KIPS Tr. Software and Data Eng.
Vol.5, No.8 pp.377~384 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.8.377

Developing a Sentiment Analysing and Tagging System

Lee Hyun Gyu† · Lee Songwook††

ABSTRACT

Our goal is to build the system which collects tweets from Twitter, analyzes the sentiment of each tweet, and helps users build a sentiment tagged corpus semi-automatically. After collecting tweets with the Twitter API, we analyzes the sentiments of them with a sentiment dictionary. With the proposed system, users can verify the results of the system and can insert new sentimental words or dependency relations where sentiment information exist. Sentiment information is tagged with the JSON structure which is useful for building or accessing the corpus. With a test set, the system shows about 76% on the accuracy in analysing the sentiments of sentences as positive, neutral, or negative.

Keywords : Sentiment Analysis, Twitter, Sentiment Tagged Corpus

감성 분석 및 감성 정보 부착 시스템 구현

이 현 규† · 이 성 욱††

요 약

본 연구의 목적은 트위터에서 수집된 트윗들의 감성을 분석하고 각 문장의 감성 정보를 반자동으로 부착하여 감성 말뭉치를 구축할 수 있는 시스템의 구현이다. 트위터 API를 이용해 트윗을 수집한 후 각 트윗이 어떤 감성을 갖는지 감성사전을 이용해 분석한다. 사용자는 감성 분석 결과를 확인하고 누락된 감성 정보를 추가하거나 의존구조 사이에 존재하는 감성 정보를 추가할 수 있다. 감성 정보는 JSON 구조로 부착함으로써 감성 말뭉치 구축 및 활용에 용이하게 하였다. 제안 시스템은 긍정, 부정, 중립 문장에 대한 감성 분석 결과 약 76%의 성능을 보였다.

키워드 : 감성 분석, 트위터, 감성 정보 부착 말뭉치

1. 서 론

대표적인 사회관계망 서비스인 트위터에서는 다양한 도메인에 대해 사용자의 의견과 감성이 표출된다. 사용자의 의견을 트위터와 같은 온라인 데이터로부터 추출하여 마케팅이나 사회적 문제 등의 해결이나 예측에 활용하려는 많은 시도가 있었다. 트윗 데이터를 분석해 미래 선거 결과를 예측하거나[1], 주식 시장과 영화 박스오피스 수익을 예측할 수 있다[2, 3]. 이외에도 신문 기사로부터 최근 동향을 추출하여 그 주제에 대한 트윗들의 감성을 판단하여 대중들의 여론을 분석하기도 하였다[4].

감성 분석 문제는 사용자들이 상품이나 영화와 같은 특정 주제에 대해 긍정과 부정 중 어떤 의견을 가지고 있는지 파악하는 문제이며 이러한 감성 분석 문제는 상품평, 뉴스, 블로그, 트위터 등의 도메인에서 주로 연구되어 왔다. 감성 분석을 위해 대량의 원시 말뭉치를 수집한 후 휴리스틱, 규칙 사전, 기계학습 기법 등을 이용하는 방법이 일반적인 분석 방법이다. 영어권에서는 영화 상품평 도메인에서 추출된 스탠포드 감성 트리 말뭉치[5]가 존재한다. 이와 같은 말뭉치를 이용하면 감성 지식 자원의 구축에 드는 시간을 단축함으로써 신속한 시스템 개발이 가능할 뿐만 아니라 여러 감성 분석 기법의 성능을 객관적으로 비교할 수 있다.

본 연구의 목적은 한국어 감성 부착 말뭉치를 구축하기 위한 도구 시스템을 구현하는 것이다. 먼저 트위터로부터 원시 말뭉치를 수집한 후, 트윗들의 감성을 분석한다. 사용자는 분석된 결과를 확인 및 수정하여 올바른 감성 정보를 부착할 수 있다. 현재 시스템은 크게 수집, 분석, 태깅, 저장 등의 단계로 구성되어 있으며, 감성 분석이 끝나면 문장 단위로 감성 분석 결과를 출력하며, 전체 문장의 감성비율을 원형그래프로 출력한다. 문장의 감성 정보는 데이터 활용이

※ 이 논문은 2014년도 한국교통대학교 교내학술연구비의 지원을 받아 수행한 연구임.
† 비 회 원 : 한국교통대학교 컴퓨터정보공학과 학사과정
†† 정 회 원 : 한국교통대학교 컴퓨터정보공학과 부교수
Manuscript Received : March 21, 2016
First Revision : April 25, 2016
Accepted : April 28, 2016
* Corresponding Author : Lee Songwook(leesw@ut.ac.kr)

편리한 JSON 구조로 저장한다.

본 논문의 구성은 다음과 같다. 2장에서 감성 분석에 관련된 연구들을 소개하고, 3장에서 본 시스템의 구조와 처리 방법에 대해 설명하고 4장에서 사용자 인터페이스를 설명한 후, 5장에서 결론을 맺는다.

