언어과학연구 61 (2012)

감성 분석 연구의 현황과 말뭉치에 기반한 사례 분석
- 영화평 자료를 중심으로 -

조은경(연세대학교)

Jo, Eunkyoung. 2012. The Current State of Affairs of the Sentiment Analysis and Case Study Based on Corpus. *The Journal of Linguistic Science* 61, 259-282. There are many researches on the sentiment analysis as the application of the semantic analysis in the area of computational linguistics and the application of linguistic informatics. We have reviewed the international researches and national researches on the approaches. And we have performed a case study based on the movie reputation corpus. We have observed the lexicon which takes a key role in the sentiment analysis by showing how the sentiment analysis is successfully performed in a corpus based on the sentimental lexicon extracted from the other experimental corpus. (Yonsei University)

Key words 말뭉치(corpus), 감성 분석(sentiment analysis), 어휘 자원(lexical resources), 전산 의미론(computational semantics), 전산 어휘론(computational lexicology)

1. 머리말

오늘날 누리꾼들이 블로그, 카페, 인터넷 쇼핑, 소셜 네트워크 등의 인터넷 서비스를 통해 생산한 전자 텍스트는 인터넷을 사용하는 여러 소비자 계층에 의해

두루 읽히고 공감되는 자원이다. 이러한 전자 텍스트는 하루에도 셀 수 없는 정도의 많은 양이 만들어지고 있고, 검색 포털같은 콘텐츠 제공 서비스를 통해 그 내용이 제시거나 검색되는 일이 많다. 이러한 정보 제공 서비스에는 제시된 어떤 문서에 대한 텍스트 수준의 의미를 더 잘 이해하게 하는 자동화된 정보 분석이 더해지고 있는데, 어떤 주제에 대한 긍정적이거나 부정적인 의견 표출에 대한 요약된 정보를 제시해주거나 어떤 주제에 대한 보다 상세한 항목에 대한 평가를 요약해서 제시해 주는 등이 그것이다. 이러한 의미 분석의 응용은 '감성 분석(sentiment analysis)', '의견 분석(opinion mining)', '감정 분석(emotion analysis)' 등으로 불린다.[1)]

감성 분석에 관한 연구는 전산언어학협회(Association of computational linguistics)에서 매년 주최하는 학술대회에서 최근 활발하게 이뤄지고 있으며, 기계컴퓨팅협회(Association for computing machinery)에서 주관하는 여러 학술 대회 중 전자 텍스트를 대상으로 하는 것들에서 활발한 연구가 이뤄지고 있다. 감성 분석의 결과는 어떤 대상에 대한 긍정적인 내용이나 부정적인 내용을 요약해주는 용도로 쓰일 수 있고, 긍정적이거나 부정적인 느낌을 갖게하는 대상을 요약해주는 용도로 활용된다. 언어 표현이 갖는 의미를 자동으로 처리 가능하도록 하는 문제는 언어학적 사고 체계와 연구의 성과를 정보 기술 분야에서 활용할 수 있게 하기 위해 전산 의미론(computational semantics)이라는 세부 분야 이름으로 오랜 동안 진행되고 있는 숙원 과제이다.

본 논문에서는 '감성 분석'이라는 주제로 이뤄지고 있는 연구와 응용의 현황을 살핌으로써 한국어 의미론적인 연구가 이루어질 수 있는 상황을 살피고, 말뭉치에 기반한 감성 분석으로서의 사례 분석을 수행하고자 한다. 이 사례 분석 과정에서 텍스트 자동 범주 분류에 많이 쓰이는 방법을 활용함으로써, 말뭉치 구성에서 있어 긍정과 부정의 부류를 잘 구성할 수 있는 방식을 보일 것이다. 이어 최

---

1) 여러 참고 문헌을 통해 볼 때, 영어 논문에서 opinion mining이라는 용어가 종종 쓰이지만 한국어 용어 '의견 분석'이라는 용어는 잘 쓰이지 않는다. opinion mining(의견 분석)이라는 용어에 비해 sentiment analysis(감성 분석)이라는 용어가 어휘 의미에 대한 보다 언어학적이며 섬세한 처리를 할 때에 쓰이는 경향이 있는 듯하다. 전산 언어학 분야에서 나오는 논문들은 간혹 'emotion'이라는 용어를 쓸 때도 있지만 대개 'sentiment(al) analysis'라는 용어를 쓴다. 영어 'sentiment'는 '감정'보다는 '감성'이라는 한국어와 더 잘 어울리고 'emotion'은 감정이라는 단어와 더 잘 어울리는 것 같다. 이러한 의미에서 본고에서는 '감성 분석'이라는 용어를 사용하고자 한다.

적의 긍정과 부정의 두 부류로 나눠진 말뭉치에서 긍정과 부정적인 어휘를 추출하는 방식을 보일 것이다. 그리고, 말뭉치 구성과 어휘 추출 과정에 사용된 말뭉치와는 별도의 평가 말뭉치를 통하여 실험 평가함으로써 어휘 자원이 감성 분석에 있어 매우 기본적이며 중요한 역할을 함을 보일 것이다.

사례 분석에서 실험 연구의 대상으로 삼은 자료는 네티즌들의 영화에 대한 평가글로서 검색 포털인 daum, naver에서 수집한 것이다. 분석과 어휘 자원을 추출하기 위한 과정에 사용한 말뭉치는 daum의 자료로서 643,856건을 수집하였고, 어휘 자원의 유용성을 평가하기 위한 과정에 사용한 말뭉치는 naver의 자료로서 20,000건을 수집하였다.

최경봉(2010:449)은 '코퍼스를 이용한 계량적 분석은 점유 비율의 변화에 따른 문체의 변화와 같은 양적 변화를 설명하는 데 적합한 것이며, (중략).....'라고 하였다. 본고는 감성 분석이라는 의미 연구를 하되, 어떤 언어 현상에 대한 설명보다는 말뭉치 언어학의 계량적 분석 과정을 활용함으로써 의미 분석의 응용을 위한 어휘 의미론적 자원을 만들어가는 과정을 보이고자 한다. 이에 감성 분석에 있어 말뭉치에 기반한 연구를 함으로써 어휘 의미 자원을 만들어가는 방법론적인 처리 과정과 이를 통해 구축한 어휘 의미 자원이 언어 분석과 응용에 있어 얼마나 중요한 역할을 하는지를 보일 것이다. 임지룡(2006:112)에서의 '의미의 응용에서 언어 공학의 경우, 의미가 어떻게 관여하고 응용되는가를 탐구'하는 한 사례이기도 하다.

2. 감성 분석의 연구 현황과 응용 사례

2.1 감성 분석에 관한 연구

국외에서 감성 분석에 대한 연구는 2000년대 초반부터, 국내에서 혹은 한국어를 대상으로 한 감성 분석에 대한 연구는 2008년을 즈음으로 연구 논문들이 쏟아져 나오고 있다. 감성 분석에 관련한 연구는 크게, 어휘적 특성에 중점을 둔 연구, 구문적 특성에 중점을 둔 연구, 감성 혹은 감정의 의미적 분류 체계 그 자체에 중점을 둔 연구 등이 있다.

