시공간을 뛰어넘어 수많은 데이터와 정보를 생산하고 공유한다. Go-Globe의 블로그[1]에 따르면 페이스북(Facebook)에서는 1분당 243,000개의 사진이 올라오고, 트위터(Tweeter)에서는 1분당 350,000개의 트윗(tweet)이 게시된다. 또한 유튜브(Youtube)에는 1분당 총합 400시간 이상의 분량에 해당하는 동영상들이 올라오며, 동시에 70만 시간의 동영상들이 보여진다. 이러한 데이터 사이에서 의미있는 정보를 빠르고 정확하게 찾고 분석하는 일은 대단히 중요하다. 본 논문은 이러한 데이터에서 생산자(producer or author)의 감정을 분석하는 일을 다룬다.

※ 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원(NRF-2016S1A5A2A03927611)과 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)과 산업통상자원부 산업전문인력역량강화사업의 재원으로 한국산업기술진흥원(KIAT)의 지원(N0001884, 2019년 임베디드SW 전문인력양성사업)을 받아 수행된 연구임.
† 비 회 원 : 한국해양대학교 컴퓨터공학과 박사과정
†† 비 회 원 : 한국전자통신연구원 인공지능연구소 책임연구원
††† 종신회원 : 한국해양대학교 컴퓨터공학과 교수
Manuscript Received : July 8, 2019
First Revision : September 16, 2019
Accepted : October 15, 2019
* Corresponding Author : Jae-Hoon Kim(jhoon@kmou.ac.kr)

1) https://www.go-globe.com/blog, 2017년 8월 기준.

이를 감정분석(sentiment analysis)이라고 하며, 감정분석은 문서 또는 대화상에서 주어진 주제에 대한 태도와 의견을 이해하는 과정이다[2]. 일반적으로는 긍정 혹은 부정, 좋음 혹은 싫음, 찬성 혹은 반대 등과 같은 이진 형식으로 문서를 분류하지만, 두려움, 화남, 싫어함 등 다양하게 분류할 수도 있다[1-3]. 감정분석 방법에는 기계학습 기반 접근법[4, 5], 사전 기반 접근법[6, 7], 심층학습 기반 접근법[9] 등 다양한 방법이 있다[2, 3]. 본 논문은 사전 기반 접근법에 필수적인 감정사전에 대해 다룬다.

감정사전을 생성하는 방법은 크게 집단지성을 통한 사전 제작 방법과 자동 생성 방법이 있다[6, 7]. 집단지성을 통한 사전 제작 방법은 다양한 계층의 여러 전문가들이 각 감정어에 극성이나 점수를 부여하게 되므로 그 신뢰성이 높다는 장점이 있으나, 사전을 제작하는데 매우 큰 노력과 시간이 소비된다는 단점이 존재한다. 자동 생성 방법은 초기 감정사전(seed dictionary)을 이용하여 동의어(synonym), 반의어(antonym) 등을 추출하고 이를 확장하거나, 이중언어 사전, 말뭉치 등을 이용하여 감정사전을 생성한다[10]. 이러한 방법은 비교적 쉽고 빠르게 감정사전을 제작할 수 있다는 장점이 있지만 그 신뢰성이 다소 부족하다는 단점이 존재한다.

두 방법의 장점을 계승하고 단점은 최대한 줄이기 위하여 본 논문에서는 신뢰성이 어느 정도 검증된 영어 감정사전으로부터 한국어 감정사전을 자동으로 생성하는 방법을 제안한다. 대표적인 영어 감정사전으로는 VADER[6]가 있으며, 이 사전은 집단지성으로 구축되었으며 다른 기계학습 모델들과의 다섯 가지 분야에서의 감정분석을 통해 그 신뢰성이 검증되었다. 그러므로 본 논문에서는 VADER를 활용하여 한국어 감정사전을 자동으로 생성하는 방법을 제안한다. 제안된 방법은 세 단계로 구성된다. 첫 번째 단계는 한영 병렬 말뭉치를 사용하여 한영 이중언어 사전을 제작한다. 제작된 이중언어 사전은 VADER 감정어와 한국어 형태소 쌍들의 집합이다. 두 번째 단계는 그 이중언어 사전을 사용하여 한영 단어 그래프를 생성한다. 세 번째 단계는 생성된 단어 그래프 상에서 레이블 전파 알고리즘[11]을 실행하여 새로운 감정사전을 생성한다. 본 논문에서는 이와 같은 과정으로 생성된 한국어 감정사전의 유용성을 보이기 위해 몇 가지 실험을 수행하였으며 기존의 기계학습 기반 방법보다 우수한 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 감정사전 생성과 관련된 기존 연구들을 조사하고, 3장에서는 본 논문에서 제안된 레이블 전파 알고리즘을 이용한 감정사전 생성 방법을 기술한다. 4장에서는 생성된 감정사전을 기반으로 감정분석 시스템을 구현하고 다른 감정분석 시스템과 비교·분석한다. 마지막으로 5장에서 결론을 맺고 향후 연구 방향을 제시한다.