2. 관련 연구

감성 분석 방법에 관한 연구는 크게 규칙에 기반한 방법과 통계에 기반한 방법으로 나눌 수 있으며 감성 정보 부착 말뭉치의 사용 유무에 따라 지도학습과 비지도학습으로 나눌 수도 있다[6].

규칙에 기반한 방법은 감성 규칙을 구축하는 방법에 따라 다양한 연구가 수행되어 왔다. [7]은 상품평에서 한국어의 본용언의 의미에 영향을 주는 보조용언을 감성 정보의 강도에 따라 분류한 후, 감성 구문에 보조용언이 나타나는 패턴에 따라 그 구문의 극성을 수정할 수 있도록 연관규칙을 정의하였다. [8]은 온라인 상품평에 존재하는 장점과 단점 단락을 추출한 후, 상품의 속성을 나타내는 자질을 추출하고, 이 상품 속성 자질을 포함하는 트라이그램을 이용함으로써 규칙을 생성하고 감성을 분류하였다. [9]는 상품 속성 자질을 명시적 자질로 사용하였고 의존관계와 목적어-술어 관계를 이용하여 추출한 규칙을 암시적 자질로 사용하여 감성이 포함된 구문을 추출하였으며 완화 레이블링(relaxation labeling) 기법을 이용하여 그 구문의 감성을 결정하였다. [10]은 소비자 신뢰도나 정치적 의견에 대한 설문조사 결과와 트위터에 표출된 감성어휘의 빈도 사이에 존재하는 연관성에 대해 연구하였으며 설문조사 결과와 트위터 사이에는 특정 사안에 대해 80% 이상의 연관성이 있음을 알아냈다. [3]은 영화 개봉일 전후의 트위터를 수집한 후 트위터에 존재하는 영화 관련단어와 감성 정보를 이용하여 박스오피스 수익을 예측하였다.

영어의 경우 WordNet에 감성 정보를 부착한 SentiWordNet[11]이 구축되어 있으나, 한국어의 경우 공개된 감성사전이 존재하지 않으므로 감성사전을 구축하는 연구가 진행되었다. 감성 사전 구축의 문제점은 동일 어휘의 감성이 문장의 쓰임에 따라 긍정과 부정 모두에 나타나는 것이다. 이를 해결하기 위해 사전이 사용될 도메인에 맞춰 사전을 각각 구축하거나, 상품의 속성에 따라 서로 다른 감성 사전을 구축하는 연구들이 있었다[8, 9, 12].

한편 통계에 기반한 연구들은 문장에서 추출한 n그램 등의 자질을 기계학습 기법으로 학습하고 분석한 연구들이 많다. [5]는 상품평 도메인 말뭉치에 트리구조로 감성정보를 부착하고, 감성트리 말뭉치를 이용하여 문장의 구문 관계 사이에서 전이되는 감성 정보를 재귀적 신경 텐서망(Recursive Neural Tensor Network)을 이용하여 처리하였다. [13]은 자동차, 여행, 은행 등의 여러 도메인의 상품평에 대해 형용사, 부사 어휘가 포함된 구문을 추출한 후, 이 추출된 구문과 단어 'excellent' 사이의 상호정보량(Mutual Information)과 추출된 구문과 단어 'poor' 사이의 상호정보량을 비교함으로써 그 구문의 감성을 결정하였다. [14]는 OpinionFinder 도구와 구글 무드 상태 프로파일(Goole-Profile of Mood States) 도구를 이용하여 트위터에 표출된 대중들의 무드 정보를 추출하였고, 추출된 무드 정보를 매일 주식시장 마감시의 다우존스산업평균지수(Dow Jones Industrial Average)의 등락을 예측하는데 이용하였으며, 예측을 위한 기계학습도구로 자동구성 퍼지신경망(Self-organizing Fuzzy Neural Network)을 이용하였다. [15]는 트위터 문서를 주관적(subjective) 문서와 객관적(objective) 문서로 구분한 후 각각의 문서의 감성을 나이브베이시언(Naive Bayes) 분류기로 분류하였는데, 주관적 문서와 객관적 문서의 감성이 서로 다른 특성을 지니고 있음을 밝혀냈다. [16]은 트위터에 사용된 이모티콘 정보를 이용하여 트위터 문서의 감성 정보를 자동으로 부착함으로써 지도학습에 필요한 데이터를 자동으로 생성한 후, 여러 가지 기계학습방법을 이용하여 감성을 분류하였다. [17]은 스탠포드 감성 트리 말뭉치에서 다양한 언어 자질들을 추출한 후 지지벡터기계를 이용하여 감성을 분류하였다.

