

사전기반의 한국어 상품 리뷰 의견표현 자질 추출 및 분류시스템

육상근

고려대학교 컴퓨터정보통신대학원
tazy@dreamwiz.com

Dictionary-Based Opinion Features Extraction and Classification of Korean Product Reviews

Sanguen Yuk

Graduate School of Computer and Information Technology, Korea University

요 약

인터넷을 이용한 사람들의 사회 참여가 확대되면서 다양한 의견(Opinion)들이 급속도로 증가하고 있으며 이러한 의견을 분석하여 유용한 정보로 활용하기 위한 연구가 활발히 진행되고 있다. 그 중에서도 상품리뷰는 기업에서 연구, 개발, 마케팅의 주요 자료로 사용되고 있으며 사용자가 상품의 구매를 결정하는 중요한 요인 중 하나로 작용하고 있다. 본 논문에서는 한국어로 이루어진 상품 리뷰를 분석하여 의견 자질(Feature)을 추출하고 분류(Classification)하는 시스템을 설계하고 구현하였다. 한글 의견 자질 추출을 위하여 먼저 한글 상품 리뷰를 분석하여 의견 사전을 구축하였다. 의견 사전으로는 의견 자질과 의견 어휘, 독립의견어휘, 의견 숙어, 부정어 등의 각기 다른 세부 사전을 구축하여 리뷰 분석 시 단계적으로 적용하여 정확도를 높일 수 있도록 설계하였다. 이렇게 구현된 시스템을 평가하기 위하여 각기 다른 3개의 도메인에서 실제 한국어 리뷰를 수집하여 실험을 수행하였으며 자질 추출에서는 평균 78.86% 정확률, 61.41% 재현율을, 극성 분류에서는 평균 69.46% 정확률, 42.26% 재현율을 나타냈다.

1. 서 론

‘참여, 개방, 공유’의 슬로건을 내건 WEB 2.0의 시대가 오면서 인터넷을 이용한 사람들의 의견(Opinion) 표현이 적극적이고, 다양해지며, 그 양도 급속도로 증가하고 있다. 또한 사람들의 의견은 특정 분야에 국한되지 않고 정치, 경제 등 사회의 모든 분야에서 활발하고 적극적으로 나타나고 있다. 특히 상품 리뷰는 많은 기업에서 분석을 통하여 얻어지는 정보를 이용하여 상품 개발, 개선, 마케팅 전략 등과 같은 많은 부분에서 활용하고 있다. 이를 위하여 인터넷 상품 리뷰를 자동으로 수집, 분석하여 이용자의 의견을 추출한 후 유용한 정보로 가공할 수 있다면 그 쓰임은 다양할 것이다.

현재 이러한 사용자 의견 분석시스템 구현을 위한 연구가 영어권을 중심으로 활발히 진행되고 있는 반면, 한국어를 대상으로 한 연구는 매우 초보적인 수준에 머물러 있다. 이에 따라 본 논문에서는 한글 의견 사전(Opinion Dictionary)을 구축하고, 이를 이용하여 한국어 상품 리뷰에서 사용자의 의견을 파악하고 의견 극성을 분류하는 시스템을 제안한다.

한글 의견 사전에 수록된 어휘는 웹에서 수집한 사용자 리뷰 문서로부터 추출하였으며, 수록되는 어휘의 성격에 따라 총 4개의 사전을 구축해 의견 극성(Polarity) 분류 시스템에서 단계적으로 사용할 수 있도록 하였다. 의견 극성 분류 시스템은 주어진 문장에서 사전에 등록된 어휘가 출현했는지 여부를 통해 문장에 나타난 의견이 긍정적인

의견인지, 부정적인 의견인지를 판단한다.

사전을 이용한 방법은 비교적 소량의 말뭉치만을 필요로 하며, 구축된 사전은 향후 다른 연구에서 중요한 자료로 활용될 수 있다는 점에서 현재 관련 언어 자원이 충분치 않은 한국어 Opinion Mining에 적용하기에 적절한 방법이라 판단된다.

본 논문의 구성은 다음과 같다. 2장에서는 이와 관련된 기존의 연구들에 대하여 알아보았으며, 3장에서는 본 논문에서 구현한 ‘의견 자질 추출 및 분류시스템’의 구성에 대하여 설명한다. 4장에서는 구현된 시스템에 대한 실험 및 결과 분석, 마지막으로 5장에서는 결론 및 향후 과제에 대하여 논의하고자 한다.