어휘적 특성과 그 분포에 중점을 둔 감성 분석에 관한 연구들이 가장 많다.

Turney(2002)는 자동차, 영화, 은행, 여행 등의 특정 영역의 리뷰 텍스트를 감성 분석함에 있어서 'excellent', 'poor' 등의 명백한 감성적 가치가 있는 작은 수의 어휘와 함께 쓰인 단어 간의 상호정보량을 활용하여 텍스트의 감성적 경향성(sentiment orientation)을 찾아내었다. Beineke et al(2004)은 리뷰 텍스트의 감성 분류를 위해 중요한 요인이 되는 단어들을 자동으로 추출하기 위한 나이브 베이즈 모델(Naive Bayes Model)과 여러 가지 통계적 분포 모델을 사용하였다. 명재석 외(2008)는 온라인 쇼핑몰에서의 상품평을 실험용 데이터로 하고, 각 상품의 특징을 표현하는 어휘와 각 어휘들의 극성(Polarity) 정보로서 반자동으로 구축된 의미 사전을 사용했다. 문맥에 따른 다른 의미를 갖는 어휘들도 사전에 정의하여 활용하였다. 고민수 외(2010)는 네이버 영화에서 확보한 20,000개의 리뷰(영화에 대한 평점과 감상문으로 구성됨)를 실험용 데이터로 하고, 반자동으로 구축된 감정 어휘 의미 사전을 이용한 연구이다. 이 연구는 글 단위의 감성 분류를 함에 있어 SVM이라는 기계 학습 방식을 사용하였다. 그러나, 학습 데이터가 20,000건으로 다소 소규모인데다가 감성 분류를 위한 범주의 목표값이 0~10점이라는 평점 그 자체이다. 네티즌이 어떤 영화에 대해 좋다는 느낌을 표현하면서 9점을 줄지 10점을 줄지는 수의적일 수 있는 문제가 있다. 안애림(2011)은 감성 분석에 있어 매우 중요한 어휘 자원을 구축하는 방식에 대한 연구를 하였는데, DECO라는 어휘사전에 기반하여 어휘 자체의 긍정적, 부정적 극성 값과 문맥 의존적인 긍정적, 부정적 극성 값을 의미 태그로 부착함으로써 감성 분석에 쓰이는 어휘 의미 정보의 체계적인 구축 방식을 보여 주었다.

구문적 특성에 중점을 둔 감성 분석에 관한 연구들도 있다. Wilson et al(2005)은 어떤 표현이 감성적으로 중립적인지, 어떤 감성적 극성을 띠고 있는지가 모호한 때의 중의적인 감성의 해결을 위해 구문 수준에서 문맥 극성(contextual polarity)을 찾는 연구를 하였다. Moilanen and Pulman(2007), Yejin Choi et al(2008)는 의미적 합성성 원칙의 관점에서 감성 분석은 단어들의 순서나 문맥 정보가 상호 작용함으로써 전체적인 감성 극성을 결정지을 수 있다며, 성분들 간의 문장내적(subsentential) 상호 작용을 살폈다. Hayeon Jang, Hyopil Shin(2010)은 한국어, 일본어, 터키어와 같이 다양한 형태적 변화를 하는 언어에 대한 감성 분석에서 언어적 자질이 중요함을 말하고, 특히, 구문 분석을 감성 분석의 단계로 넣음으로써 의미 관계에 영향을 주는 성분 요소들 간의 관계를 추출하여 감성 분석에 적용하였다.

감성 혹은 감정의 의미적 분류 체계 그 자체에 중점을 둔 연구는 향후 감성 분석에 응용될 수 있는 언어적 속성의 기반을 마련할 수 있다. 윤애선 외(2010)는 감성 분석을 위한 기반 언어 자원이 되는 감정 온톨로지의 모형에 대한 제안을 한 연구이다. 제안하는 감정 온톨로지는 감정 표현의 범주를 기술 대상과 방식에 따라 6개 범주로 분류하고, 이들 간 상호 대응관계를 설정함으로써, 하나의 텍스트를 단위로 한 감성 분석뿐만 아니라 상호 작용형 의사 소통 환경(멀티모달)에 적용할 수 있도록 하였다. 여기에 국제표준의 태그세트를 수용함으로써, 다국어 처리에 활용을 극대화할 수 있도록 고려했다.

위의 여러 가지 연구들은 어휘 정보의 직접적인 활용이나 어휘 수준 이상의 언어 정보의 추가적인 활용을 통해 감성 분석의 성과를 보여주고 있다. 그런데, 위의 연구들은 가장 기본이 되는 어휘 자원은 대개 가용한 기존의 사전이나 시소러스의 활용에 대한 언급이 있을 뿐 어휘 자원 자체가 감성 분석에 있어 얼마만큼의 중요성을 갖는지에 대한 분석은 없었을 뿐만 아니라 가장 기본이 되는 어휘 자원을 충분히 확보하기 위한 노력에 대해서는 구체적인 언급이 없었다. 본고는 어휘 자원의 활용이 감성 분석에 있어 얼마나 중요한 역할을 할 수 있는지를 말뭉치를 이용한 사례 분석을 통해 보일 것이며 이 과정에서 감성 분석에 활용할 수 있는 어휘 자원을 확보하기 위한 계량적인 분석 방법을 보일 것이다.

2.2 감성 분석의 응용 사례

한국어 감성 분석의 응용 사례는 검색 포털 서비스인 네이버와 다음에서 찾아볼 수 있는데[2], 예시를 보이면 [그림 1]과 같다. [그림 1]의 내용은 '7광구'라는 영화를 검색어로 입력했을 때이며, '7광구'와 관련된 문서들을 감성 분석한 결과이다. 긍정적인 부류로 결정된 것과 부정적인 부류로 결정된 것의 분포가 각각 54%, 46%로 제시되고 있다. 글 자체에 감성적인 평가 표현을 표시해주고 있지 않아서 정확히는 알 수 없다. 그러나 요약 제시된 내용을 보면, '재미', '총평', '연기력', '출연진', '영상', '평가' 등의 단어가 쓰이고, 이 단어들과 함께 '좋다', '나쁘다'와 같은 단어가 쓰였는지를 보아 긍정적이거나 부정적인 글로 분류했을 것으로 추정된다.