3. 시스템 구성

본 시스템의 구조도는 Fig. 1과 같다. 먼저 트윗 데이터를 수집하고, 수집된 데이터는 전처리 과정과 형태소분석 과정을 거친다. 그 후 감성사전을 이용하여 감성 정보를 부착한다. 사용자가 사용자 인터페이스를 통해 감성 분석 결과를 확인하고, 수정하면 최종적으로 감성 정보 부착 말뭉치가 구축된다.

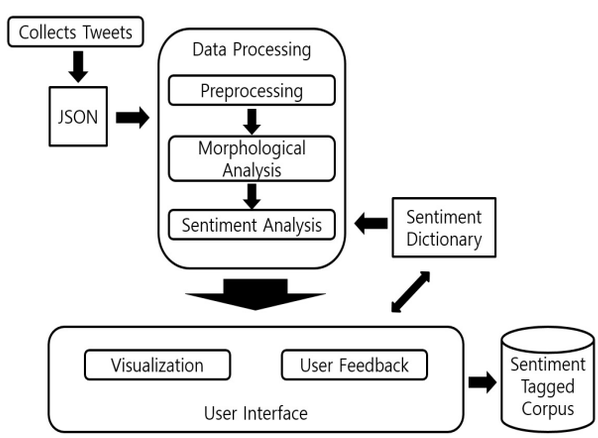

Fig. 1. The System architecture

3.1 트윗 수집

트위터에서 제공하는 API를 이용해 트윗들을 수집한다. 트윗 API는 질의요청을 받으면 한 트윗의 본문을 포함해 작성날짜, 작성자 등의 정보가 포함된 JSON 객체로 응답한다. 주로 삼성, 애플, LG 등이 제조한 스마트기기들에 대한 트윗을 수집하였으며 수집된 JSON 객체들을 모아 원시 말뭉치를 구성하였다.

3.2 전처리 및 형태소 분석

트윗 JSON 객체에서 본문은 “text” 키의 값에 저장되어 있다. 전처리 과정에서 본문을 추출하고 분석에 불필요한 요소들을 처리한 후 정제된 텍스트만 추출한다. 트윗에서 다른 사용자의 글을 전파하는 기능이 있는데 이를 리트윗이라 하고 본문에 RT로 표시된다. 이것은 중복된 트윗이기 때문에 분석대상에서 제외한다. 해쉬태그는 사용자가 입력한 주제어를 나타낼 때 사용하는 것으로 ‘#키워드’ 형식으로 표현되는데 감성 분석에 영향을 끼칠 수 있으므로 제거한다. URL주소와 ‘@사용자ID’ 형식으로 표현되는 사용자 id도 제거한다. 여러 개의 문장으로 이뤄진 트윗도 있으므로 정규식을 이용하여 문장분리를 한다.

문장의 전처리가 끝난 후, CNUMA[18]를 사용하여 형태소 분석을 한다.

3.3 감성 분석

형태소 분석이 끝난 파일의 형태소들이 문장 감성 분석의 기준이 되며 [4]의 감성사전을 이용하여 분석한다. 감성사전은 Fig. 2와 같이 약 3,800여개의 어휘 및 구문들과 극성(polarity)값의 강도로 구성되어 있다. 감성 정보는 극성값으로 나타내는데, 극성값은 -2 ~ +2의 값을 가지며 각각 매우 부정, 부정, 중립, 긍정, 매우긍정의 5단계를 나타낸다[5].

	단어	평가
2	가련하다	0
3	가망 없다	-2
4	가멸다	1
5	가문의 영광	2
6	가소롭다	-2
7	가엾다	-1
8	가장 빼어나다	2
9	가장 좋아하다	2
10	가장 친하다	2
11	가장 큰 선물	2
12	가지고 싶다	0
13	가치 가지다	1
14	가치 높다	1
15	가치 높아지다	1
16	가치 드러나다	1
17	가치 발휘하다	1

Fig. 2. Examples of the sentiment dictionary

문장의 모든 어절들은 감성사전을 이용하여 극성값을 결정한다. 먼저 문장의 모든 실질형태소가 사전의 실질형태소와 일치하는지 판별하여 그 극성값을 부여한다. 사전에 없는 어휘는 중립으로 간주한다. 이 과정을 거치면 각 문장의 모든 어휘들의 극성값을 얻을 수 있다.

문장의 감성은 각 어절이 가진 극성값의 총합으로 결정하는데 강조의 의미를 갖는 수식어와 부정문이 존재할 때에는 다음과 같은 휴리스틱을 적용한다.

‘가장’, ‘완전’, ‘제일’, ‘아주’, ‘정말’ 등과 같이 강조의 의미를 갖는 수식어는 먼저 오른쪽에서 가장 가까운 어절 중 + 또는 - 극성값을 가진 어절을 찾는다. 그리고 그 어절이 가진 극성값에 2의 가중치를 곱해 준다. 짧은 문장이 대다수인 트윗의 특성을 고려한 휴리스틱이다. 강조 어구는 가중치를 곱함으로써 긍정이나 부정 감성에 영향을 준다. 다음은 강조 어구의 처리의 예이며 괄호 안은 감성 극성값을 나타낸다.