2) https://monkeylearn.com/sentiment-analysis/

2. 관련 연구: 감정사전 생성

감정사전을 생성하는 방법은 세 가지로 나뉘며 사전 기반(lexicon-based) 방법과 말뭉치 기반(corpus-based) 방법 그리고 집단지성 기반(collective intelligence-based) 방법으로 구분할 수 있다. 각각의 방법들을 소개하고 그에 따른 주요 알고리즘에 관해 기술한다.

2.1 사전 기반 감정사전 생성

사전 기반 감정사전 생성 방법은 일반적인 사전으로부터 감정어를 추출하여 감정사전을 제작하는 것이다. 일반적으로 사전에는 각 단어의 의미, 그리고 동의어와 반의어 같은 단어 사이의 관계 정보가 수록되어 있다. 따라서 내용이 검증된 사전을 보유하고 있다면 비교적 쉽게 다양한 감정어들을 얻어낼 수 있다. Fig. 1은 사전을 기반으로 감정사전을 제작하는 과정을 보인다.

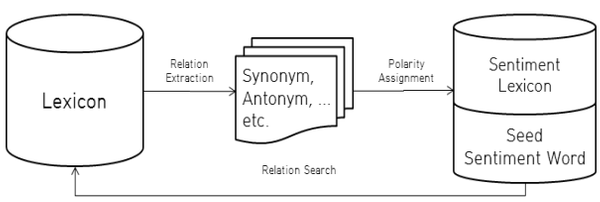

Fig. 1. The Processing of Generating a Sentiment Lexicon using a Lexicon-based Method

Fig. 1에서 초기 감정어는 전문가에 의해서 미리 정해진 최소한의 감정어를 의미한다. 이와 같은 초기 감정어와 WordNet[12]을 이용하여 동의어와 반의어를 추출하여 감정어를 확장한다. 이렇게 확장된 감정어에서 동의어는 같은 극성을 부여하고 반의어는 반대 극성을 부여하여 감정사전에 추가한다. 이것을 반복하여 감정사전의 크기를 확장하는 연구가 진행되었다[13, 14]. 추가적으로 접두사, 접미사와 같은 언어적 특징을 활용하여 다양한 관계어들을 찾아내기도 했다[15].

사전 기반 감정사전 생성 방법은 비교적 쉬운 방법으로 감정어를 확장할 수 있으나 언어에 따라 온톨로지를 쉽게 찾을 수 없어서 감정어를 확장하는 데 다소 어려움이 있을 수 있다. 이와 같은 문제점을 보완하기 위해 말뭉치로부터 감정사전을 생성하는 방법이 등장하였다.

2.2 말뭉치 기반 감정사전 생성

말뭉치 기반 감정사전 생성 방법은 전문가에 의해 초기 감정어를 선정하는 건 사전 기반 감정사전 생성 방법과 동일하나 감정어들이 말뭉치로부터 확장된다는 차이가 있다. 즉 초기 감정어의 연어정보(collocation information)를 이용해서 감정어를 추출하며, 그렇게 추출된 감정어는 초기 감정어

와 동일한 감정의 극성을 부여하여 감정사전을 생성한다. 그러나 연어라고 해서 항상 같은 극성을 가지는 것은 아니므로 부여된 극성에 오류가 존재할 가능성이 있다. 이와 같은 오류를 최소화하기 위해 특정 분야의 말뭉치만을 활용하여 그 분야에 특화된 감정사전을 제작한 연구가 있었다[16].

2.3 집단지성 기반 감정사전 생성

집단지성 기반 감정사전 생성 방법은 전문가에 의해서 초기 감정어가 정해지는 것은 여느 방법과 다르지 않으나 감정어의 확장보다는 극성의 정확성을 높이는 것이 그 목적이다[17]. 감정어의 극성은 전문가의 투표를 통하여 부여되며 극성과 동시에 극성의 강도도 함께 부여되는 것이 일반적이다. 채택된 감정어는 일반적으로 말뭉치 기반 방식과는 달리 특정 분야에 국한되지 않고 범용적인 감정 정보를 지닌다. 특정한 알고리즘이나 메커니즘을 통해 자동으로 감정사전을 제작했던 앞선 두 방식과 달리 직접적인 사람의 수고와 노력이 들어가는 방식으로, 제작 시간과 비용이 상대적으로 높다. 하지만 평가하는 인적 집단이 커질수록, 평가된 감정 점수의 신뢰도가 높아지므로 앞선 두 방식보다 검증되고 정확한 감정사전을 제작할 수 있는 방법이라고 할 수 있다.