2. 관련 연구

2.1 Opinion Mining

의견(Opinion)이란 특정의 문제나 사항, 인물이나 사물에 대한 태도를 언어로 표현한 것이다[1]. Opinion Mining은 문서에서 작자가 해당 주제(Product)에 대해 표현한 의견(Opinion) 혹은 감성(Sentiment)을 찾아내고 파악하는 것이다[2]. 이러한 의견을 파악하는 것은 기존의 다양한 연구 분야와도 깊은 연관성을 가지고 있다. 사용자가 원하는 정보를 검색, 요약, 분류한다는 점에서 기존의 정보 검색, 문서 요약, 문서 분류 등의 주제에서 사용하는 방법이 적용 가능하다. 그러나 앞서 언급한 분야들이

주로 문서 단위로 사용자의 필요 정보를 추출했다면 Opinion Mining은 대상의 특징을 기준으로 검색, 요약, 분류를 수행한다는 다른 점이 있다. 데이터 단위의 관계에 주목한다는 점에서 데이터마ining 분야와도 깊은 연관을 맺고 있으며 Opinion Mining은 여기에 언어적인 기반과 사용자의 선호 여부를 나타내는 극성을 부여함으로써 좀 더 세부적인 분야로 자리 잡았다. 또한 사용자의 의견을 분석하는데 언어학적인 자원을 활용함으로써 전산언어학 분야와도 깊은 관계를 맺고 있다[3].

이러한 Opinion Mining의 연구 분야는 다음과 같이 나누어 볼 수 있다.

- **Opinion Classification** : 의견을 담고 있는 문서안의 긍정 또는 부정의 의견을 분류하는 것
 - 사전기반 : 사전에 정의된 의견들의 긍정, 부정의 값을 이용하여 분류.
 - 규칙기반 : 수집된 규칙을 적용하여 분류.
 - 기계학습기반 : 미리 분류되어 있는 의견 학습데이터를 이용한 기계학습을 통한 분류.
- **Opinion Retrieval** : 많은 문서 중에서 주어진 질의어에 적합한 의견 문서를 찾는 것.
- **Opinion Summary** : 의견문서를 분석하여 다양한 의견 자질을 기준으로 요약하는 것.

본 연구는 다양한 Opinion Mining 분야에서 Opinion Classification에 해당되며, 방법론 측면에서 사전 기반에 해당한다.

3. 한국어 의견표현 자질 추출 및 분류시스템

3.1. 시스템 개요

(그림 1) 시스템 구성도

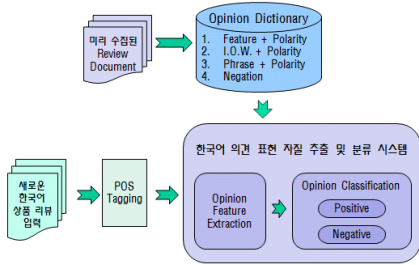


그림 1은 한국어 의견표현 자질 추출 및 분류시스템의 전체적인 구성을 보여준다. 가장 먼저 선행되는 작업은 의견 자질 추출과 분류에 사용될 의미 사전의 구축이다. “의견 자질(Feature)+의견 어휘(Opinion Word)”, “독립의견어휘(IOW:Independent Opinion Word)”, “의견 숙어(Opinion Phrase)”, “부정어(Negation)” 등 총 4가지 사전을 구축하였다. 부정의 사전의 경우 하나의 사전으로 처리 단계별 중복 사용이 가능하도록 설계하여 효율을 높였다.

분석할 한국어 상품 리뷰 문장이 입력되게 되면 우선 형태소 분석기를 거친 후 우선적으로 사전에 등재되어 있는 자질을 추출하게 된다. 하지만 사전에 모든 자질을 등록하여 활용할 수 없으므로 미등록 자질을 유추하여 추출하게 된다. 이때는 의견 어휘를 중심으로 어휘간 거리 구

칙을 정하여 일정 범위 안의 특정 품사의 단어를 자질로 선정하여 추출하게 된다. 의견 분류의 단계에서는 이러한 자질의 의견어휘가 가지고 있는 극성 값을 찾고, 주변을 검색하여 부정어 출현 여부를 파악한 후 이를 반영하여 해당 자질이 문서에서 실제로 어떤 의견 자질인지, 즉 긍정적인 의견을 나타내는 자질인지, 부정적인 의견을 나타내는 자질인지를 분류한다.

3.2. 의견 사전(Opinion Dictionary)

본 시스템에서 사용되는 의견 사전은 “의견이 표현된 단어 또는 구와 분류값의 쌍(Pair)으로 구성된 목록”으로 정의 된다.