가장(0, 강조어구) 품질이(0) **좋아(1)**

=> 가장(0) 품질이(0) **좋아(1*2)**

강조 어구 ‘가장’의 영향으로 어절 “좋아” 의 극성값이 1에서 2가 되어 강한 긍정을 의미하게 된다. 만약 음수값을 갖는 어절이 강조 어구의 수식을 받는다면 더 강한 부정을 나타내게 될 것이다.

부정문은 ‘안’, ‘못’, ‘않다’, ‘아니다’ 등의 부정을 나타내는 어절이 포함된 문장을 의미한다. 이런 어절을 포함하면 0이 아닌 극성값을 가진 가장 가까운 어절을 찾은 후, 극성값에 -1을 곱하여 그 값을 반전시킨다[15]. 다음은 부정문의 처리 예를 나타낸다.

안(0, 부정문) 좋아(1)

=> 안(0) **좋아(1*-1)**

원래는 긍정 감성을 나타내는 “좋아”가 부정문의 존재에 따라 값이 반전되어 부정의 감성을 나타내게 된다.

이렇게 강조와 부정문에 대한 휴리스틱을 적용한 후, 문장의 모든 어절의 극성값의 총합을 문장의 극성값으로 결정하고, 이 극성값이 양수이면 긍정, 음수이면 부정, 0이면 중립으로 분석한다. 200개 트윗(210개 문장 중 긍정 65개, 부정 29개, 중립 116개)으로 구성된 실험데이터로 성능을 평가해본 결과 어절 기준 약 83%, 문장기준 약 76%의 정확도를 보였으며, 이 때 긍정은 약 76%, 부정은 약 86%, 중립은 약 75%의 정확도를 보였다. 다음 Table 1은 다른 시스템의 감성 분석 성능을 비교한 표이다.

Table 1. The accuracy of other systems

	proposed	[4]	[12]	[5]
Accuracy(%)	76	74.5	81.8	85.4

Table 1의 결과는 실험 데이터가 모두 달라 객관적 비교는 어렵다. 다만 본 시스템의 성능은 한국어에 대해 기계학습 방법을 적용한 [4]의 성능 74.5%에 거의 필적한 결과라 할 수 있다. [12]와 [5]의 경우 중립 극성을 분석에 고려하지 않은 실험 결과라 대체로 제안 시스템보다 높은 성능을 보였는데, [12]는 ‘길이’ ‘디자인’ ‘재질’ ‘배송’ 등의 4가지 상품 속성에 대해서만 수행한 감성 분석 결과이며, [5]는 영어권에서 최고 성능을 보인 시스템이다.

다음 문장들은 본 시스템의 감성 분석 오류의 예이다.

1) 아이폰은 갤럭시보다 **작아** 한손에 들어온다. (-1: 부정)
2) 소비자를 **농락**하는(-1) 기술이 **뛰어나다**(2). (1: 긍정)
3) 이벤트 매장음악 **서비스** 브랜드라디오 알리고, 아이패드 **받자!** (2:매우 긍정)
4) 아이폰 사진 편집앱 아주 쩌는거 두 개 발견함 (0: 중립)
5) 그거하나 돈받는다고 삼성**불매**냐? (0: 중립)

본 시스템과 같이 사전에 기반한 감성 분석에서 발생하는 대표적인 오류는 문장 1)과 같이 평가 항목에 따라 동일 평가 어휘의 극성이 달라지는 문제라 할 수 있다. 문장 1)의 '작다'는 사전에 부정 어휘로 등록되어 있으나 그 평가 대상이나 문장의 쓰임(비교)에 따라 긍정의 의미를 가지는 것을 말한다.

또한 사전 기반 방법은 단순히 단어주머니(Bag of Words)의 극성만 이용하므로 문장 구조에 내포된 감성 정보를 반영하기 어려운 문제가 있다. 문장 2)와 같이 '기술이 뛰어나다'는 긍정의 속성이나 기술을 수식하는 관형어가 부정의 극성을 띈 경우에 문장 전체는 부정의 감성을 띄게 된다. 하지만 문장 구조를 고려하지 않는 사전 기반 방법으로는 이 문제를 해결할 수 없다. 문장 3)의 경우도 올바른 구문 분석없이 사전의 '서비스 받다(매우 긍정)'를 탐색하여 발생하는 오류이다.

문장 4)의 경우 '쩔다'가 감성 사전에 등록되지 않아 발생한 오류인데 자료 부족 문제에 기인한 오류이다.