VADER[6]는 이러한 집단지성 방법으로 구축된 영어 감정사전이며, 당시 기존의 집단지성 방법으로 제작된 감정사전[18-19]에 포함된 감정어들을 감정어 후보로 사용하여 10명의 평가자, 2명의 전문가가 감정 점수를 부여하여 감정사전을 구축했다. VADER 감정사전을 통해 4,200개의 트윗을 사용한 소셜 미디어 분야, 10,605개의 영화평을 사용한 영화평론 분야, 3,708개의 상품평을 사용한 상품평론 분야, 5,190개의 뉴스 기사를 사용한 사설 분야의 네 가지 분야에서 실험을 진행하였고 소셜 미디어 분야에서는 0.96의 F1 점수를 받아 사람보다도 높은 정확도를 보였다.

본 논문에서는 이렇듯 검증된 감정사전인 VADER를 활용하여, 영어 감정어의 감정 점수를 한국어 감정어 후보에 전파하여 한국어 감정사전을 자동으로 생성한다.

3. 감정 점수 전파를 통한 감정사전 생성

본 논문에서는 한영 단어 그래프를 기반으로 영어 감정 점수를 한국어 감정어 후보로 전파하여 한국어 감정사전을 구축하는 방법을 제안한다. 제안된 방법은 세 단계로 구성된다. 첫 번째 단계는 한영 병렬 말뭉치를 사용하여 한영 이중언어 사전을 제작한다. 제작된 이중언어 사전은 VADER 감정어와 한국어 형태소 쌍들의 집합이다. 두 번째 단계는 그 이중언어 사전을 사용하여 한영 단어 그래프를 생성한다. 세 번째 단계는 생성된 단어 그래프 상에서 레이블 전파 알고리즘을 실행하여 새로운 감정사전을 구축한다.

3.1 한영 이중언어 사전 구축

한영 이중언어 사전은 한국어 형태소와 영어 감정어 쌍의 집합이며, 영어 감정어의 의미에 부합하는 한국어 감정어 후보를 찾기 위한 목적으로 사용된다. 공개된 한영 이중언어 사전이 존재하지 않으므로 본 논문에서는 자체적으로 한영 이중언어 사전을 구축하여 사용할 것이다. Fig. 2는 영어 감정어 'cute'를 통해서 한영 이중언어 사전을 구축하는 과정의 한 부분을 보인다. Fig. 2에서처럼 일반적인 이중언어 사전 구축 방법은 원시언어(source language)의 문맥벡터와 목적언어(target language)의 유사도를 이용해서 구축된다[20, 21]. 이와 같은 문맥벡터를 구축하기 위해서 본 논문에서는 한영 병렬 말뭉치[22]를 이용한다. 영어 감정어의 문맥벡터는 한영 병렬 말뭉치에서 영어 부분을 이용하여 영어 단어 사이의 PMI(Point-wise Mutual Information) 점수[23]를 구하여 문맥벡터로 사용하고, 한국어 형태소의 문맥벡터는 한영 병렬 말뭉치 전체를 이용해서 영어 단어와의 PMI 점수를 구하여 문맥벡터로 사용한다. 이와 같은 방법으로 구해진 두 문맥벡터의 코사인 유사도(cosine similarity)를 이용해서 대역어를 선정한다. 본 논문에서는 가장 높은 10개의 단어를 대역어로 선정한다[20, 21].

