감성과학
제18권 1호, 2015
<연구논문>

pISSN 1226-8593
eISSN 2383-613X
Sci. Emot. Sensib.,
Vol.18, No.1, pp.103-110, 2015
http://dx.doi.org/10.14695/KJSOS.2015.18.1.103

그래프 기반 준지도 학습 방법을 이용한 특정분야 감성사전 구축

The Construction of a Domain-Specific Sentiment Dictionary Using Graph-based Semi-supervised Learning Method

김정호*† · 오연주* · 채수환**

Jung-Ho Kim*† · Yean-Ju Oh* · Soo-Hoan Chae**

***한국항공대학교 컴퓨터공학과**

***Department of Computer Engineering, Korea Aerospace University**

****한국항공대학교 전자 및 정보통신공학부**

****The School of Electronics and Telecommunication, Korea Aerospace University**

Abstract

Sentiment lexicon is an essential element for expressing sentiment on a text or recognizing sentiment from a text. We propose a graph-based semi-supervised learning method to construct a sentiment dictionary as sentiment lexicon set. In particular, we focus on the construction of domain-specific sentiment dictionary. The proposed method makes up a graph according to lexicons and proximity among lexicons, and sentiments of some lexicons which already know their sentiment values are propagated throughout all of the lexicons on the graph. There are two typical types of the sentiment lexicon, sentiment words and sentiment phrase, and we construct a sentiment dictionary by creating each graph of them and infer sentiment of all sentiment lexicons. In order to verify our proposed method, we constructed a sentiment dictionary specific to the movie domain, and conducted sentiment classification experiments with it. As a result, it have been shown that the classification performance using the sentiment dictionary is better than the other using typical general-purpose sentiment dictionary.

Key words: sentiment lexicon, sentiment dictionary, sentiment classification, sentiment word, sentiment phrase, graph, semi-supervised learning

요 약

감성어휘는 텍스트로 감성을 표현하거나, 반대로 텍스트로부터 감성을 인식하기 위한 특징으로써 감성분류 연구에 필수요소이다. 본 연구는 감성어휘의 집합인 감성사전을 자동으로 구축하는 그래프 기반 준지도 학습 방법을 제안한다. 특히 감성어휘가 사용되어지는 분야에 따라 그 감성이 변하는 중의성 문제를 고려하여 분야 별 감성사전을 구축하고자 한다. 제안하는 방법은 어휘와 어휘들 간의 밀접도를 토대로 그래프를 구성하고, 사전에 학습 된 일부 소량의 감성어휘들의 감성을 구성된 그래프 전체에 전파하는 방식으로 모든 어휘의 감성을 추론한다. 감성어휘는 대표적으로 감성단어와 감성구문이 있으며, 본 연구에서는 이들 각각에 대한 그래프를 구성하고 감

※ 본 논문은 미래창조과학부 및 정보통신산업진흥원의 ICT융합고급인력과정지원사업(NIPA-2014-H0401-14-1021)의 연구결과로 수행되었음.

† 교신저자 : 김정호 (한국항공대학교 컴퓨터공학과)
E-mail : natul2@kau.ac.kr

성을 추론하여 전체 감성사전을 구축하였다. 제안하는 방법의 성능을 검증하기 위해 영화평 분야의 감성사전을 구축하고, 이를 이용한 영화평 감성분류 실험을 수행하였다. 그 결과 기존 범용 감성사전의 어휘들을 이용한 감성분류보다 더 높은 분류 성능을 확인하였다.

주제어: 감성어휘, 감성사전, 감성분류, 감성단어, 감성구문, 그래프, 준지도 학습

1. 서론

요즈음 스마트폰과 같은 정보통신기기의 발달과 보급으로 남녀노소 누구나 시간과 장소에 불문하고 자신의 의견을 온라인 텍스트로 표현할 수 있다. 온라인 텍스트에 담긴 의견 정보는 점점 더 발전하는 정보화 시대에서 매우 가치 있는 지적 자원으로 여겨지고 있으며, 기업의 위기관리(risk management)나 소비자의 의사결정(decision making)과 같이 여러 분야에서 다양한 용도로 사용되어진다. 온라인 텍스트 데이터가 담고 있는 사람들의 의견을 인식하고 긍정과 부정 두 감성 범주로 분류하는 것을 텍스트 감성 분석이라 한다.