문장 5)의 경우 '삼성불매'라는 복합명사를 형태소 분석기가 올바로 분석하지 못해 발생한 오류이다.

향후 문장 1)과 2)와 같은 문장 구조의 오류를 해결하기 위해서는 영어권의 연구[5]와 같이 구문분석을 수행하고 각 트리의 단말 노드에서부터 감성을 분석하여 점진적으로 문장 전체의 감성을 결정하는 방법을 한국어에도 적용하기 위한 연구가 필요하다.

3.4 사용자 확인 및 수정

감성 분석이 끝나면 사용자는 사용자 인터페이스를 통해 분석 결과를 확인하고 수정할 수 있다. 사용자 인터페이스는 4장에서 설명한다. 사용자는 시스템이 분석한 결과를 확인하고 그 극성값을 수정함으로써 어휘의 극성값 오류를 수정할 수 있다. 그 외 감성사전에 누락된 어휘는 사용자가 직접 사전에 극성값을 추가할 수 있다.

또한 사용자는 의존관계에 대해 감성 정보를 추가할 수 있다. 상품의 서로 다른 속성에 대해 동일한 어휘가 문장에 따라 긍정과 부정 모두에 사용되는 경우가 있다[8, 9, 12]. 이런 경우, 우리는 의존관계에 극성값을 추가하여 서로 다른 쓰임을 구분하고자 한다. 예를 들어 '빠르다'는 주로 긍정의 극성값을 가지는데 "방전이 빠르다"와 같이 사용되면 부정의 감성을 표현하게 된다. 이 때, 사용자가 의존관계 '방전이-> 빠르다'를 묶어 부정의 극성값을 입력할 수 있게 한다.

3.5 감성 정보 부착 말뭉치 구축

감성 정보 부착 말뭉치의 저장 형식은 분석과 사용이 편리한 JSON 구조로 저장한다. 감성 정보 부착 말뭉치 구축 형식에 대한 표준화는 아직 공개적으로 논의가 되지 않은 상태이나 향후 표준 형식이 제정된다면 우리가 제안하는 JSON 구조의 특성상 간단히 다른 형식으로 변환이 가능할 것이다. JSON은 객체 구조로서 키와 값이 쌍을 이루며 값은 다른 JSON 객체를 포함할 수 있다. 한 문장이 저장되는 JSON 형식을 Table 2와 같이 정의한다.

Table 2. The JSON structure of a sentence

key	value
id	index of the sentence
sentiment	sentiment of the sentence
text	the sentence
eojeols	array of eojeol(JSON)
relations	array of dependency relation(JSON)
sentence_polarity	polarity of the sentence

Table 2에서 'id'는 문장의 색인번호이며, 'sentiment' 키는 문장의 감성 정보로 'positive', 'neutral', 'negative'의 값을 가진다. 'text'는 트윗에서 추출하여 전처리 과정을 거친 원시 문장, 즉 분석대상이 되는 문장이다. 'eojeols'는 문장의 어절 집합이다. 'relations'는 사용자가 추가한 의존관계이다. 'sentence_polarity'는 감성 분석 결과로 계산된 문장의 극성값이다. 'eojeols'는 어절 JSON 객체의 배열로 정의하였으며 원소인 어절은 Table 3과 같이 정의하였다.

Table 3. The JSON structure of an 'eojeols'

key	value
id	index of the eojeols array
lex	lexicon of eojeol
morps	array of morpheme(JSON)

'id'는 어절의 색인번호이고, 'lex'는 어절이며, 'morps'는 어절의 형태소들을 저장하고 있는 형태소 JSON 객체의 배열이다. 한 어절에 대한 형태소 분석 정보를 담고 있는 'morps'의 원소인 형태소는 Table 4와 같이 정의하였다.

Table 4. The JSON structure of a 'morps'

key	value
id	index of the morps array
morpheme	morpheme
pos	POS tag
polarity	polarity of the morpheme

'id'는 'morps' 배열의 색인번호이다. 'morpheme'은 형태소를 저장하고, 'pos'은 그 품사 태그이다. 'polarity'는 형태소의 극성값을 나타낸다. 의존관계에 대한 극성값을 저장하기 위한 'relations' 배열을 구성하고 있는 의존관계에 대한 JSON 구조는 Table 5와 같다.

Table 5. The JSON structure of a 'relations'

key	value
id	index of the relations array
words	words pair of the dependency relation [modifier, head]
location	index pair of the relation [index of eojeols, index of morphs],[index of eojeols, index of morphs],
polarity	polarity of the relation

'id'는 'relations' 배열의 색인번호을 나타낸다. 'words'는 의존관계가 있는 단어 쌍을 저장한다. 'locations'은 의존관계 쌍의 수식어와 피수식어가 'eojeols 배열'과 'morps' 배열의 몇 번째 원소인지를 각각 색인번호로 나타낸다. 'polarity'는 의존관계의 극성값을 의미한다.