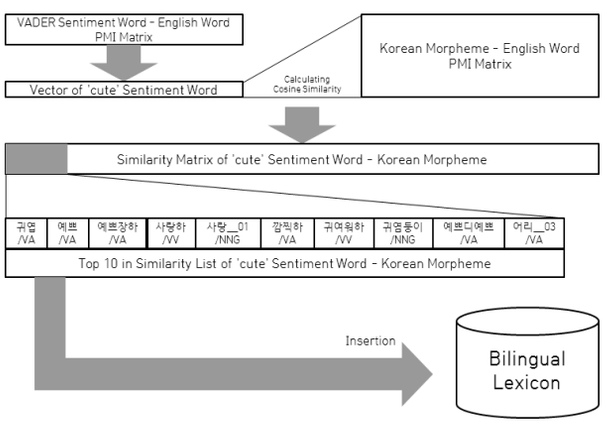

Fig. 2. An Example of Generating a Korean-English Bilingual Lexicon

3.2 한영 이중언어 단어 그래프 생성

한영 이중언어 단어 그래프는 3.1절에서 생성된 한영 이중언어 사전에 포함된 영어 감정어와 한국어 형태소 쌍을 이용하여 이중언어 단어 그래프를 생성한다. 이 그래프의 정점(node)은 영어 감정어와 한국어 형태소이다. 각 단어 사이의 간선(edge)은 두 종류로 나눌 수 있다. 첫째, 영어 감정어와 한국어 형태소 사이의 간선은 3.1절에서 구한 한영 이중언어 사전에 포함된 쌍에 대해서만 연결하고 연결강도를 1.0으로 고정한다. 둘째, 한국어 형태소 사이의 간선은 FastText[24]로 단어표상을 구하여 단어표상 간의 코사인 유사도를 연결강도로 사용한다. Fig. 3은 한국어 형태소, '사랑', '시샘하', '더

없이', '귀엽', '증오'와 영어 감정어 'cute', 'hate'에 대해서 한영 이중언어 단어 그래프를 생성한 예시를 보인다. 참고로 Fig. 3에서 각 정점의 레이블은 감정점수이며 영어 감정어를 제외하고는 모두 0.0으로 초기화하고, 3.3절에서 언급할 감정점수의 점파를 통해서 최종 감정점수를 구할 것이다.

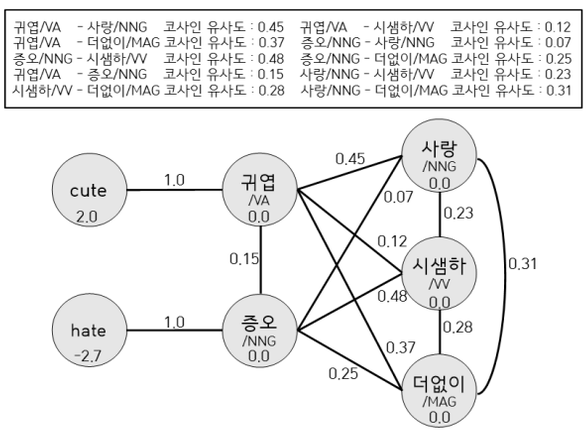

Fig. 3. An example of Initializing a Korean-English Bilingual Graph

3.3 감정 점수의 전파

레이블 전파 알고리즘[11]은 그래프 상에서 레이블이 존재하지 않는 정점에 레이블을 부여하는 알고리즘이다. 본 논문에서는 레이블 전파 알고리즘의 마지막 단계인 레이블 할당 단계를 수행하지 않고 전파 단계를 통하여 모든 정점의 감정점수가 수렴할 때까지 전파를 진행한다. 수렴된 각 정점의 감정점수가 한국어 형태소의 감정점수로 부여되며 이를 한국어 감정사전으로 사용한다.

Fig. 3에서 보듯이 영어 감정어 'cute'의 감정점수는 2.0이지만 레이블 전파를 반복적으로 수행한 결과, 한국어 형태소 '귀엽/VA'의 감정점수는 2.0로 수렴되지 않는다. 두 단어가 사전적 의미로는 매우 유사하나 완전히 동일한 감정점수로 수렴하지 않는 이유는 영어와 한국어라는 언어적 차이, 말뭉치의 한계 등이 고려된다.

4. 실험 및 평가

이 장에서는 생성된 한국어 감정사전의 유용성을 평가하기 위하여 외부평가 방법인 감정분석 시스템을 이용한다. 생성된 한국어 감정사전에 대해서 살펴보고, 실험 환경으로 감정 말뭉치와 평가 척도에 대하여 기술한다. 마지막으로 객관적인 평가를 위해서 동일한 실험 환경에서 다른 기계학습, 심층학습 모델들과의 평가를 수행한다.

4.1 생성된 한국어 감정사전의 구성

Table 1은 본 논문에서 생성된 한국어 감정사전에 포함된 감정어의 수이며, 일반적인 감정사전[6, 25]과 유사하게 부정 감정어가 긍정 감정어보다 좀 더 많은 비중을 차지한다.