---

2) 네이버의 경우 다음의 유알엘에서 사례를 볼 수 있다. http://s.lab.naver.com/posneg/

[그림 1]. 블로그 글 감성 분석 사례 예시1 :
http://search.daum.net/search?w=blog&m=board&f=section

[그림 2]는 해외에서 만들어진 감성 분석의 응용 사례이다. 최근 들어, 트위터, 페이스북 등 인터넷 이용자들의 참여가 유도되는 소셜 서비스들이 속속 등장함으로 인해, 이를 통해 양산된 데이터를 감성 분석함으로써 감성적 트랜드를 파악하고, 누리꾼들끼리 공감되는 의견을 자동으로 추출하려는 요구도 확산되고 있다. tweetfeel은 트위터 메시지에서 검색어 'Korean'에 대해 긍정적인지, 부정적인지에 대한 분석 결과를 보이고 있다. 단문 형태의 메시지가 하나의 글 단위이기 때문에 감성 분석에서 어떤 언어적 요소들을 보는지가 쉽게 드러난다. 'love', 'best'와 같은 긍정적인 감성 단어의 출현이 있으면 긍정적인 것(웃는 모양 이모티콘)으로, 'fail', 'sucks'와 같은 부정적인 감성 단어의 출현이 있으면 부정적인 것(울상의 이미티콘)으로 표시하고 있는 듯하다.

[그림 2] 소셜 텍스트에서의 감성 분석 응용 예시 : http://www.tweetfeel.com/

위와 같은 예시들에서처럼 감성 분석이라는 주제로 이뤄지는 응용 사례들은 공통적으로 감성 분석에 있어서 가장 기본적이며 핵심적인 요소는 감성적인 의미를 담고 있는 어휘의 사용임을 보이고 있다고 할 수 있다. Valentin Jijkoun et al(2010)의 연구를 보더라도 감성 분석의 효용성 측면에서도 감성 분석의 양적 범위를 넓히는 데에 가장 큰 역할을 하는 것은 어휘이다.

3. 사례 분석

이 장에서는 말뭉치에 기반한 감성 분석의 사례 분석으로서, 3.1에서는 감성 분석의 처리 과정을 살피고, 3.2에서 사례 분석을 위한 말뭉치를 어떻게 구성할지, 3.3에서 어떻게 어휘 정보를 추출하고 어휘 자원으로 구축할 수 있을지를 살핀다. 그리고, 3.4에서 어휘 정보의 활용으로 감성 분석이 이뤄지는지 과정을 살피고 사례 분석 과정에 구성한 말뭉치와는 별도로 구성된 말뭉치를 가지고 실험 평가해 본다.

3.1 감성 분석의 처리 과정

감성 분석의 대상이 대개 하나의 완결된 텍스트를 단위로 이뤄진다고 할 때, 어휘 정보에 중점을 둔 감성 분석의 모형을 기술하고 본고에서의 실험 과정에 사용될, 텍스트 범주 분류에 자주 쓰이는 분류 방법론인 나이브베이즈 분류 방법에 대한 간략한 소개를 한다.

감성 분석을 언어 처리 응용의 마지막 단계인 화용 분석으로까지 활용한다고 한다면, 형태소 분석에서부터 구문 분석, 담화 분석까지 진행될 수 있으며 어휘 정보뿐만 아니라 문장 형태 정보와 담화 구조 정보까지 활용되어야 하는 분석 단계를 둘 수 있을 것이다. 그러나, 감성 분석을 어휘 분석에 주로 의존한 의미 분석으로 본다면, [그림 3]과 같은 분석 단계를 둘 수 있다.

0. 문서(텍스트)
->단계 1. 문장 단위로 자름.
->단계 2. 형태소 분석[3], 감성 어휘 추출
->단계 3. 감성 어휘 강도, 감성 어휘 결합에 의한 강도 계산
->단계 4. 감성 부류 결정

[그림 3] 감성 분석 처리 단계

나이브베이즈 분류는 베이즈 정리(Bayes theorm)에 기반해 있다. 베이즈 정리는 어떤 부류(C)에 속해 있는 단어(W)의 확률값을 알면, 어떤 단어가 어떤 부류에 속할 확률값을 추정할 수 있다는 것이다. 베이즈 정리[4]의 수식은 다음과 같다.

$$P(W|C) = P(W)*P(C|W) / P(C)$$

3) 본고에서 사용한 형태소 분석기는 모란소프트(http://www.moransoft.com/)의 것이다. 형태소 분석기는 정확도는 어절 단위 기준으로 98% 이상으로 알려져 있다.

4) 베이즈 정리는 아래의 사이트에서도 자세히 설명되어 있다. http://en.wikipedia.org/wiki/Bayes'_theorem

그런데, 하나의 글은 여러 단어들(W_1,W_2,W_3, ...,Wn)로 이뤄져 있다. 따라서, 글 단위가 어떤 부류(C)에 속할 확률값은 다음과 같이 표현할 수 있다.

$P(C|W_1, W_2, W_3, ..., Wn)$

이것은 베이즈 정리에 의해 다음과 같이 쓸 수 있다.

$P(C|W_1,W_2,W_3, ...,Wn) = P(C) * P(W_1,W_2,W_3, ...,Wn|C) / P(W_1,W_2,W_3, ...,Wn)$

여기서, 오른 항에 있는 분모에 있는 W_1,W_2,W_3, ...,Wn라는 단어 연속체의 확률 값은 모두 알아내기도 불가능하고 어떤 부류를 결정하는 데에 영향을 미치지 못한다고 보아 소거된다. 이제 '어떤 부류의 확률값 P(C)와 어떤 부류에 속한 단어 연속체의 확률값 $P(W_1,W_2,W_3, ...,Wn|C)$'을 알게 된다면, 왼 항의 단어 연속체인 글이 어떤 부류에 속할지 확률값으로 결정이 가능하다.

이 같이 자료를 가지고 목표 부류를 추정하는 방식에 있어서, 이미 어떤 부류에 속한 단어들의 확률값을 계산하는 과정을 학습과정이라고 하고, 구해 둔 확률값을 가지고 단어 연속체인 글의 부류를 결정하는 과정을 분류 과정이라고 한다. 이 분류 과정에서 이미 부류의 정답 값을 가지고 있다면 분류가 얼마나 잘 되는지를 평가할 수 있고 이를 평가 과정이라고 한다. 다시 말해, 나이브베이즈 분류는 기존에 있는 어떤 글의 '단어'들이 속한 부류의 확률값을 저장해 두었다가 가장 큰 값을 갖는 부류를 새로운 글의 부류로 결정하는 방식이다. 이를 위해서는 각 부류의 데이터를 학습용과 평가용으로 나눈 다음, 학습용 자료에서 각 부류의 글에서 나온 단어들이 어떤 부류에 속할지의 확률값을 만들고, 평가용 자료의 각 단어에 그 확률값을 적용함으로써 올바르게 자동 분류되는지를 평가하게 된다. 이러한 과정을 정리하면 [그림 4]와 같다.

단계1. 학습과 평가에 사용할 부류가 있는 말뭉치를 수집한다. 학습과 평가에 사용되는 말뭉치는 각기 독립적이어야 한다.
단계2. 학습 과정을 통해 단어가 부류에 속할 확률값을 구한다.
단계3. 분류 과정을 통해 학습한 부류 조건부 단어 확률값(P(W|C))으로 정확한 부류값(P(C|W))을 찾는지 평가한다. 평가 결과가 좋은 조건 부 단어 확률값을 어휘 확률 정보로 사용한다.