<표 1> 사전 구축 기초 데이터

제품(Product)	핸드폰, 디지털카메라	
대상 리뷰	제품별 각 30개 리뷰 문서	
출처	핸드폰	세터즌(www.cetizen.com) - 30개 리뷰 문서
	디지털 카메라	디시인사이드(www.dcsinside.com) - 10개 리뷰 문서 SLR클럽(www.slrclub.com) - 20개 리뷰 문서
수집된 의견문장	총 573개	

사전 구축을 위한 단어는 문장에서 사용된 원형 그대로를 모두 추출한 후, 형태소 분석기를 이용하여 품사 부착(POS Tagging)을 거쳐 발견되는 중복성을 제거한 후 사용되었다. 이러한 이유는 <표2>와 같이, 같은 의미의 단어 “높”이 동일한 의견 자질인 “수신/NGG”에서 활용된 형태나 다음 단어에 따라 다르게 분석되는 차이점을 파악하여 사전 구축 시 이를 모두 포함하여 구축하기 위함이다.

<표 2> 동일 의견 자질에서 의견 어휘의 다른 품사 태깅

수신률이 높아졌습니다	수신/NGG+를/XSN+이/JKS 높 /VA+아/EC+지/VX+였/EP+습니다/EC
수신률이 높았다	수신/NGG+를/XSN+을/JKO 높 이/VV+였/EP+다/EC

3.2.1. Feature + Opinion Word

자질(Feature) : 의견의 대상이 되는 제품의 특성을 나타내는 단어이다. 품사는 대부분 명사(NN)이며 동사(VV), 부사(MA), 외국어(SL)도 존재한다.

의견 어휘(Opinion Word) : 의견을 담고 있는 단어로서 자질과 쌍을 이룬다. 품사는 명사(NN), 동사(VV), 형용사(VA)가 대부분이며 어근(XR)도 상당부분 출현한다. 이때 같은 의견 어휘라도 주인이 되는 자질에 따라 그 분류값(Polarity)이 달라진다.

<표 3> 자질 + 의견 어휘 + 분류값

<예문1> 메뉴를 이용할 때는 화면이 작은 느낌을 지을 수 없었다.		
<예문2> 나름 작은 크기와, 좋은 화질로 인기가 많았던 녀석이죠.		
Feature	Opinion Word	Classification
화면/NGG	작/VA	부정(Negative)
크기/NGG	작/VA	긍정(Positive)
화질/NGG	좋/VA	긍정(Positive)
인기/NGG	많/VA	긍정(Positive)

3.2.2. I.O.W.(Independent Opinion Word)

독립의견어휘(IOW:Independent Opinion Word) : 자질이 생략된, 단독으로 사용된 의견을 담고 있는 단어이다. 품사는 명사(NN), 동사(VV), 형용사(VA)등 다양하게 존재한다.

<표 4> I.O.W + 분류값

<예문>때문에 기존의 애니콜 키 배지에 익숙한 사용자들은 상대적으로 불편함을 느낄 수도 있을 것 같다
불편/NNG (I.O.W) : 부정(Negative) - '사용(Feature)'이 생략됨

3.2.3. Opinion Phrase

의견 속어(Opinion Phrase) : 사람의 의견이나 감정을 나타내는 구이다. 이때 단어는 하나하나의 의미로 해석되지 않고 그룹으로 묶여서 전혀 다른 의미로 사용된다. 보통 2개의 단어가 쌍으로 붙어서 출현하게 되며, 특정 품사에 국한되지 않고 다양한 품사가 다양한 조합을 통하여 등장하게 된다. 이러한 의견 속어 사전은 “단어1+단어2”의 형식으로 제작 하였으며 분류값(Polarity)을 가지게 된다.

<표 5> Opinion Phrase + 분류값

<예문>수출 제품에 비해 내수용 휴대폰은 '걸음마 수준'이라고 해도 과언이 아닐 것이다	
걸음마/NNG + 수준/NNG	부정(Negative)

3.2.4. Negation

부정어(Negation) : 부정의 의미를 가진 단어이다. 부정어가 출현하게 되면 그 분류값이 역전된다. 한국어에서 부정어의 위치는 의견을 나타내는 단어의 앞, 뒤, 또는 문장의 맨앞, 맨뒤 어느 곳이나 올수 있으며, 그 영향력이 미치는 범위 1단어, 2단어 이상, 해당 문장 전체 등 다양하다.