Table 1. The Number of Words in the Generated Korean Sentiment Lexicon

Sentiment word	Numbers	Percentage
Positive	20,758	41.5%
Negative	29,309	58.5%
Total	50,067	100.0%

Table 2는 본 논문에서 생성된 한국어 감정사전에 포함된 감정어의 품사의 분포를 나타낸다. 명사 감정어의 비중이 약 76%를 차지하며 용어에 해당하는 동사와 형용사의 비중이 약 20%를 차지한다. 이는 일반적인 사전에 포함된 단어의 분포와도 비슷하다.

Table 2. The Distribution of Parts-of-speech of Words in the Generated Korean Sentiment Lexicon

Part-of-Speech	Numbers	Percentage
Nouns	38,073	76.1%
Verbs	7,923	15.8%
Adjectives	2,171	4.3%
Adverbs	1,557	3.1%
Interjections	343	0.7%
Total	50,067	100.0%

4.2 실험 환경

1) 개발 환경

본 절에서는 실험 및 평가를 위해서 본 논문에서 사용된 개발환경을 기술한다. 먼저 소프트웨어는 Python 3.6 Scikit-Learn과 TensorFlow를 이용하였으며 하드웨어는 CPU가 Intel Core I7-4790 3.6Hz이고 주기억장치는 3.2GB이다.

2) 감정 말뭉치

본 논문에서는 객관적인 평가를 위해 공개된 KMU 감정 말뭉치(KMU/SC)[3][4]와 네이버 영화 감정 말뭉치(NAVER/SC)[4]를 사용하며, 전자는 뉴스 댓글이며 후자는 영화평이다. Table 3은 감정 말뭉치에 포함된 문서 수를 보인다. KMU/SC는 긍정(positive), 부정(negative), 중립(neutral)로 분류되어 있고, NAVER/SC는 긍정과 부정으로만 분류되었다.

3) https://docs.google.com/file/d/0BwROrx84nPqDMjJBZmRCMlBKbEU/edit?usp=drive_web
4) https://github.com/e9t/nsmc

Table 3. Statistics of the Two Korean Sentiment Corpus

	KMU/SC	NAVER/SC
Positive	788	90,449
Negative	7,739	92,094
Neutral	301	0
Total	8,828	182,543

4.3 감정사전 기반 한국어 감정분석 시스템 구현

3장에서 생성된 한국어 감정사전의 유용성을 보이기 위해서 영어 VADER 감정분석 시스템[6]을 수정하여 한국어 감정분석 시스템(KorVADER)을 구현하였다. KorVADER는 먼저 입력된 문서를 Utagger[26]를 이용하여 형태소로 분리한다. 분리된 각 형태소는 3장에서 생성된 한국어 감정사전을 이용해서 감정점수를 부여한다. 이렇게 부여된 점수를 평균하여 각 문서의 감정값을 구한다. 이렇게 구해진 감정값은 -1.0에서 1.0 사이의 분포를 가지며 이진 분류를 수행할 때는 -1.0 이상 0.0 미만의 감정값이 측정되면 부정으로, 0.0 초과 1.0 이하의 감정값이 측정되면 긍정으로 분류하였다.

4.4 한국어 감정분석 시스템의 성능 비교

4.3에서 구현한 KorVADER 외에도 다양한 기계학습 모델들[27](Naïve Bayes(NB), Maximum Entropy(ME), Support Vector Machine-Classification(SVM-C), bi-Long Short-Term Memory(bi-LSTM))을 구현하여 성능을 비교하고 분석하였다. 참고로 KorVADER은 4.3절에서 설명한 기준에 따라 감정의 극성(긍정과 부정)을 분류하지만 다른 기계학습 모델들은 모두 분류모델로서 분류기에 의해서 바로 결정된다. Table 4는 각 시스템의 분류 정확도(accuracy)를 통한 성능을 보이고 있으며, 본 논문에서 제안한 KorVADER가 다른 모델에 비해 우수한 결과를 보이고 있다.

Table 4에서 본 논문을 제외하고는 모두 학습말뭉치가 필요하다. 평가를 위해서 학습말뭉치와 평가말뭉치를 80:20의 비율로 나누었으며 KMU/SC의 경우에 중립(neutral)은 사용하지 않았다. Table 4의 결과는 모두 같은 평가말뭉치를 대상으로 평가되었다.