텍스트 감성 분석은 주로 텍스트의 감성어휘(sentiment lexicon)를 단서로, 텍스트에 내포된 감성을 인식하고 분류한다. 이들은 텍스트가 가지는 감성을 대표하는 특징이라 하여 감성특징(sentiment feature)라고도 부른다. 대표적인 감성어휘로는 감성단어(sentiment word)와 감성구문(sentiment phrase)이 있다. 감성단어는 감성을 직접 표현하기 위해 사용되는 단일 단어로 '좋다' 또는 '싫다'와 같은 단어가 이에 해당된다. 감성구문은 단일 단어로는 직접적으로 감성을 표현하지는 못하지만 두 단어 이상이 의미상으로 결합하여 특정 감성을 표현하는 구문으로 '화면이 밝다', '평점이 낮다'와 같은 구문이 감성구문의 예이다.

감성어휘는 특정 감성을 대표하는 특징이기 때문에 정확한 감성 분석을 위해서는 사전에 잘 정의된 양질의 감성어휘들이 필요하다. 이들을 감성어휘집합(sentiment lexicon set) 또는 감성사전(sentiment dictionary)라 하며, 각 감성어휘와 감성어휘가 가지는 감성 쌍의 집합이다. WordNet을 기반으로 만든 Senti-WordNet이 대표적인 감성사전이다.(Baccianella et al., 2010) 하지만 위 감성사전은 분야에 상관없이 범용으로 만들어진 사전으로, 감성어휘가 사용되어지는 분야에 따라 표현하는 감성이 변하는 언어의 중의성을 전혀 고려하지 않는다. 예를 들면, '슬프다'와 같은 단어는 상품 평에서는 주로 부정의 감성을 표현하지만 영화 평에서는 보통 긍정의 감성을 표현할 때 사용된다. 이를 고려하지 않을 경우 특정 분야에서 높은 감성 분석 성능을 기대하기 어렵다.

범용 감성사전이 가지는 한계를 극복하기 위해 특정 분야에 맞춘 감성사전을 구축하는 많은 연구가 수행되어졌다. 대부분 전통적인 기계학습 방법을 이용하여 일반 사전과 양질의 학습 데이터로부터 어휘들을 수집하고, 어휘들 사이의 관계로부터 각 어휘의 감성을 추측하여 감성어휘를 학습하였다. 하지만 이들은 단어가 가지는 사전적 의미와 유의어 관계를 바탕으로 감성어휘를 학습하기 때문에 정작 특정 분야에서 어휘들이 가지는 서로간의 관계를 고려할 수 없다.

본 연구는 어휘들이 가지는 사전적 의미와 유의어 관계가 아닌 특정 분야에서의 어휘들 사이의 밀접도를 고려하여 감성어휘를 학습하고 감성사전을 구축하는 그래프 기반 준지도적 학습 방법을 제안한다. 제안하는 방법은 사전과 학습 데이터로부터 추출한 어휘들과 어휘들 간의 밀접도를 이용하여 그래프를 구성하고, 사전에 정의한 소수의 감성어휘들의 감성을 어휘들 간의 밀접도를 통해 전파하는 레이블 전파(label propagation)방법에 기반 한다.(Zhu & Ghahramani, 2002) 다양한 분야에서 사전에 잘 지도 된 양질의 학습 데이터를 얻기 힘들기 때문에 제안하는 준지도적 학습 방법이 분야별 감성사전을 만드는데 효율적이다. 그리고 기존 그래프 기반 방법들이 사용하는 어휘들 간의 사전적 의미관계가 아닌 사용되는 분야에서의 밀접도를 고려하기 때문에 특정 분야에 더욱 적합한 감성어휘들을 학습할 수 있다. 본 연구는 제안하는 방법을 이용하여 영화평 분야의 감성사전을 구축하고 감성분석을 실험을 수행해 봄으로써 그 성능을 검증하였다.

2. 관련연구

감성어휘는 텍스트 내에서 긍정 또는 부정의 감성을 표현하기 위해 사용되어진다. 그렇기 때문에 텍스트 감성분석에서 감성어휘의 역할이 매우 크다. 감성어휘들의 집합을 감성사전이라 하고 양질의 감성어휘들을 가진 감성사전을 사용할수록 좋은 감성분석 결과를 기대할 수 있다. 대표적인 감성사전으로는 Senti-WordNet이 있다.(Baccianella et al., 2010) 유의어 사전인 WordNet의 각 어휘에 특정 감성을 부여하여 만든 사전으로, 감성어휘뿐만 아니라 어휘들 사이의 유의어 관계도 포함하고 있어 다양한 감성분석 연구에서 사용하였다. 그 외에도 여러 연구에서 WordNet을 기반으로 하여 감성사전을 구축하였다.(Esuli & Sebastiani, 2005, Kamps et al., 2004, Andreevskaia & Bergler, 2006) 하지만 이들 감성사전은 어휘가 특정 분야에서 사용될 시 표현하는 감성이 변할 수 있는 중의성을 고려하지 않은 범용 감성사전이다. 범용 감성사전을 이용하였을 경우 특정 분야에서 좋은 분석 성능을 기대하기 어렵다.