<표 6> Negation + 분류값

<예문>숫자 키패드들의 색상이 조금씩 달라 역시 단조로운 음을 지혜롭게 극복하였다
색상/NNG(Feature) + 단조롭/VA(Opinion Word) + 극복/NNG (Negation) => 긍정(Positive)

3.3. 한국어 상품 리뷰의 처리 과정

한국어 상품 리뷰는 가장 먼저 형태소 분석기를 거쳐 “형태소/품사”의 쌍의 분석된 값을 가지게 된다. 이렇게 형태소 분석 후 품사 태깅을 하는 이유는 사전의 미등록 자질 추출에 활용하기 위해서 이다. 형태소 분석이 완료된 한국어 상품 리뷰는 이후 의견 어휘 사전을 이용하여 분석에 들어가게 된다.

<표 7> 구축된 의견 사전

의견자질+의견어휘	6,723개
독립의견어휘	249개
의견속어	25개
부정어	35개

3.3.1. 의견 자질 및 의견어휘 추출

의견 자질 및 의견어휘 추출은 우선적으로 문장의 맨 앞 쪽부터 탐색이 시작되며 다음의 단계를 거치게 된다.

- Step1**, 사전에 등록된 ‘자질’의 추출.
- Step2**, 추출된 자질을 기준으로 전후 3단어 이내의 ‘의견어휘’ 추출
- Step3**, Step2까지 제외된 단어를 대상으로 ‘독립의견어휘’ 추출
- Step4**, 추출된 독립의견어휘를 기준으로 전후 3단어 이내의 ‘자질(명사)’를 추출
- Step5**, Step4까지 제외된 단어를 대상으로 ‘의견속어’ 추출

<표 8> 의견 자질 및 어휘 추출 결과

<예문>SCH-C210의 단점을 보완하여 전면의 네비게이션 버튼이 크고 키감이 좋아 손가락이 큰 사람이 사용해도 불편함이 없으며 수화부 걸 스피커도 장착되어 있는 것이 눈에 띈다.
<Find1> 버튼(NNG) : ‘자질사전’
<Fine1-1> 크(VA) : ‘버튼(NNG)’과 거리가 ‘1’인 ‘의견어휘사전’
<Find2> 키감(NNG) : ‘자질사전’
<Find2-1> 좋(VA) : ‘키감(NNG)’과 거리가 ‘1’인 ‘의견어휘사전’
<Find3> 단점(NNG) : ‘I.O.W사전’
<Find4> 불편(NNG) : ‘I.O.W사전’
<Find4-1> 사용(NNG) : ‘불편(NNG)’과 거리가 ‘1’인 명사 유추
<Find5> 눈(NNG) : ‘속어사전’
<Find5-1> 띄(VV) : ‘눈(NNG)’과 거리가 ‘1’인 ‘속어사전’

3.3.2. 극성 분류

극성 분류는 ‘3.3.1. 의견 자질 및 의견어휘 추출’의 단계에 ‘부정어’ 추출을 추가하여 진행하게 된다. 이때 분류값은 사전에 정의된 값을 이용한다. 이때 추가되는 과정은 다음과 같다.

- Step1**, ‘자질+의견어휘’ 추출 후, ‘의견어휘’의 전후 3단어 이내의 ‘부정어’ 추출
- Step2**, ‘독립의견어휘’ 추출 후, 전후 3단어 이내의 ‘부정어’ 추출
- Step3**, ‘의견속어’ 추출 후, 전후 3단어 이내의 ‘부정어’ 추출

<표 9> 극성 분류 결과

* 표8과 동일 예문 사용	분류값
추출된 어휘	
<Find1> 버튼(NNG)+크(VA)	긍정(Postive)
<Find2> 키감(NNG)+좋(VA)	긍정(Postive)
<Find3> 단점(NNG)	부정(Negative)
<Find3-1> 단점(NNG)+보완(NNG)	긍정(Postive)
<Find4> 사용(NNG)+불편(NNG)	부정(Negative)
<Find4-1> 사용(NNG)+불편(NNG)+없(VA)	긍정(Postive)
<Find5> 눈(NNG)+띄(VV)	긍정(Postive)

4. 실험

4.1. 수행 방법 및 결과

본 시스템의 의견 자질 추출 및 분류의 결과를 확인하기 위하여 사전 구축 때 사용되었던 것과 “같은 도메인(핸드폰)”, “유사 도메인(네비게이션)”, “다른 도메인(화장품)” 이렇게 3개의 도메인으로 나누어 실험을 수행하였

다.