Table 4. Comparison of Classification Performance between the Proposed Sentiment Analysis System (KorVADER) and Other Machine Learning-based Models

Approaches	KMU/SC	NAVER/SC
NB	0.55	0.67
ME	0.57	0.68
SVM-C	0.51	0.52
bi-LSTM	0.82	0.81
KorVADER	**0.85**	**0.88**

4.5 한국어 감정사전의 생성 방법의 비교 분석

본 절에서는 한국어 감정사전 생성 방법을 비교하고 분석하고자 한다. 2장에서 언급했듯이 감정사전 생성 방법은 사전 기반 생성 방법(dictionary-based approach), 말뭉치 기반 생성 방법(corpus-based approach)과 집단지성 기반 생성 방법(collective intelligence-based approach)으로 나뉜다. Table 5는 각 방법들의 특징을 보이고 있다. 이하에서는 Table 5에 언급된 방법들을 구체적으로 하나하나씩 기술한다.

첫 번째는 집단지성 기반으로 생성된 KOSAC(KOrean Sentiment Analysis Corpus)[28]이다. KOSAC은 세종 구문 분석 말뭉치에서 추출한 7,713개 문장에 감정분석을 수행하였고, 문장 내 등장하는 모든 형태소들을 네 가지의 극성(복잡함, 부정, 중립, 긍정)과 네 가지의 강도(높음, 중간, 낮음, 없음)를 조합하여 감정값을 부여하였다.

두 번째 역시 집단지성 기반 생성 방법으로 제작된 감정사전[17]으로 표준국어대사전[5]에 수록된 단어 중 명사, 형용사, 동사, 부사를 우선순위로 소셜네트워크 사이트 상에서 투표를 받아 단어에 감정 극성을 부여하였다. 극성은 부정, 중립,

Table. 5 Comparison of Korean Sentiment Dictionaries

Source	Method	Tools	Origin	Polarity and Score range
1. [28]	CIB	Annotation of Sentence polarity	Sejong Corpus	Polarity(Pos/Neg), Score(0 ~ 1.0)
2. [17]	CIB	Voting polarity on SNS	STD*	Polarity(Pos/Neg/Neu)
3. [29]	CIB	Voting polarity by Researchers	KET[30]	10 categories
4. [10][16]	CB	LP algorithm / word graph w. PMI	Movie Reviews	Polarity(Pos/Neg)
5. [30]	CB	Bi-LSTM	STD	11 categories, Score(2.80~10.0
Proposed	DB	LP algorithm / word graph w. FastText	VADER[6]	Score(-1.0 ~ 1.0)

※ CIB: Collective intelligence-based / CB: Corpus-based / DB: Dictionary-based
Pos: Positive / Neg: Negative / Neu: Neutral / VNeg: Very Negative / VPos: Very Positive
LP: Label propagation / PMI: Pointwise Mutual Information
STD: Standard Korean Dictionary / KET: Korean emotion terms
* https://stdict.korean.go.kr/

긍정의 세 가지를 사용하며 강도는 부여되지 않았다. [5)]

세 번째는 집단지성 기반 생성 방법으로 제작된 감정사전[29]으로 국문학 전공자와 감정연구자 12명을 통해 총 504개의 감정어 후보들을 선정하고 80명의 대학생을 대상으로 11개(슬픔, 분노, 기쁨, 놀람, 공포, 역겨움, 지루함, 관심, 고통, 중립, 기타)의 극성으로 투표하게 했다. 각 감정어별로 투표수가 가장 많은 극성이 부여되었고 해당 극성의 투표 비율에 따라 최소 2.80점에서 최대 10.0점까지 강도가 책정되었다.

네 번째는 말뭉치 기반 생성 방법으로 제작된 감정사전[10, 16]으로 특정 분야의 텍스트 자료로부터 단어를 추출하여 단어 그래프를 생성한다. 생성된 그래프의 정점은 미리 정의된 초기 감정어와 감정어 후보 단어들로 구성되며 정점 사이의 간선은 단어 사이 PMI 점수가 0이 아닐 시에 연결된다. 이러한 PMI 점수가 0.5를 초과하면 같은 극성, 미만이면 다른 극성으로 판단하여 각 정점 별로 레이블을 부여한다. 이와 같이 초기화된 그래프에 레이블 전파 알고리즘을 사용하여 감정사전을 생성한다. 감정사전 내의 감정어들은 긍정과 부정의 두 가지 중 하나의 극성만 부여받는다.

다섯 번째도 말뭉치 기반 생성 방법으로 제작된 범용 감정사전, 표준국어대사전의 명사, 형용사, 동사, 부사들의 뜻풀이(gloss)를 추출하여 연구자 세 명이 각 뜻풀이의 긍정 및 부정 여부를 사전에 정의하였다[30]. 그렇게 정의된 뜻풀이 데이터를 Bi-LSTM 모델에 학습데이터로 활용하여 감정 어휘 추출을 수행했다.