이러한 문제를 해결하기 위해 특정 분야에 맞춘 감성어휘의 감성을 파악하는 방법들이 제안되었다. 이들 방법은 크게 사전기반 방식과 말뭉치 기반 방법으로 구분할 수 있다. 두 방법은 모두 어휘의 감성을 다른 감성어휘로부터 추론하는 방법이다. 단지 추론하고자 하는 대상인 어휘의 출처에 차이가 있으며, 또한 감성 추론 시에 사전적인 의미 관계를 고려했는지 여부에 차이가 있다.

사전기반 방법의 대표적인 사례는 (Hu & Liu, 2004)에서 보인다. 소수의 감성어휘를 기준으로, WordNet으로부터 이들의 유의어를 추출하여 새로운 감성어휘를 정의하였다. 그 외의 사전기반 방법은 사전에 정의한 감성어휘와 WordNet으로부터 구한 이들의 유의어로 그래프를 구성한 후 다양한 그래프 기반 방법들로 감성어휘를 구하였다.(Blair-Goldensohn et al., 2008, Rao & Ravichandran, 2009, Hassan et al., 2010) 사전기반 방법은 사전으로부터 쉽고 빠르게 대량의 감성어휘를 구할 수 있다는 장점이 있지만 의미적인 관계만을 고려하기 때문에 실제 특정 분야에서 단어들끼리 가지는 관계를 고려하진 못한다.

말뭉치 기반 방법은 말뭉치로부터 추출한 어휘들의 감성을 추론하기 위해, 이들과 감성어휘간의 관계를 다양한 요소로부터 추측하여 어휘의 감성을 추론한다. Hatzivassiloglou와 Mckeown(1997)은 감성어휘와 ‘and’, ‘or’, ‘but’ 등과 같은 접속사로 연결 된 다른 어휘들을 추출하고 접속 관계에 따라 추출한 어휘의 감성을 추론하였다. Qiu 등(2009)은 특정 객체와 관련된 어휘들을 추출하고 클러스터링을 통해 감성어휘 집합을 생성하였다. Turney(2002)는 말뭉치에서 대표적인 긍정, 부정 어휘와 공기관계를 가지는 어휘들을 추출하고 이들의 감성을 추론하였다. 말뭉치 기반 방법은 분야 별 말뭉치로부터 어휘들의 타당한 감성을 추론할 수 있다. 하지만 접속사, 또는 특정 객체와 같이 고려하는 요소들이 실제 말뭉치 내에서 발생하는 경우가 드물고, 감성을 추론할 때 직접적으로 관련 있는 감성어휘와의 단편적인 관계에만 의존한다.

Tai와 Kao(2013)는 사전 기반과 말뭉치 기반 방법을 혼합한 하이브리드 방식으로 감성사전을 구축하였다. 사전과 말뭉치로부터 각각 구한 단어 간의 유사도를 모두 고려하여 그래프를 생성하고, 레이블 전파 방법을 이용해 단어의 감성을 추론하였다. 하지만, 말뭉치 보다는 여전히 사전에 많이 의존하였기 때문에, 한글을 포함한 다른 많은 언어에 대해서 WordNet과 같이 공개된 유의어 사전이 없는 경우 사용하기 어렵다. 그리고 구문을 고려하지 않은 단어에 국한된 방식으로 보다 정확한 감성분석에 제한적인 감성사전을 구축하였다.

본 연구에서는 분야에 따른 어휘의 보다 정확한 감성을 파악하기 위해, 특정 분야에서 모든 어휘들 사이의 관계를 말뭉치로부터 추측하고 이를 그래프로 표현하였다. 그리고 그 위에서 어휘들이 가지는 감성을 인접한 어휘들로 전달하여 감성을 추론함으로써 단편적인 감성 추론이 아닌 전체적인 어휘들 사이의 관계를 고려한 감성 추론을 수행하였다.