[그림 4] 나이브 베이즈 분류 과정

본 논문의 사례 분석에서는 [그림 3]과 같은 어휘 정보에 중점을 둔 감성 분석을 수행할 것이며, 이를 위해 최적의 말뭉치를 구성하기 위한 방법으로 텍스트 범주(부류) 분류에 가장 많이 쓰이는[5] [그림 4]와 같은 방법을 사용할 것이다. 나이브베이즈 분류 방법은 어떤 단어가 범주에 속할 확률 값을 만들어 두면, 범주가 없는 텍스트의 출현 단어를 가지 텍스트의 범주를 결정할 수 있다. 따라서 어떤 단어가 범주에 속할 확률값을 잘 만드는 학습 과정은 범주가 없는 텍스트 혹은 학습 과정에 쓰인 말뭉치와는 별도의 말뭉치에서도 정확한 범주 결정을 잘 하도록 만드는 과정이다. 따라서 말뭉치의 적절한 구성이 무엇보다 중요하다.

3.2 말뭉치 구성

실험 과제를 위해 누리꾼들이 작성한 영화 평가문을 말뭉치로 수집하였다.[6] 이는 <표 1>과 같이 영화 제목, 평점(평가 점수), 평가문으로 구성되어 있다. 누리꾼이 평가문을 작성하고 평점을 기입한 것이기 때문에, 평점을 기준으로 평가문이 긍정적인지 부정적인지의 두 부류로 나눌 수 있다. 이같은 평점을 가진 평

5) 나이브 베이즈 분류에 대한 보다 자세한 내용은 책 Manning Christopher D., Schütze Hinrich (1999)의 16장 'Text categorization'과 Manning Christopher D., et al(2008)의 13장 'Text classification and Naive Bayes'을 참조하면 자세한 내용을 알 수 있다. 혹은 온라인으로는 'http://nlp.stanford.edu/teaching.shtml'라는 사이트를 방문해서, Foundations of natural language processing이라는 책을 클릭하면 16장 Text categorization에 대한 학습 사이트를 볼 수 있고, Introduction to information retrieval이라는 책을 클릭하면 13장 Text classification and Naive Bayes에 대한 학습 사이트를 볼 수 있다.

6) 검색 포털 다음의 영화 섹션(http://movie.daum.net/review/)에서 네티즌 평점이 있는 텍스트 데이터를 크롤링하였다.

가문 말뭉치를 두 가지의 부류로 나눔으로써 긍정적인 평가를 할 때의 어휘 표현과 부정적인 평가를 할 때의 어휘 표현을 추출할 수 있다.

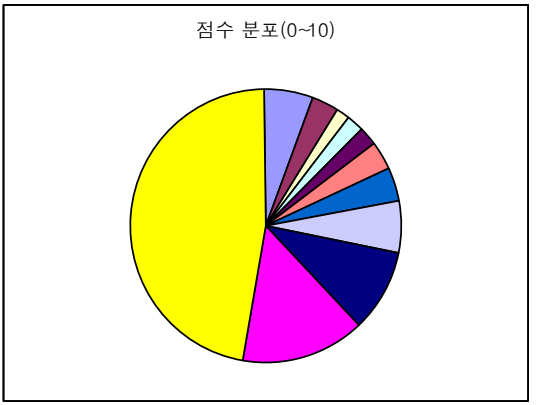
<표 1> 분석 데이터 예시

영화 제목	평점	평가문
포화속으로	4	이거친구들한테 보자해서 봤다가 욕만 얻어먹었습니다 뭔가 더 나와야 할 장면이 끊기는 기분 많이 들고요 마지막 30분 정도 전쟁신만 볼만하고 전까지는 지루합니다..탑얼굴 보러 갈 여자분만 보세요
포화속으로	3	연출력 최악.. 잼없고 졸림.. 탑 연기 때문에 3점 줌..
포화속으로	3	실화를 바탕으로 했다지만... 너무 억지스런 상황연출... 영화 자체는 좀 어이가 없네요.. 물론 6.25때 고생하신 여러분들은 참 자랑스럽지만..
포화속으로	10	후회 없었던 영화였습니다.. 영화 끝나고 사람들이 일어나서 기립박수를 치더군요. 캐스팅도 좋았고 정말 마음으로와 닿는걸 보면 영화가 주는 메시지도 잘 전달되었다고 봅니다.
쿵푸팬더	9	너무 재미있네요^^ 간만에 모든 걸 잊고 웃을 수 있었습니다!!
쿵푸팬더	9	너무너무 재밌었어요. 애니메이션이라 그리 큰 기대는 안했는데 보는 내내 웃느라 정신 없었어요~ 팬더두 너무 너무 귀엽고요~ 이런 영화가 많이 나왔음 좋겠네요^^
좋은 놈 나쁜 놈 이상한 놈	10	....송강호 좀 짱인 듯! 완전 추천!!!

평가문을 평점을 기준으로 긍정적인 부류와 부정적인 부류의 2가지 부류로 나눈다고 할 때, 2부류로 나눌 수 있는 평점의 기준을 몇 점으로 할지가 문제이다. 누리꾼의 평가는 위의 <표 1>에서처럼 어떤 영화에 대해 0점에서 10점까지의 평점과 이 점수에 상응하는 평가문으로 구성된다. 평가문별 평점 분포는 <표 2>, [그림 5]와 같다.

<표 2> 평점 분포 표

평점	문서 수
0	36,938
1	21,779
2	11,134
3	13,172
4	13,047
5	22,123
6	24,582
7	40,120
8	62,249
9	93,293
10	305,419
합	643,856



[그림 5]. 평가문 점수 분포
(가운데 위에서 시계 방향으로 0점부터 10점까지)

자료를 긍정과 부정의 두 가지 부류로 나눌 때, 양적인 기준으로만 나눈다면 점수 분포표에서 보이듯이 반 정도의 자료로 나뉘어지는 점수인 10점이나 9점을 기준으로 두 가지로 나눌 수 있겠다. 하지만, 긍정과 부정이라는 두 부류의 텍스트로 나눌 때 가장 잘 분류될 수 있는 기준 평점을 찾는다면, 긍정적인 느낌의 텍스트에 쓰이는 언어 현상과 부정적인 느낌의 텍스트에 쓰이는 언어 현상의 차이를 보다 잘 관찰할 수 있을 것이다. 따라서, 1~9점까지를 각각 기준 평점으로 설정하여 부정적인 글과 긍정적인 글로 나눈 다음 가장 잘 나누어지는 점수를 찾아내도록 하였다. 즉, 1점을 기준으로 평점이 0점인 글을 부정적인 글로 보고, 평점이 1~10점인 글을 긍정적인 글로 보아서 잘 나누어지는지 살피고, 다음은 2점을 기준으로 평점이 0~1점인 글을 부정적인 글로 보고, 평점이 2~10점인 글을 긍정적인 글로 보아서 잘 나누어지는지 살피고, 다음은 3점을 기준으로 평점이 0~2점인 글을 부정적인 글로 보고, 평점이 3~10점인 글을 긍정적인 글로 보아서 잘 나누어지는지 살피는 식으로, 이후 기준 점수에 대해서도 차례차례 수행한 다음 가장 잘 나누어지는 점수대를 찾는 방식이다. 이 때, 두 가지 부류로 나누기 위한 기준 평점을 찾기 위해 텍스트 범주 분류에 자주 쓰이는 나이브베이즈 분류 방식을 활용하였다. <표 3-1>은 점수대별 실험 결과이다. 각 평점의 문서 수에 의한 자료의 치우침을 피하기 위해 평점별 동일한 문서 수로 구성하였

다.