본 논문에서 제작한 감정사전은 다른 방법과 다르게 사전 기반 생성 방법을 수행했으며, 감정 강도를 수치로 부여하였다. 따라서 분석 대상이 되는 문서의 감정 극성과 강도를 가장 세부적으로 정확하게 분석이 가능하다. 또한 특정 분야에 특화된 네 번째 감정사전과 달리 범용적으로 사용이 가능하며 네 번째 감정사전은 소규모의 임의로 정의된 초기 감정어에서 다른 감정어 후보 단어들로 감정 극성을 전파한 반면 본 논문에서는 7,517개의 검증된 감정어에서 감정 점수를 전파하였다.

5. 결론 및 향후 연구

본 논문에서는 영어 감정 점수를 이용하여 한국어 감정사전을 자동생성 방법을 제안했다. VADER 감정사전과 한영 병렬 말뭉치를 활용하여 한영 이중언어 사전, 이중언어 단어 그래프를 생성하였고, 생성된 이중언어 단어 그래프 상에서 레이블 전파 알고리즘을 수렴할 때까지 반복 수행하여 한국어 감정사전을 생성하였다.

생성된 한국어 감정사전을 기반으로 새로운 감정분석 시스템(KorVADER)을 구현하여 감정분석 시스템의 성능을 통해서 생성한 감정사전을 유용성을 평가하였다. 또한 다양한 기계학습 모델(NB, ME, SVM-C, bi-LSTM-CRFs)과 성능을 비교함으로써 생성된 한국어 감정사전의 유용성을 더욱 확실히 볼 수 있었다.

향후에 KorVADER에 한국어 특성을 고려한 언어적 규칙을 추가한다면 좀 더 좋은 성능을 기대할 수 있을 것이다. 감정사전이나 감정분석 시스템에서 다루는 감정 극성의 종류, 점수의 범위 등도 역시 추가적인 연구가 필요할 것이다. 본 논문에서는 긍정과 부정의 두 가지 극성에 대해서만 감정사전을 제작하였으나 인간의 다양한 감정 종류 모형에 따른 다각적인 분석을 위한 감정사전도 필요할 것이다.

5) https://stdict.korean.go.kr/main/main.do

References

[1] TTA, Sentiment Ontology for Social Web, Telecommunications Technology Association Report TTAK.KO-10.0639/R1, 2013 (in Korean).

[2] B. Liu, "Web Data Mining," Springer, 2007.

[3] D. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University - Engineering Sciences*, Vol.30, No.4, pp.330-338, 2018.

[4] K.-J. Lee, J.-H. Kim, H.-W. Seo, and K.-S. Ryoo, "Feature weighting for opinion classification of comments on news articles," *Journal of the Korean Society of Marine Engineering*, Vol.34, No.6, pp.871-879, 2010 (in Korean).

[5] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *Proceedings of the Association for Computational Linguistics*, pp.417-424, 2002.

[6] E. Gilbert and C. J. Hutto, "VADER : A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the 8th International Conference on Weblogs and Social Media*, pp.216-225, 2014.

[7] M. Taboada and J. Brooke, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, Vol.37, No.2, pp.272-274, 2011.

[8] Ł. Augustyniak, P. Szymański, T. Kajdanowicz, and W. Tuligłowicz, "Comprehensive study on lexicon-based ensemble classification sentiment analysis," *Entropy*, Vol. 18, No.1, pp.1-29, 2015.

[9] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," arXiv:1801.07883, 2018.

[10] C. Heo and S.-Y. Ohn, "A novel method for constructing sentiment dictionaries using word2vec and label propagation," *The Journal of Korean Institute of Next Generation Computing*, Vol.13, No.2, pp.93-101, 2017 (in Korean).

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

[12] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, and K. Miller. "WordNet: An online lexical database," *International Journal of Lexicograph*, Vol.3, No.4, pp.235-244, 1990.

[13] A. Hassan, V. Qazvinian, and D. Radev, "What's with the attitude? identifying sentences with attitude in online discussion," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.1245-1255, 2010.

[14] E. C. Dragut, C. Yu, P. Sistla, and W. Meng, "Construction of a sentimental word dictionary," *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp.1761-1764, 2010.

[15] S. Mohammed, C. Dunne, and B. Dorr, "Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.599-608, 2009.

[16] J.-H. Kim, "The Graph-based Method for Construction of Domain-oriented Sentiment Dictionary," MS. Thesis, Dept. of Computer Engineering, Korea Aerospace University, Seoul, Republic of Korea, 2015 (in Korean).