3. 분야별 감성사전 구축

본 장은 자동으로 감성사전을 구축하기 위해 제안하는 그래프 기반 준지도적 학습 방법을 소개한다. 감성어휘인 감성단어와 감성구문 각각에 대한 그래프를 구성하고, 이들 위에서 특정 감성어휘의 감성을 다른 어

휘들로 전파하여 모든 감성어휘들의 감성을 추론한다.

3.1. 그래프 구성

3.1.1. 단어 그래프 구성

단어 그래프는 단어와 단어들 사이의 밀접도를 각각 정점(vertex)과 변(edge)으로 정의한 그래프이다. 단어 그래프 $G_w = (V, E)$는 다음과 같이 정의한다.

$V = \{x_1,...x_l,x_{l+1},...,x_n\},\ x_i \in X,\ n = |X|$
$E = \{(x_i,x_j) \mid x_i,x_j \in X, i \neq j\}$
$S^w = \{s_1^w,...,s_n^w\},\ n = |X|$

노드 V는 단어 집합 X의 단어들로 구성된 집합이며, 두 개의 하위집합 $L = \{x_1,...,x_l\}$과 $U = \{x_{l+1},...,x_n\}$로 구분된다. 집합 L은 사전에 감성을 알고 있는 소수의 감성단어 집합이고, 집합 U는 아직 감성을 알지 못하는, 감성을 추론해야하는 단어 집합이다. S^w의 s_i^w는 단어 x_i가 가지는 감성 값이며, -1과 1 사이의 실수를 갖는다. 초기에는 감성을 미리 알고 있는 집합 L의 단어들만 긍정 부정에 따라 양수 또는 음수의 특정 감성 값을 가지며, 나머지 집합 U의 단어들은 감성 값을 모르며 초기 값으로 0을 가진다.

변 E는 밀접한 관계를 가지는 두 단어 쌍의 집합이다. 각 변의 가중치는 두 단어 사이의 밀접도이며 다음 식1의 PMI(Pointwise Mutual Information)로 계산되어진다. PMI는 다량의 텍스트 데이터 내에서 서로 다른 두 단어의 공기(co-occurrence) 정보를 토대로 이들 사이의 밀접도를 계산한다.

$$w_{ij} = PMI(x_i,x_j) = \log \frac{p(x_i,x_j)}{p(x_i)p(x_j)} \quad (1)$$

w_{ij}는 두 단어 x_i와 x_j 사이에 정의된 변의 가중치인 밀접도이며, 밀접도가 존재할 경우($w_{ij} > 0$) 하나의 변으로 정의하며, 밀접한 정도는 계산된 밀접도의 크기에 따른다.

3.1.2. 구문 그래프 구성

구문 그래프는 단어 그래프와 동일하게 단일 단어를 하나의 정점으로 정의한다. 하지만 하나의 변은 단어 그래프와 다르게 두 단어 사이가 하나의 구문을 이루는지 여부로 변을 정의한다. 구문 그래프 $G_p = (V, E)$는 다음과 같이 정의한다.

$V = \{x_i \mid x_i \in X\}$
$E = \{(x_i,x_j) \mid x_i,x_j \in X, i \neq j\}$
$S^p = \{s_1^p,...,s_m^p\},\ m = |E|$

구문 그래프는 단어 그래프와 동일하게 단어를 정점으로 정의한 집합 V을 가진다. 반면, 집합 E의 각 변은 단어들 간의 밀접 관계가 아닌 두 단어가 하나의 구문을 이루는 구문 관계로 정의된다. 그리고 구문의 감성 S^p는 구문을 이루는 두 단어 사이의 변에 존재한다. s_i^p는 i 번째 변(구문)의 감성을 의미한다.

3.2. 감성 전파

3.2.1. 단어의 감성 전파

본 연구에서는 단어 그래프를 구성하는 모든 단어의 감성을 추론하기 위해 LP(Label Propagation) 방법을 이용한다.(Zhu & Ghahramani, 2002) 이는 서로 인접한 단어들은 서로 유사한 감성을 가질 것이라는 가정 하에, 각 단어가 가지고 있는 감성을 인접한 다른 단어들로 전파함으로써 인접한 단어들이 유사한 감성을 가지도록 한다.