<표 3-1> 긍정 부정 부류 결정을 위한 실험 결과 표

평점	1점	2점	3점	4점	**5점**	6점	7점	8점	9점
문서 수	23,488	23,488	23,488	23,488	**23,488**	23,488	23,488	23,488	23,488
틀린 개수	6,706	4,886	4,431	4,106	**3,800**	3,838	4,137	4,872	6,280
맞힌 개수	16,782	18,602	19,057	19,382	**19,688**	19,650	19,351	18,616	17,208
정확도	71.44	79.20	81.13	82.51	**83.82**	83.65	82.38	79.25	73.26

<표 3-1>을 보면, 5점까지의 평점을 가진 글을 부정적인 글로 분류하고 6점 이상의 평점을 가진 글을 긍정적인 글로 분류했을 때 가장 잘 분류됨을 보여준다. 이것으로 긍정적인 글과 부정적인 글로 특징을 살피기 위한 평점 기준을 5점으로 하였으며, <표 3-2>와 같이 구성하였다.

<표 3-2> 긍정 부정 부류 말뭉치 구성 표

부류	부정 부류 문서 (0점~5점)	긍정 부류문서 (6점~10점)	합
문서 수	118,193	525,663	643,856

3.3 감성 어휘 정보 추출

이 절은 이전 절의 말뭉치 구성과 분석을 토대로 어떻게 어휘 정보를 추출하고 어휘 자원을 구축할지에 대해 살피고자 한다.

감성 분석에서 가장 기본이 되는 것은 감성을 드러내는 어휘이다. 부정적이거나 긍정적인 뜻의 단어는 텍스트의 감성을 결정하는 단서로서 가장 기본적인 역할을 한다. 말뭉치에 기반하여 어휘 정보를 추출하고 어휘 자원으로 구축하려 한다면, 어떤 단어가 긍정적인 글에 쓰인 어휘 확률과 부정적인 글에 쓰인 어휘 확률을 만드는 것부터 시작할 수 있다. 이때, <표 3-2>와 같이 부정 부류의 문서가 긍정 부류의 문서보다 훨씬 많은 즉, 부정 부류 문서와 긍정 부류 문서의 양이 치우침이 심하면 유의미한 어휘 확률값을 구할 수 없다. 어느 부류이든 자료량이 큰 쪽의 어휘 확률값이 대부분 작을 것이기 때문이다. 따라서, 긍정과 부정에 대한 유의미한 어휘 확률값을 내기 위해서는 두 부류의 자료량이 비슷하여

말뭉치가 균형성을 가질 수 있도록 <표 3-3>과 같이 구성하였다. 그리하여, 643,856 건의 문서 중에서 236,386건의 문서만 사용하였다.

<표 3-3> 긍정 부정 부류 말뭉치 구성표

부류	부정 부류 문서 (0점~5점)	긍정 부류 문서 (6점~10점)	문서 수 합(어절 수)
문서 수	118,193	118,193	236,386(2,524,770)

부정 부류와 긍정 부류의 균형된 말뭉치를 가지고, 각 단어에 대한 긍정과 부정에 대한 어휘 확률을 구하면 <표 4>와 같다. 보다 큰 확률 값을 갖는 경우에 굵은 글씨체로 표시하였는데, 긍정 부류(good)에 대해 확률 값이 크다는 것은 긍정적인 글이 될 가능성을 키우는 단어이고, 부정 부류(bad)에 대해 확률 값이 크다는 것은 부정적인 글이 될 가능성을 키우는 단어이다. 부정에 대한 확률 값이 큰 단어들 예를 들어, '2'의 '쓰레기', '5'의 '최악'이라는 단어는 직관적으로도 부정적인 느낌을 주는 단어들이다.

<표 4> 부정 부류와 긍정 부류에 대한 어휘 확률값

1	극장판	bad	0.00003
	극장판	**good**	**0.00005**
2	**쓰레기**	**bad**	**0.00196**
	쓰레기	good	0.00011
3	애니메이션	bad	0.00008
	애니메이션	**good**	**0.00025**
4	**모르다**	**bad**	**0.00211**
	모르다	good	0.00205
5	**최악**	**bad**	**0.00275**
	최악	good	0.00006
6	**없다**	**bad**	**0.00004**
	없다	good	0.00001
7	카리스마	bad	0.00004
	카리스마	**good**	**0.00018**
8	**허접하다**	**bad**	**0.00036**
	허접하다	good	0.00005
9	좋다	bad	0.00310
	좋다	**good**	**0.00765**
10	**실망**	**bad**	**0.00224**
	실망	good	0.00030

위와 같이 어떤 단어가 긍정적인 부류의 글에 쓰일 확률값과 부정적인 부류의 글에 쓰일 확률값을 가진다면, 그 단어에 대한 의미적 경향성을 다음과 같이 구해볼 수 있다.

긍정 부류에 대한 강도 = 긍정에 대한 확률값 - 부정에 대한 확률값

위와 같이 긍정에 대한 값에서 부정에 대한 값을 제한 값 즉, 긍정에 대한 값만 남게 되는 것을 긍정 부류에 대한 의미적 경향성(강도)이라고 할 수 있겠다.

역으로, 부정 부류에 대한 의미적 경향성(강도)은 다음과 같이 표현할 수 있다.

부정 부류에 대한 강도 = 부정에 대한 확률값 - 긍정에 대한 확률값
= -긍정 부류에 대한 강도

<표 5>의 왼쪽은 <표 4>에 있는 긍정(good)에 대한 확률값에서 부정(bad)에

대한 확률값을 뺀 값 즉, 긍정에 대한 강도이고, <표 5>의 오른쪽은 그 값에 따라 정렬하여 순위를 매겨 본 것이다.