다음 Fig. 3은 "웹서핑 중점이시면 아패 동감 중심이시면 갤탭 추천드려요~"이라는 문장에 감성 정보를 부착한 JSON 구조의 예이다.

4. 사용자 인터페이스

본 시스템은 Java 언어로 개발되었으며 Fig. 4는 제안 시스템의 메인화면이다.

입력데이터 선택은 ①에서 입력파일을 선택하거나 ②에서 문장을 직접 입력할 수 있다. ③은 분석 진행 상황과 오류 등의 로그를 출력하는 영역이다. ④의 버튼 3개는 분석과정에서 생성된 중간파일들을 확인할 수 있다. 각 버튼 클릭 시 해당 파일을 직접 열어 볼 수 있다. ⑤버튼은 분석에 사용된 감성사전의 내용을 볼 수 있으며 Fig. 5와 같은 사전 뷰어가 실행되며 특정 단어의 감성 정보를 간단하게 검색할 수 있다.


```
{"id" : 5,
 "sentiment" : "positive",
 "sentence_polarity" : 4,
 "text" : " 웹서핑 중점이시면 아패 동감 중심이시면 갤탭 추천드려요~ ",
 "eojeols" : [
      {
          "id" : 0,
          "lex" : "웹서핑",
          "morps" : [
              { "id" : 0, "morpheme" : "웹", "pos" : "NNG", "polarity" : 0},
              { "id" : 1, "morpheme" : "서핑","pos" : "NNG", "polarity" : 0}
          ],
      },
      {
          "id" : 1,
          "lex" : "중점이시면",
          "morps" : [
              { "id" : 0, "morpheme" : "중점", "pos" : "NNG", "polarity" : 0},
              { "id" : 1, "morpheme" : "이시면", "pos" : "P", "polarity" : 0}
          ],
      },
      {
          "id" : 2,
          "lex" : "아패",
          "morps" : [
              { "id" : 0,"morpheme" : "아패", "pos" : "UNK", "polarity" : 0}
          ],
      },
     ...중략...

      {
          "id" : 6,
          "lex" : "추천드려요",
          "morps" : [
              { "id" : 0, "morpheme" : "추천", "pos" : "NNG", "polarity" : 2},
              { "id" : 1, "morpheme" : "드리", "pos" : "VV", "polarity" : 0}
              { "id" : 2, "morpheme" : "어요", "pos" : "E", "polarity" : 0}
          ],
      },
      {
          "id" : 7,
          "lex" : "~",
          "morps" : [
              { "id" : 0, "morpheme" : "~", "pos" : "SW", "polarity" : 0}
          ],
      }
 ],
 "relations" : [
      {
          "id" : 0,
          "words" : ["아패","추천"], "location" : [[ 2, 0], [6,0]],
          "polarity" : 1
      },
      {
          "id" : 1,
          "words" : ["갤탭","추천"], "location" : [[ 5, 0], [6,0]],
          "polarity" : 1
      }
 ]
}
```