한 단어의 감성은 다른 인접 단어들로부터 전달받은 감성의 가중평균(weight average)으로 계산된다. 이때, 가중평균의 가중치는 인접한 단어들과의 밀접도를 일반화한 확률 값을 사용한다. 식2는 단어 x_i와 x_j 사이의 밀접도를 일반화하는 계산식이고, 식3은 이를 가중치로 사용하여 단어 x_i의 감성 s_i을 계산하는 식이다.

$$s_i^w = \sum_{j \in A_i} t_{ij}\, s_j^w \quad (2)$$

s_i : 단어 x_i가 가지는 감성 ($s_i^w \leftarrow [-1,1]$)
A_i : 단어 x_i와 인접한 단어들의 집합 ($i \notin A_i$)
t_{ij} : 단어 x_i와 x_j 사이의 일반화된 밀접도(가중치)

$$t_{ij} = P(j \rightarrow i) = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}} \quad (3)$$

각 단어의 감성은 식2를 통해 특정 감성 값에 수렴

할 때 까지 반복 계산되어진다.

3.2.2. 구문의 감성 전파

구문의 감성은 구문을 이루는 두 단어의 관계에 존재한다. 그렇기 때문에 단어의 감성 전파와는 달리 구문의 감성 전파는 다음의 원리에 따라 이루어진다. 세 단어 A, B, C가 서로 A-B와 A-C 두 구문을 이룰 때, 두 구문의 감성은 단어 B와 C의 밀접 관계를 통해 서로에게 전달된다. 즉, 단어 그래프의 단어들 간 밀접 관계를 참고하여 구문의 감성을 전파한다. 예를 들면, (평점, 낮다)와 (평점, 떨어지다) 두 구문이 있고 '낮다'와 '떨어지다' 두 단어가 서로 밀접한 관계가 있을 때, 두 구문은 서로 유사한 감성을 가지며 각자의 감성을 서로에게 전파한다. 두 구문이 공통으로 가지고 있는 단어 '평점'을 제외하고 보았을 경우 단어 '낮다'와 '떨어지다' 두 단어 사이의 감성 전파와 동일하게 볼 수 있다.

즉, 한 구문의 감성은 공통 단어를 가지는 구문들의 감성의 가중평균으로 구할 수 있다. 이때 가중평균을 구하기 위해 사용하는 가중치는 공통 단어를 제외한 나머지 단어들 간의 일반화 된 밀접도이다. 식4는 단어 x을 공통 단어로 가지는 구문들$\{(x,x_1),\ (x,x_2),\ \cdots,(x,x_m)\}$의 감성을 구하기 위한 계산식을 보이며, 이를 모든 구문에 적용하여 전체 구문의 감성을 구한다.

$$s_i^p = \sum_{k \in B} t_{ij}\, s_j^p \qquad (4)$$

s_i^p : 단어 x를 포함하는 i 번째 구문 (x,x_i)이 가지는 감성 ($s_i^p \leftarrow [-1,1]$)

B : 단어 x와 구문 관계를 가지는 단어들의 집합

t_{ij} : 단어 x_i와 x_j 사이의 일반화된 밀접도(가중치)

(x_i: 단어 x를 포함하는 i번째 구문의 나머지 단어)

4. 감성사전 구축 및 감성 분석 실험

4.1. 실험 개요

본 연구에서는 제안하는 방법의 성능을 검증하기 위해 영화평 분야의 감성사전을 구축하고, 이를 이용한 영화평의 감성분석을 수행하였다. 실험데이터로 네이버 랩(lab.naver.com/research)에서 제공하는 오피니언 마이닝 연구 데이터인 영화평을 사용하였다. 각 영화평은 1~10의 평점을 가진다. 본 연구에서는 평점이 1~2인 영화평을 부정 영화평으로, 평점 9~10인 영화평을 긍정 영화평으로 사용하였다. 긍정 영화평 1,000개와 부정 영화평 1,000개, 총 2,000개의 영화평 중 특정 감성으로의 치우침(bias)을 방지하기 위해 각각 500개씩의 긍정, 부정 영화평을 학습 데이터로 사용하여 감성사전을 구축하였으며, 나머지 1,000개의 영화평 중 임의의 500개 영화평으로 감성분석을 수행하였다. 감성사전 구축 및 감성분석 실험의 구성 및 과정은 다음 그림 1과 같다.