<표 5> 긍정에 대한 강도 예시

긍정에 대한 강도			긍정 강도에 따른 순위		
극장판	0.00002		1	좋다	0.00455
쓰레기	-0.00178		2	애니메이션	0.00017
애니메이션	0.00017		3	카리스마	0.00014
모르다	-0.00006		4	극장판	0.00002
최악	-0.00269	=>	5	없다	-0.00003
없다	-0.00003		6	모르다	-0.00006
카리스마	0.00014		7	허접하다	-0.00031
허접하다	-0.00031		8	쓰레기	-0.00178
좋다	0.00455		9	실망	-0.00194
실망	-0.00194		10	최악	-0.00269

<표 6>은 위와 같은 과정을 통해 구축한 감성 어휘 자원이다. 긍정과 부정 각 부류에서 30개씩을 추출하였다. 부정에 대한 강도가 큰 순으로 상위 30개와 긍정에 대한 강도가 큰 순으로 상위 30개를 추출한 것이다. 굵은 글씨로 표시된 단어들은 말뭉치를 통해 각 부류에 대한 경향을 구해보지 않더라도 부정적이거나 긍정적인 느낌을 주는 단어라고 할 수 있다. 그리고, 긍정보다는 부정적인 감성을 더하는 어휘가 조금 더 두드러짐을 보이는데, 이는 박인조, 민경환(2005), 임지룡(2006)에 의한 조사와도 일치하는 결과이다.[7)]

7) 이지영(2009) p.206 각주 5)의 재인용. 박인조, 민경환(2005)에 의하면 감정을 표현하는 우리말은 흔히 쓰는 말만 430여 개쯤 되고 그 중에서 불쾌한 감정을 표현하는 낱말이 훨씬 많다. 사랑, 행복, 기쁨처럼 '쾌(快)'를 표현하는 말은 전체의 30%도 안 되고 참담·배신 등 '불쾌'를 나타내는 단어가 70%가 넘는다.
이지영(2009) p.206 각주 6)의 재인용. 임지룡(2006)에서는 신체 부위에 따른 감정의 생리적 반응 중에서 부정적 감정이 긍정적 감정에 비해 압도적으로 많이 나타난다고 한다. 곧, 217개의 생리적 반응 가운데 부정적인 감정인 '화, 두려움, 슬픔, 부끄러움, 미움, 긴장, 걱정'은 186개로 85.71%이며, 긍정적인 감정인 '기쁨, 자부심, 경탄, 감동'은 31개로 14.29%이다.

<표 6> 감성 어휘 자원 예시

부정 강도가 큰 단어		긍정 강도가 큰 단어	
없다	0.01041	**있다**	0.01094
아깝다	0.00497	**재밌다**	0.00654
돈	0.00450	**최고**	0.00562
최악	0.00268	정말	0.00486
지루하다	0.00226	**좋다**	0.00454
아니다	0.00210	보다	0.00442
뭐	0.00200	연기	0.00347
ㅡㅡ	0.00197	감동	0.00305
실망	0.00194	ㅋㅋ	0.00292
평점	0.00191	**잘**	0.00203
쓰레기	0.00185	수	0.00199
왜	0.00184	감동적	0.00191
이건	0.00171	**괜찮다**	0.00179
별로	0.00166	가슴	0.00170
그냥	0.00164	**강추**	0.00168
어이	0.00124	눈물	0.00156
알바	0.00123	**대박**	0.00155
짜증	0.00114	ㅋ	0.00145
주다	0.00114	너무	0.00144
무슨	0.00113	**재미**	0.00144
나오다	0.00109	않다	0.00143
말다	0.00101	울다	0.00142
시간	0.00100	**ㅎㅎ**	0.00134
ㅡ	0.00098	역시	0.00134
감독	0.00090	웃다	0.00129
차라리	0.00090	느끼다	0.00128
안되다	0.00088	슬프다	0.00126

지금까지 어휘가 긍정적인 부류에 속할 수 있는 확률 값과 부정적인 부류에 속할 수 있는 확률 값을 계산함으로써, 긍정적인 의미 강도와 부정적인 의미 강도를 구해 보았다. 긍정적인 의미 강도나 부정적인 의미 강도는 어휘들 간에 감성적 강도의 차이를 드러내는 역할을 한다. 이를테면, '실망'은 '쓰레기'나 '별로'라는 단어에 비해 더 강한 부정적인 강도를 지니고, '쓰레기'는 '실망'보다는 약하지만 '별로'라는 단어보다는 더 강한 부정적인 강도를 지닌다.

3.4 어휘 자원에 기반한 감성 분석 과정과 실험 평가

어떤 텍스트를 단어 출현 여부를 기준으로 감성적 부류를 결정한다고 한다면, 출현한 어휘가 부정적인 부류에 속하는지 긍정적인 부류에 속하는지, 그리고 그 각 부류의 어휘가 몇 번 정도의 횟수로 출현했는지에 따라 결정 가능하다. 1~3)과 같은 글들이 있고, 이들은 각각 괄호 속의 부류에 속할 수 있도록 감성 분류를 해야 한다고 하자.

(1) 해운대 : 완죤 재밌어요~특히 김휘박사님 카리스마 대박~ㅋ (긍정)
(2) 해운대 : 시도는 좋았으나 냉정하게 점수 주기도 아깝다.. 이건 장르가 무엇인가? . (부정)
(3) 박쥐 : 모르는사람에겐 쓰레기영화 아는사람에겐 최고였던 영화. (긍정)

이 때, <표 7-1>과 같은 단순한 수준의 긍정 부류와 부정 부류의 어휘부가 있다고 하자.

<표 7-1> 긍정/부정 부류 단순 어휘부

부정 부류	긍정 부류
쓰레기	카리스마
아깝다	최고
모르다	대박
최악	좋다
실망	재밌다
...	...

긍정/부정 부류의 단어 출현만으로 긍정과 부정을 분류할 때, (1)은 성공하겠지만, (2)와 (3)은 긍정이나 부정을 결정할 수 없어 실패할 것이다. '좋다'는 긍정 어휘이고 '아깝다'는 부정 어휘이며, '최고'는 긍정 어휘이고 '쓰레기'는 부정 어휘일 텐데, 긍정 어휘와 부정 어휘가 동일한 빈도로 출현을 했기 때문이다. 그런데 이때, 이전 장에서 구축한 감성 강도 정보가 있는 <표 7-2>와 같은 자원이 있다고 하자.

<표 7-2> 긍정/부정 강도 어휘부

부정 부류	부정 강도	긍정 부류	긍정 강도
쓰레기	0.00185	카리스마	0.00014
아깝다	0.00497	최고	0.00562
모르다	0.00006	대박	0.00155
최악	0.00268	좋다	0.00454
실망	0.00194	재밌다	0.00654
...		...	

<표 7-2>의 각 어휘가 가지고 있는 긍정 부류와 부정 부류에 대한 감성 강도를 활용한다면, (2), (3)과 같은 경우에도 성공적으로 분류할 수 있다. <표 7-2>를 보면, '좋다'의 긍정 강도에 비해 '아깝다'의 부정 강도가 더 크다. '좋다'는 0.00454이고, '아깝다'는 0.00497이다. 즉, (2)는 긍정의 강도보다 부정의 강도가 더 큰 어휘에 의해 부정 부류로 결정될 수 있다. 또, '쓰레기'의 부정 강도에 비해 '최고'의 긍정 강도가 더 크다. '쓰레기'는 0.00185이고, '최고'는 0.00562이다. 즉, (3)은 부정의 강도보다 긍정의 강도가 더 큰 어휘에 의해 긍정 부류로 결정될 수 있다. 이처럼, 말뭉치로부터 추출한 어휘와 이들 각각의 감성 강도는 감성 분류를 더 잘 할 수 있게 한다.