Fig. 3. A JSON example of a sentiment information tagged sentence

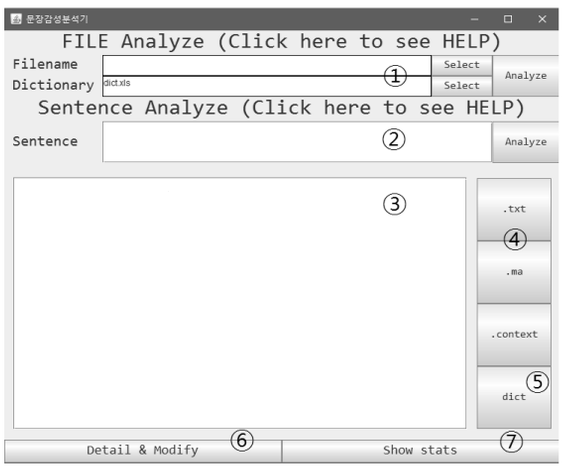

Fig. 4. Main view of the proposed system

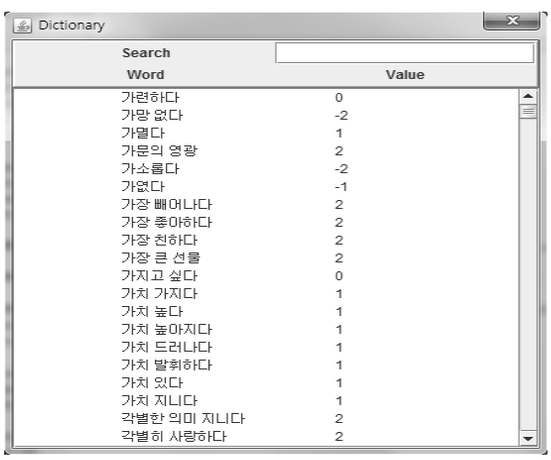

Fig. 5. A window for the sentiment dictionary

Fig. 4의 ⑥번 버튼은 Fig. 6과 같이 사용자가 문장을 확인하고 감성 사전 탐색 결과를 수정할 수 있는 창을 띄운다. 문장 감성 분석 결과는 Fig. 6의 ①번에서 긍정은 초록색, 중립은 검정색, 부정은 빨강색으로 각각 표시한다. Fig. 6의 ②번 버튼을 클릭하면 Fig. 7과 같이 현재 문장의 모든 형태소들의 극성값을 확인하고, 콤보박스의 조작을 통해 값을 수정할 수 있는 창이 실행된다.

사용자가 의존관계에 존재하는 극성을 추가하고자 할 때에는 사용자가 특정 단어를 클릭하면 된다. Fig. 7과 같은 창에서 사용자가 '아패'를 클릭하면 '아패'와 의존관계가 가능한 피수식어 리스트가 팝업메뉴로 생긴다. 이 때 올바른 피수식어인 '추천'을 클릭하면 '아패'와 '추천' 사이의 의존관계에 대한 극성값을 입력할 수 있는 창이 실행된다. 향후 다른 구문분석기와 통합한다면 의존관계에 존재하는 감성 극성값의 입력이 더욱 편리하도록 개선할 수 있을 것이다. Fig 6의 텍스트영역에는 현재 문장에서 극성값이 0이 아닌 형태소를 출력한다. Fig. 6의 ③버튼은 현재까지 수정된 분석 결과를 말뭉치에 저장한다.

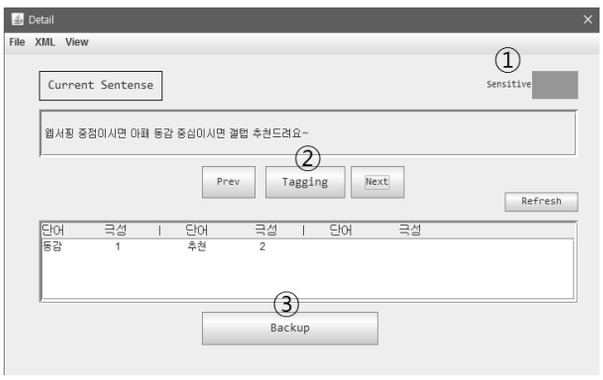

Fig. 6. A window for showing the sentiment of each sentence

Fig. 7. A window for polarity modification

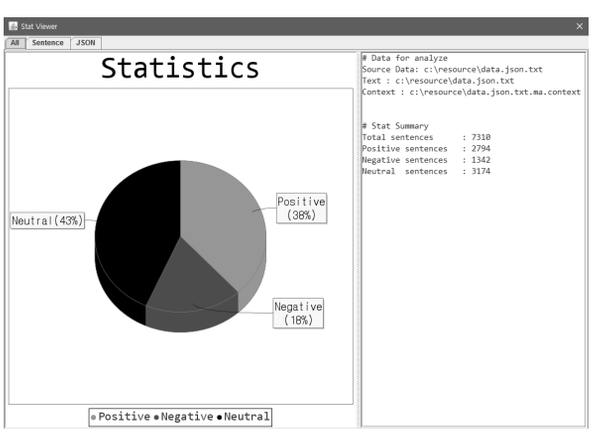

Fig. 8. A window for showing statistics of sentiment Analysis

한편, Fig. 4의 ⑦버튼은 분석 결과를 시각화하여 입력파일 전체에 나타나는 감성문장의 분포를 보여주는 Fig. 8과 같은 창을 띄운다. Fig. 8의 첫 번째 탭에서는 감성 분석 결과를 원형 그래프를 이용하여 긍정, 부정, 중립 문장의 비율을 나타내고, 우측의 텍스트영역에 실제 분석된 문장의 개

수를 보여준다. 두 번째 탭에선 모든 문장을 긍정, 부정, 중립 문장 등으로 분류하여 출력한다. 마지막 세 번째 탭은 말뭉치에 저장될 최종 JSON 객체의 내용을 Fig. 3과 같이 보여준다.

5. 결론 및 향후 과제

우리는 트위터에서 수집한 트윗에 대한 감성 정보 부착 말뭉치를 구축할 수 있는 시스템을 구현하였다. 먼저 트위터 API를 이용하여 원시 트윗 데이터를 수집하였다. 트윗 데이터는 간단한 전처리과정과 형태소분석 단계를 거친 후에 감성 사전을 이용하여 각 어절의 감성 극성값을 결정하고 문장에 존재하는 각 극성값의 총합이 양의 값이면 긍정, 음의 값이면 부정, 0의 값이면 중립으로 분석한다. 감성 분석 결과를 사용자가 수정할 수 있도록 사용자 인터페이스를 구성하였으며 사용자는 어휘 및 의존관계에 극성값을 부여할 수 있다. 감성 정보 부착 말뭉치 구축을 위한 각 문장의 감성 정보는 JSON 구조로 정의하였다. 제안 시스템은 구축 말뭉치의 감성 분포를 출력할 수 있으며 사전기반 감성 분석의 정확도는 어절 단위 약 83%, 문장 단위 약 76%이다.