[17] J. An and H.-W. Kim, "Building a Korean sentiment lexicon using collective intelligence," *Journal of Intelligence System*, Vol.21, No.2, pp.49-67, 2015 (in Korean).

[18] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of LIWC2007," Austin, TX: LIWC net, 2007.

[19] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 417-422, 2006.

[20] H.-W. Seo, H.-S. Kwon, and J.-H. Kim, "Extended pivot-based approach for bilingual lexicon extraction," *Journal of the Korean Society of Marine Engineering*, Vol.38, No. 5, pp.557-565, 2014.

[21] J.-H. Kim, H.-S. Kwon, and H.-W. Seo, "Evaluating a pivot-based approach for bilingual lexicon extraction," *Computational Intelligence and Neuroscience*, Vol.2015, pp.1-13, 2015.

[22] H.-W. Seo, H.-C. Kim, H.-Y. Cho, J.-H. Kim, and S.-I. Yang, "Automatically constructing Korean-English parallel corpus from Web documents," *Proceedings of the 26th KIPS Fall Conference*, Vol.13, No.2, pp.161-164, 2006 (in Korean).

[23] K. W. Church and P. Hanks, "Word association norms, mutual information and lexicography," *Computational Linguistics*, Vol.16, No.1, pp.22-29, 1990.

[24] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, Vol.5, pp.135-146, 2017.

[25] T. Loughran, and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries and 10-Ks," *The Journal of Finance*, Vol.66, No.1, pp.35-66, 2011.

[26] J.-C. Shin and C.-Y. Ock, "A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary," *Journal of The Korean Institute of Information Scientists and Engineers*, Vol.39, No.5, pp.415-424, 2012 (in Korean).

[27] A. Geron, "Hands-On Machine Learning with Scikit-Learn & TensorFlow," O'Reilly, 2017.

[28] M.-H. Kim, Y.-M. Jo, H.-Y. Jang, and H.-P. Shin, "KOSAC: Korean Sentiment Analysis Corpus," *Proceedings of the Korean Institute of Information Scientists and Engineers*, pp.650-652, 2013 (in Korean).

[29] S. Sohn, M.S. Park, J.-E. Park, J.-H. Sohn, "Korean emotion vocabulary: Extraction and categorization of feeling words," *Korean Journal of the Science of Emotion & Sensibility*, Vol.15, No.1, pp.105-120, 2012 (in Korean).

[30] S.-M. Park, C.-W. Na, M.-S. Choi, D.-H. Lee, and B.-W. On, "KNU Korean sentiment lexicon: Bi-LSTM-based method for building a Korean sentiment lexicon," *Journal. of the Intelligence Information Systtem*, Vol.24, No.4, pp. 219-240, 2018 (in Korean).

[31] I. J. Park and K. H. Min, "Making a list of Korean emotion terms and exploring dimensions underlying them," *Korean Journal of Social and Personality Psychology*, Vol.19, No. 1, pp.109-129, 2005 (in Korean).

박 호 민

https://orcid.org/0000-0001-7324-387X
e-mail : homin2006@hanmail.net
2017년 한국해양대학교 컴퓨터정보공학과 (학사)
2019년 한국해양대학교 컴퓨터공학과 (석사)
2019년 ~ 현 재 한국해양대학교 컴퓨터공학과 박사과정
관심분야 : Natural Language Processing, Information Retrieval, Sentiment Analysis

김 창 현

https://orcid.org/0000-0001-7692-0733
e-mail : chkim@etri.re.kr
1991년 홍익대학교 전계계산학과(학사)
1993년 한국과학기술원 전산학과(석사)
2001년 한국과학기술원 전산학과(박사수료)
2001년 ~ 현 재 한국전자통신연구원
인공지능연구소 책임연구원
관심분야 : Natural Language Processing, Machine Translation, Dialogue Processing

김 재 훈

https://orcid.org/0000-0001-8655-2591
e-mail : jhoon@kmou.ac.kr
1986년 계명대학교 전계계산학과(학사)
1988년 한국과학기술원 전산학과(석사)
1996년 한국과학기술원 전산학과(박사)
1988년 ~ 1997년 한국전자통신연구원
선임연구원
2001년 ~ 2002년 Information Sciences Institute USC
방문연구원
2007년 ~ 2008년 Beckman Institute UIUC 방문연구원
1997년 ~ 현 재 한국해양대학교 컴퓨터공학과 교수
관심분야 : Natural Language Processing, Information Retrieval, Corpus Linguistics, Sentiment Analysis