지금까지 우리는 감성 부류 결정을 위해 말뭉치에서 감성 어휘 정보를 추출하고 추출한 어휘의 감성적 강도를 구해보았으며, 감성 분석의 처리 과정을 통해 어휘 정보가 감성 분석과 감성 부류의 결정에 매우 중요하며 가장 기초적인 자원이 됨을 말했다. 그렇다면, 기초적인 자원인 어휘 정보에 의한 감성 분류는 어느 정도의 커버리지를 갖는지 양적인 평가를 해 볼 수 있다. 이를 위해 활용한 자료는 네이버의 영화 평가문 말뭉치[8]를 확보하였는데, 이는 이전까지의 분석과 실험에 사용한 자료와는 영화라는 주제 영역이 같을 뿐 완전히 독립적인 말뭉치임을 보장하기 위해서이다. 이 말뭉치도 평가글과 함께 1~10점이라는 평점으로 구성되어 있다. 이것도 3.1절에서 구한 긍정과 부정으로 나누기 위한 최적 기준을 찾는 방식을 써서 최적 기준인 5점을 기준으로 긍정 부류와 부정 부류의 말뭉치로 구성하였으며 그 구성 결과는 <표 8>과 같다. 평가 자료의 전체 문서 수

8) http://lab.naver.com/research/라는 사이트를 통해 20,000건의 연구용 데이터 지원이 가능하다.

는 20,000건이지만, 부정 부류와 긍정 부류의 개수를 동일하게 맞추어서 각각 7,068 건씩으로 샘플링하여 14,136건으로만 구성하였다.

<표 8> 평가 자료 구성

	점수	개수
부정 부류	1~5점	7,068
긍정 부류	6~10점	7,068
전체	1~10점	14,136

어휘 정보에 의한 감성 분류 실험은 <표 9>와 같다.

<표 9> 감성 어휘 정보에 의한 감성 분류 평가

	개수	옳게 분류한 개수	틀리게 분류한 개수	정확도
부정 부류	7,068	5,441	1,627	76.98%
긍정 부류	7,068	5,222	1,846	73.87%
전체	14,136	10,663	3,473	75.42%

어휘 정보만으로 긍정 부류에서는 76.98%와 부정 부류에서는 73.87%로 분류하였으며, 전체적으로는 75% 정도로 분류되었다. 어휘에만 의존한 감성 분석의 결과가 75% 정도로 담보될 수 있으며, 어휘 자원과 어휘 정보는 감성 분석에 있어 매우 기초적이며 강한 역할을 한다.

3.5. 오류 분석과 향후 연구

지금의 사례 분석 방법은 어휘 추출과 이를 적용한 감성 분석의 방식으로 성공 정확도가 그리 높지는 않다. 보다 구체적으로, 실험 평가 결과에서의 오분류는 어떠한 양상을 띄는지 살펴봄으로써 향후 연구를 위한 오분류의 예시를 살펴보고 현재 연구의 한계에 대해서도 짚어보도록 한다. 오분류된 글의 분포는 <표 10>과 같다.

<표 10> 오분류 문서 분포

	부정 부류					긍정 부류					합
평점 수	1	2	3	4	**5**	**6**	7	8	9	10	
틀린 문서 수	896	146	134	204	**247**	**312**	220	175	138	1002	3,473

<표 10>에서 굵게 표시한 부정 부류 5점대와 긍정 부류 6점대의 오류는 예측 가능한 오류였다. 긍정과 부정을 나누는 범주 경계에 있는 평점 5, 6점 글은 긍정이나 부정의 부류로 명백히 나누기가 모호할 것이기 때문이다. 부류 결정이 모호한 경우는 모호한 대로 분류하지 않는 것이 바람직할 수도 있을 것이다. 예시 4)는 평점이 5점이어서 부정 부류에 속해 있는 경우이고, 예시 (5)는 평점이 6점이어서 긍정 부류에 속해 있는 경우이다.

(4) 박쥐 : 작품성은 좋다고 하더라도 재미는 덜하다 (부정 부류)
(5) 해운대 : 지루한스토리였음. 기대를 너무 많이한듯... - -ㅜ한계인가 (긍정 부류)

그런데, 평점 1점과 10점에서 오류가 가장 많은 것은 예상하기 어려운 것이었다. 잘 분류된 1점이나 10점의 글은 매우 극단적으로 부정적이거나 긍정적인 표현을 담고 있는 것이 일반적이다. 그러나, 오분류된 1점이나 10점의 글 표현을 살펴보면, 반어적 표현을 하여 나쁜 평가 표현을 쓰면서도 점수를 10점을 주거나 좋은 평가 표현을 쓰면서도 점수는 1점을 주는 일이 있었다. (6)은 일반적인 형태의 부정 부류의 글이고 (7)은 일반적인 형태의 긍정 부류의 글이다. (8)은 긍정 표현을 쓰고 있으면서도 부정적인 평가 점수인 글이다.

(6) 해운대 : 엉성한 연기. ㅜㅜㅜ 최악! (일반적인 부정 표현)
(7) 해운대 : 이 영화 너무 감동적입니다. (일반적인 긍정 표현)
(8) 해운대 : 우와정말재밌다 (평점이 1인 부정 부류 : 반어적인 표현)

또, (9)~(11)은 '조잡하다', '쓰레기', '재미 없다', '별로다'와 같은 전형적인 부정 표현이 쓰였으나 모두 평점이 10점이 글들인데 부정 부류로 분석이 되어 실패한 경우이다. 이들은 평점과 긍정 부류 간에 일치가 되지 않는 경우들로 추측컨대, 영화 평점 작성을 아르바이트로 하는 이들의 글도 포함되었을 것으로 짐작된

다.

(9) 해운대 : 조잡한cg. 쓰레기같은 스토리. 어색한사투리. 그러나 김인권이 살렸다.
(10) 해운대 : 너무 재미없었어요.... 헬리콥터 보신분은 아실듯... 그게 CG???
(11) 해운대 : 별로였지만 짱준다

위에서 살핀 오분류의 경우들은 감성 분석의 방식을 개선하더라도 결과를 향상시키기 어려운 것들이다. 이처럼 어쩔 수 없는 표현들을 제외하고는 감성 분석의 처리 과정을 개선해 감으로써 그 정확도를 향상시킬 수 있을 것이다. 향후에 언어 표현의 특징을 고려해 감으로써 개선해가야 할 것들의 유형과 예시를 들어보면 다음과 같다.

o. 평가 표현 자체가 없는 경우

(12) 해운대 : 심형래 이 영화보고 반성좀해라..
(13) 해운대 : 예고편이랑 좀 다른듯 .....ㅠ

(12)는 영화에 대해서 나쁘거나 좋다는 감성 표현은 없지만, 추론하자면 긍정적인 평가의 뜻이고, (13)은 직접적인 표현이 없어도 뭔가 부정적인 뜻을 담고 있다. 이러한 경우는 어떤 식의 언어 처리 방식이 도입될 수 있을지 모호하다. 적어도 평가 표현이 없으므로 처리 대상이 되지 못하는 것을 판별해내는 것으로도 의미가 있을 것이다.

o. 평가 표현 어휘의 주어가 평가 대상이 아닌 경우

(14) 해운대 : 국가 대표가 훨씬 재밋던데
(15) 해운대 : cg는 정말 멋있지만 전 별로... 국가대표가 훨 낫네요.