향후 구문분석기와 통합한다면 의존관계에 존재하는 감성 정보를 부착하는데 더욱 편리해질 것이다. 또한 본 시스템을 웹서비스를 제공할 수 있도록 확장시킨다면 더 많은 사용자들이 감성 정보 부착 말뭉치를 구축하는데 참여하도록 하여 더 양질의 말뭉치를 보다 빠르게 구축할 수 있을 것이다.

References

[1] Gyoung-Ho Lee and Kong Joo Lee, “Design of a Reputation System for Twitter,” *Proceedings of the 24th Annual Conference on Human and Cognitive Language Technology*, pp.62-66, 2012.

[2] Johan Bollen, Huina Mao, and Xiao-Jun Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, Vol.2, No.1, pp.1-8, 2011.

[3] Sitaram Asur and Bernardo A. Huberman, “Predicting the Future With Social Media,” *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol.1, pp.492-499, 2010.

[4] Gyoung-Ho Lee and Kong Joo Lee, “Twitter Sentiment Analysis for the Recent Trend Extracted from the Newspaper Article,” *The KIPS Transactions Part B*, Vol.2B No.10, pp.731-738, 2013

[5] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts Potts, “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank,” *Proceedings of Conference on Empirical Methods on Natural Language Processing*, 2013.

[6] R. Feldman, “Techniques and applications for sentiment analysis,” *Communications of the ACM*, Vol.56, No.4, pp.82-89, 2013.

[7] Kong Joo Lee, “Compositional rules of Korean auxiliary predicates for sentiment analysis,” *Journal of the Korean Society of Marine Engineering*, Vol.37, No.3, pp.291-299, 2013.

[8] Bing Liu, Minqing Hu, and Junsheng Cheng, “Opinion Observer: Analyzing and Comparing Opinions on the Web,” *Proceedings of the 14th International World Wide Web Conference*, pp.342-451, 2005.

[9] Ana-Maria Popescu and Oren Etzioni, “Extracting Product Features and Opinions from Reviews,” *Proceedings of Conference on Empirical Methods on Natural Language Processing*, pp.339-346, 2005.

[10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith, “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series,” *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp.122-129, 2010.

[11] Andrea Esuli and Fabrizio Sebastiani, “SENTIWORDNET: A publicly available lexical resource for opinion mining,” in *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pp.417-422, 2006.

[12] Woo Chul Lee, Hyun Ah Lee, and Kong Joo Lee, “Product Evaluation Summarization Through Linguistic Analysis of Product Reviews,” *The KIPS Transactions Part B,* Vol.17B No.1, pp.93-98, 2010.

[13] P. D. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pp.417-424, 2002.

[14] Johan Bollen, Huina Mao, and Xiao-Jun Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, Vol.2, No.1, pp.1-8, 2011.

[15] Alexander Pak and Patrick Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pp.1320-1326, 2010.

[16] Alec Go, Richa Bhayani, and Lei Huang, “Twitter Sentiment Classification using Distant Supervision,” Technical Report CS224N, Stanford University, 2009.

[17] Songwook Lee, “Sentiment Analysis System Using Stanford Sentiment Treebank,” *Journal of the Korean Society of Marine Engineering*, Vol.39, No.3, pp.274-279, 2015.

[18] Kong Joo Lee and Songwook Lee, “Error-driven Noun-Connection Rule Extraction for Morphological Analysis,” *Journal of the Korean Society of Marine Engineering*, Vol.36, No.8, pp.1123-1128, 2012.

이 현 규

e-mail : gusrb0808@naver.com
2010년～현 재 한국교통대학교
컴퓨터정보공학과 학사과정
관심분야 : 자연언어처리, 감성정보처리

이 성 욱

e-mail : leesw@ut.ac.kr
1996년 서강대학교 전자계산학과(학사)
1998년 서강대학교 컴퓨터학과(석사)
2003년 서강대학교 컴퓨터학과(박사)
2003년～2004년 서강대학교 산업기술연구소
연구원
2003년～2005년 서강대학교 정보통신대학원 대우교수
2004년～2005년 LG전자 기술원 선임연구원
2005년～2007년 동서대학교 컴퓨터공학과 전임강사
2007년～현 재 한국교통대학교 컴퓨터정보공학과 부교수
관심분야 : 자연언어처리, 대화인터페이스, 기계번역, 인공지능