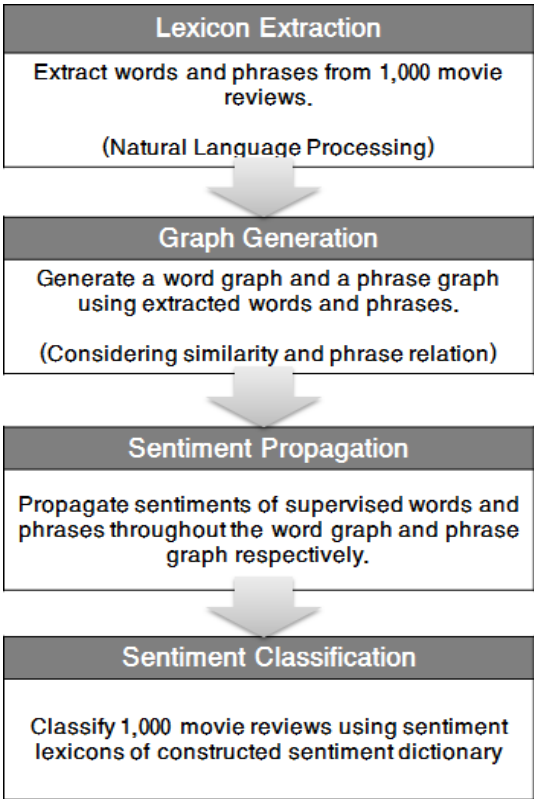

Figure 1. The process of the sentiment dictionary construction and sentiment analysis experiment

4.2. 실험 결과

4.2.1. 영화평 분야 감성사전 구축

영화평 분야의 감성사전을 구축하기 위해 학습 데이터로부터 단어와 구문을 추출하였다. 본 연구에서는 감성을 직접 표현할 수 있는 형용사, 동사, 명사 품사의 단어만 학습 데이터로부터 추출하여 단어 그래

프 및 구문 그래프를 구성하였다. 추출한 단어와 구문 중 자명한 감성을 가지는 일부 단어와 구문을 감성단어와 감성구문으로 정의하였다. 표 1은 단어와 구문 그래프를 구성하기 위해 추출한 단어와 구문의 수를 보인다. 그리고 표 2와 표 3은 각각 정의한 소수의 감성단어와 감성구문을 보인다. 28개의 단어를 사전에 정의한 감성단어 집합(L)으로 정의하였고, 60개의 구문을 사전에 정의한 감성구문 집합으로 정의하였다.

Table 1. The number of extracted words and phrases

Words				Phrases
Noun (NNG)	Verb (VV)	Adjective (VA)	Total	2,430
1,017	296	81	1,394	

Table 2. The examples of pre-defined sentiment words

Sentiment (count)	Sentiment Words
Positive (14)	좋/VA, 재밌/VA, 괜찮/VA, 슬프/VA, 좋아하/VV, 재미있/VA, 이쁘/VA, 돋보이/VV, 슬푸/VA, 끝내주/VV, 매력적/NNG, 멋있/VA, 짱하/VA, 강추/NNG,
Negative (14)	어설프/VA, 낚이/VV, 재미없/VA, 최악/NNG, 열받/VV, 짜증/NNG, 시끄럽/VA, 뻔하/VA, 거슬리/VV, 욕먹/VV, 나쁘/VA, 쪽팔리/VV, 지겹/VA, 싫/VA

Table 3. The examples of pre-defined sentiment phrase

Sentiment (count)	Sentiment Phrases
Positive (30)	(기대/NNG, 크/VA), (장면/NNG, 슬프/VA), (영화/NNG, 짠하/VA), (감동/NNG, 느끼/VV), (스토리/NNG, 괜찮/VA), (눈물/NNG, 나/VV) … (스릴/NNG, 있/VV), (표현력/NNG, 돋보이/VV), (소름/NNG, 돋/VV), (사랑이/NNG, 넘치/VV), (만점/NNG, 주/VV), (스트레스/NNG, 풀리/VV), (감명/NNG, 깊/VA),
Negative (30)	(기대/NNG, 다르/VA),(영화/NNG, 부풀리/VV), (돈/NNG, 아깝/VA), (억지/NNG, 자아내/VV), (감동/NNG, 약하/VA), (시간/NNG, 아깝/VA) … (손발/NNG, 오그라들/VV), (기분/NNG, 드럽/VA), (긴장감/NNG, 없/VA), (인위적/NNG, 지나치/VA), (몰입/NNG, 어렵/VA), (창의성/NNG, 없/VA), (한계/NNG, 느껴지/VV)

구성한 단어 그래프와 구문 그래프로부터 제안하는 감성 전파 방법에 따라 모든 단어와 구문의 감성을 추론하였다. 사전에 정의한 소수의 감성단어가 가지는 감성 값 s을 인접한 다른 단어들로 전파하는 것을 시작으로 각 단어가 가지는 감성을 반복적으로 전파하여 모든 단어의 감성을 추론하였다. 구문도 마찬가지로 소수의 감성구문의 감성을 다른 연관된 구문으로 전파하여 모든 구문의 감성을 추론하였다. 제안하는 방법을 이용한 감성단어와 감성구문의 집합인 감성사전의 구축 실험 결과는 다음 표 4와 표 5에서 제시한다.