(14), (15) 모두에서 '훨씬 재미있다', '정말 멋지다' 등의 긍정적인 표현을 사용하였으나 이 긍정적인 평가의 대상은 평가 대상이 되는 '해운대'가 아니라 다른 비교 대상을 가리키고 있다. 따라서, 평가 표현의 대상이 무엇인지에 대해 판별해내는 구문 분석의 과정이 필요하겠다.

o. 연결 어미, 종결 어미 패턴을 보아서 뉘앙스를 알아야 하는 경우

(16) 해운대 : 이게 진정 재밌나??
(17) 해운대 : 순간순간 빵터지는 코미디영화..재난영화맞어?

(16), (17)은 '재밌다', '빵터진다'와 같은 긍정적인 표현을 쓰고 있기는 하지만, 의문형 종결어미로 끝내는 것으로 일차원적인 어휘 정보만으로는 한계가 있고, 문장 분석 수준의 과정이 추가될 필요성을 보이고 있다.

이 절에서는 오류 분석을 통해 감성 분석의 처리 과정에서 어휘 자원을 이용하는 것에서 나아가 구문 분석, 문장 분석 수준의 의미 분석 과정이 필요로 함을 보여주었다. 또한, 이러한 문장 분석 또한 감성적 특징을 보이는 어휘를 중심으로 진행되어야 하는 방향에서는 변함이 없을 것이다.

4. 맺음말

본고에서는 최근 연구가 활성화되고 있는 감성 분석에 대한 소개를 하고, 영화 평점 말뭉치를 가지고 사례 분석을 수행하였다. 수많은 감성 분석의 연구들이 어휘 자원을 활용했을 뿐 어휘 자원의 활용이 감성 분석에 있어 얼마나 기여를 하는지에 대한 분석 결과는 없었으며, 어휘 자원은 대개 어디선가 활용될 수 있도록 있어야 하는 자원이었다. 이에 네티즌들이 직접 작성한 영화 평가문 말뭉치에서 감성 분석을 위해 활용될 수 있는 어휘 자원을 추출하는 과정을 보이고, 어휘 자원의 활용이 감성 분석에 있어 얼마나 중요한 역할을 하는지에 대한 실험적인 분석을 수행하였다. 사례 분석 과정에서 말뭉치를 긍정적인 감성의 부류와 부정적인 감성의 부류로 나누었을 때, 각 부류의 어휘들의 감성적 경향을 나타내는 확률값을 구해 보았고, 이를 가지고 긍정 혹은 부정에 대한 감성 강도를 양적으로 구해보았다. 의미 분석의 한 응용으로써 감성 분석을 수행하고자 할 때의 어휘 정보가 가장 기본적인 자원인지를 보이기 위해 또 다른 말뭉치에서의 감성 분석에서 활용했을 때의 커버리지를 보이는 실험을 하였다. 감성 분석 과정에서 단

순히 어휘 사전을 활용했을 때와 감성 강도 정보가 있는 어휘 사전을 활용했을 때의 차이를 기술하였으며, 감성 분석에 있어 어휘 자원만의 활용만으로도 의미 있는 분석 결과를 낼 수 있음을 보임으로써 어휘 자원의 역할이 매우 중요함을 살폈다.

참고문헌

고민수, 신효필. 2010. “감정어휘 평가사전과 의미마디 연산을 이용한 영화평 등급화 시스템”, 「인지과학」 21-4, 669-695.

명재석, 이동주, 이상구. 2008. “반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템”, 「정보과학회논문지 : 소프트웨어 및 응용」 35-6, 392-403.

박인조, 민경환. 2005. “한국어 감정단어의 목록 작성과 차원 탐색”, 「한국심리학회지: 사회 및 성격」 19-1, 109-129, 한국심리학회.

서상규 편. 1999.「언어 정보의 탐구,. 연세대학교 언어정보개발연구원, 도서출판 월인.

안애림. 2011. “한국어 오피니언 문장 분류 시스템을 위한 사전 및 구문 패턴 연구”, 언어인지과학과 석사 학위 논문, 한국외국어대학교.

윤애선, 권혁철, 2010. “감정 온톨로지의 구축을 위한 구성요소 분석”, 「인지과학」 21-1, 157-174.

이지영. 2009. “한국어 교육을 위한 감정 표현 연구”, 「한국어의미학」 29, 201-227.

임지룡. 2006. “의미교육의 학습 내용에 대하여”, 「한국어학」 33, 87-116.

임지룡. 2010. “감정의 그릇 영상 도식적 양상과 의미특성”, 「國語學」 57. 31-73.

최경봉. 2010. “50년 - 의미 연구의 성과와 전망”, 「國語學」 57. 421-468.

Beineke, Philip, Trevor Hastie, Shivakumar Vaithyanathan. 2004. “The Sentimental Factor: Improving Review Classification via Human-provided Information”, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics,* 263-271.

Choi, Yejin & Claire Cardie. 2008. “Learning with Compositional Semantics as

Structural Inference for Subsentential Sentiment Analysis", *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 793-801.

Jang, Hayeon & Hyopil Shin. 2010. "Language-specific Sentiment Analysis in Morphologically Rich Languages", *Proceedings of the 23rd International Conference on Computational Linguistics,* 498-506.

Manning, Christopher D. & Schütze Hinrich. 1999. *Foundations of Statistical Natural Language Processing,* MIT press.

Manning, Christopher D., Raghavan Prabhakar & Schütze Hinrich. 2008. *Introduction to Information Retrieval,* Cambridge University Press.

Min, Hye-Jin & Jong C. Park. 2007. "Representing Emotions with Linguistic Acuity", *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing,* 348-360.

Moilanen, Karo and Stephen Pulman. 2007. "Sentiment Composition", *Proceedings of Recent Advances in Natural Language Processing,* 378-382.

Theresa, Wilson, Janyce Wiebe, Paul Hoffmann. 2005. "Recognizing Contextual Polarity in Phrase-level Sentiment Analysis", *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347-354.

Turney, Peter. 2002. "Thumbs Up or Thumbs Down?: Semantic Orientation Aapplied to Unsupervised Classification of Review", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* 417-424.

Valentin, Jijkoun, Maarten de Rijke & Wouter Weerkamp. 2010. "Generating Focused Topic-Specific Sentiment Lexicons", *Proceedings of the 48th Annual Meeting on Association for Computational Linguistics,* 585-594.

조은경
120-140 서울시 서대문구 신촌동 134 연세대학교 언어정보학협동과정

134-715 서울 강동구 암사3동 롯데캐슬퍼스트 124동 501호
전화번호 : 010-4149-9134
전자우편 : jek.cl.nlp@daum.net

투고논문접수일	2012년 5월 9일
논 문 심 사 일	2012년 5월 17일
심 사 완 료 일	2012년 6월 12일