Table 4. The distribution of sentiments of words after sentiment propagation

	Positive words	Negative words	Neutral words	Total
# of words	741	642	11	1394
ratio	53%	46%	1%	100%

Table 5. The distribution of sentiments of phrases after sentiment propagation

	Positive phrases	Negative phrases	Neutral phrases	Total
# of phrases	1004	1382	44	2430
ratio	41%	57%	2%	100%

감성이 없는 중립 단어와 구문은 실제로 감성이 없는 어휘여서 다른 감성어휘와 연결이 없거나 학습 데이터가 충분하지 않아 다른 감성어휘들과의 관계를 발견하지 못한 경우이다. 본 연구에서는 이러한 중립 어휘들을 제외하고 나머지 감성어휘들로 감성사전을 구축하였다.

4.2.2. 영화평 감성 분석

본 연구에서는 영화평의 감성을 인식하고 분류하기 위해 대표적인 기계학습 방법인 SVM(Support Vector Machine)과 RF(Random Forest), 그리고 NNet(Neural Network)을 사용하였다. 감성사전의 감성어휘들을 특징으로 학습과 실험 데이터를 표현하였다. 그리고 학

습 데이터를 통해 각 분류기의 모델을 학습하고, 이를 이용하여 실험 데이터의 영화평의 감성을 분류하였다. 커널 함수가 가우시안(gaussian) 함수인 비선형 SVM, 트리의 수가 500개, 변수의 수가 10개인 RF, 은닉 층(hidden layer)의 수가 10개인 NNet을 각각 사용하였다.

제안하는 방법으로 구축한 감성사전의 성능을 검증하기 위해 기존 범용 감성사전과의 분류 성능을 비교하였다. 성능 평가 척도로는 재현율(recall), 정밀도(precision), 그리고 F-measure를 사용하였다. 이들은 분류 정확도를 측정하는 대표적인 평가 척도이며 아래의 분류 행렬(confusion matrix)과 식5-7을 통해 계산된다.

Table 6. Confusion matrix

		Predicted Label	
		True	False
Known Label	True	a	b
	False	c	d

$$\text{precision} = \frac{a}{a+c} \quad (5)$$

$$\text{recall} = \frac{a}{a+b} \quad (6)$$

$$\text{F-measure} = \frac{2\times(precision\times recall)}{precision+recall} \quad (7)$$

표 7은 영화평의 각 분류기 별 분류 성능을 보인다. 분류 성능은 5-겹 교차 검증(5-fold cross validation)으로 제시하였다. 각 분류기의 설정은 SVM은 대표적인 커널 함수 RBF(Radial Basis Function)을 사용하였으며, RF는 500개의 트리와 10개의 변수를 사용하였다. 그리고 NNet은 히든 레이어 노드의 수를 10개로 설정하였다.

범용사전과 본 연구에서 구축한 감성사전 각각을 이용한 감성분석 결과를 보면 다음의 특징이 있다. NNet의 분석 결과를 제외하고는 범용사전을 사용한 경우는 재현율이 정밀도보다 대체적으로 높고, 반면에 구축한 감성사전을 사용한 경우는 반대로 정밀도가 재현율에 비해 높다. 하지만 재현율과 정밀도 사이의 차이가 범용사전보다 구축한 사전이 작고, 그러므로 더 높은 F-값을 보였다. 이 결과로부터 영화평 분야에서 구축한 감성사전을 사용하였을 시 더욱 안정되고 정확한 감성분석을 수행하였음을 확인할 수 있다.

Table 7. The results of sentiment analysis for movie reviews using the proposed sentiment dictionary

Classifier	Evaluation measures	Classification Performance(%)	
		general dictionary	Our dictionary
RF	Recall	86.76	71.95
	Precision	57.28	85.51
	F-measure	69.01	78.15
SVM	Recall	86.76	62.20
	Precision	55.66	93.10
	F-measure	67.82	74.57
Nnet	Recall	66.18	85.37
	Precision	54.88	78.65
	F-measure	60.00	81.87

5. 결론 및 향후 연구

본 연구는 보다 정확한 감성분석을 수행하기 위해 필요한 감성사전을 구축하는 방법을 제안하였다. 제안하는 방법은 분야별로 어휘가 가지는 의미와 감성이 변하는 중의성 문제를 해결하고, 제한적인 학습 데이터 내에서 각 분야의 특성을 살린 감성사전을 구축하는 준지도적 학습 방법이다.

분야에 따라 감성어휘가 가지는 감성을 보다 정확히 추론하기 위해, 본 연구는 특정 분야에서 어휘들 간의 관계를 구하고, 구한 관계를 통해 모든 어휘의 감성을 추론하는 방법을 제안하였다. 제안하는 방법을 검증하기 위해 영화평 분야의 감성사전을 구축하고, 구축한 감성사전을 이용해 영화평 감성 분석 실험을 수행하였다. 그 결과 기존 범용 감성사전의 감성어휘를 사용하였을 때보다 더욱 안정적이고 정확히 영화평의 감성을 분석하였다.

본 연구에서 제안하는 방법은 특정 분야에서 어휘들 간의 관계를 기반으로 서로 밀접한 어휘들은 유사한 감성을 가지도록 한다. 하지만 어휘들 간의 관계를 파악하기 위해 단순히 어휘들의 공기 정보만을 고려한다. 공기 정보만으로는 두 단어가 가지는 감성이 같은지 다른지를 명확히 알기 어렵다. 추후 연구에서 공기 정보뿐만 아니라 두 어휘가 가지는 감성의 같음과

다름도 고려한다면 더욱 정확한 어휘의 감성을 추론할 수 있을 것이라 기대한다.

REFERENCES

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining*. Paper presented at the Seventh conference on International Language Resources and Evaluation.

Hu, M. & Liu, B. (2004). *Mining and summarizing customer reviews*, Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 168-177.

Kim, S. M. & Hovy, E. (2004). *Determining the sentiment of opinions*, Proceedings of the International Conference on Computational Linguistics, 1367-1373.

Dragut, E. C., Yu, C., Sistla, P. & Meng, W. (2010). *Construction of a sentimental word dictionary*, In Proceedings of ACM International Conference on Information and Knowledge Management, 1761-1764.

Mohammad, S., Dunne, C. & Dorr, B. (2009). *Generating highcoverage semantic orientation lexicons from overtly marked words and a thesaurus*. in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 599-608.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A. & Reynar, J. (2008). *Building a sentiment summarizer for local service reviews*. in Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era.

Rao, D. & Ravichandran, D. (2009). *Semi-supervised polarity lexicon induction*. in Proceedings of the 12th Conference of the European Chapter of the ACL, 675-682.

Hassan, A., Qazvinian, V. & Radev, D. (2010). *What's with the attitude?: identifying sentences with attitude in online discussions. in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1245-1255.

Hatzivassiloglou, V. & McKeown, K. R. (1997). *Predicting the semantic orientation of adjectives*, Proceedings of the Joint ACL/EACLConference, 174-181.

Qiu, G., Liu, B., Bu, J. & Chen, C. (2009). *Expanding Domain Sentiment Lexicon through Double Propagation*, International Joint Conference on Artificial Intelligence, 1199-1204.

Tai, Y. J. & Kao, H. Y. (2013). *Automatic Domain-Specific Sentiment Lexicon Generation with Label Propagation*. In Proceedings of International Conference on Information Integration and Web-based Applications & Services, 53-62.

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

Zhu, X. & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation*, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CALD-02-107.

Esuli, A. & Sebastiani, F. (2005). *Determining the semantic orientation of terms through gloss analysis*, Proceedings of the ACM Conference on Information and Knowledge Management, 617-624.

Kamps, J., Marx, M., Mokken, R. J. & Rijke, M. D. (2004). *Using WordNet to measure semantic orientation of adjectives*. In Proceeding of 4th International Conference on Language Resources and Evaluation, 1115-1118.

Andreevskaia, A. & Bergler, S. (2006). *Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses, Proceedings of the European Chapter of the Association for Computational Linguistics*, 209-216.

원고접수: 2014.08.25
수정접수: 2015.02.02
게재확정: 2015.02.10