<표 10> 실험 데이터

	핸드폰	네비게이션	화장품
출처	www.citizen.com	www.navinside.com	cafe.daum.net/cosmetic1
데이터	37문장	49문장	39문장
자질	109개	113개	85개

성능 평가의 방법으로는 실제의 상품평을 대상으로 한 한국어 분석 시스템이라는 점으로 인하여 다른 시스템과 직접 비교하는 것은 어려운 문제였다. 따라서 실험에 사용된 한국어 상품 리뷰에 대하여 사람이 직접 보고 의견 자질을 찾고, 분류한 값과 시스템이 분석한 결과를 비교하였다.

<표 11> 자질 추출 및 분류 결과

	자질 추출		의견 분류	
	정확률(%)	재현율(%)	정확률(%)	재현율(%)
핸드폰	91.30	77.06	74.07	58.82
네비게이션	82.22	65.48	72.41	46.66
화장품	63.06	44.70	61.90	21.31

<표 12> 자질 유추 결과

	핸드폰	네비게이션	화장품
정확률(%)	81.25	70.96	60.46
재현율(%)	72.22	61.11	41.93

4.2. 결과 분석

실험 결과 정확률의 경우에는 비교적 높은 비율을 보여 주고 있으나 재현율은 낮아 아직까지 보완할 점이 많음을 보여 주고 있다. 특히 사전 구축을 위하여 기존에 수집, 분석한 제품과 동일한 도메인에서는 정확률과 재현율 모두 비교적 높은 점수가 나왔다. 하지만 다른 도메인(화장품)에서는 재현율이 현저히 낮았다. 이는 디지털가전과 화장품의 리뷰에 사용되는 의견 자질이 매우 상이하기 때문이다. 디지털 가전에서는 주로 “액정, 화질, 해상도..” 등과 같은 단어들 이 자질로 사용 되었으나 화장품에서는 “피부, 수분, 조직..” 등의 단어가 사용됨을 보여 주었다. 하지만 “기능, 효과, 사용..” 등의 자질은 공통적으로 의견 표현에 자주 사용되는 자질이였으며, 이러한 자질의 영향으로 정확률 하락은 재현율에 비해 상대적으로 적었다.

정확률에 비하여 재현율이 현저히 낮은 것은 사전 기반 시스템에서 일반적으로 지적되는 문제이다. 이의 개선을 위하여 본 시스템에서는 의견 어휘와 의견 자질간의 거리를 이용한 미등록 의견 자질의 유추를 제시하였다.

5. 결론

의견 분석 연구의 가장 기본이 되는 것이 의견 자질의 추출과 분류이다. 이에 따라 본 논문에서 제안한 사전 기반의 의견 자질 추출 및 분류 시스템은 여러 용도로 확장하여 활용이 가능한 시스템으로 유용하게 사용되리라 생각된다.

본 연구에서는 한국어 상품 리뷰의 의견 자질 추출 및

분류를 위하여 의견 사전을 활용 하였다. 실험에서는 높은 정확률에 비해 낮은 재현율을 보여주었으며, 이는 향후 기계학습이나 규칙활용등을 통한 보완이 필요하다. 또한 다양한 도메인에서 활용할 수 있도록 의견사전을 좀 더 다양한 제품군으로 확장하고자 한다.

참고문헌

[1] 두산백과사전 EnCyber & EnCyber.com
 [2] M. Rimon, Sentiment Classification : Linguistic and Non-Linguistic Issues Hebrew University
 [3] 명재석, 이동주, 이상구, 반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템, 정보과학학회지, 2008
 [4] Bing Liu, Opinion Mining & Summarization : Sentiment Analysis, Proceedings of the 17th international World Wide Web conference(WWW-2008), April 21, 2008, in Beijing, China.
 [5] 황재원, 고영중, 감성 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템, 정보과학학회지. pp.336-340, 2008
 [5] Soo Min Kim, Identification, Classification, and Analysis of Opinions on the Web, University of Southern California 2006
 [6] Hong Yu, Vasileios Hatzivassiloglou, Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, Columbia University. 2003
 [7] B. Pang, L. Lee and S. Vaithyanathan, Thumbs up? Sentiment Classification Using Machine Learning Techniques, EMNLP, 2002.
 [8] Turney,P.D., 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), Philadelphia, Pennsylvania, 2002
 [9]K. Dave, S. Lawrence, D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," In Proceedings of WWW 2003, Budapest, Hungary, 2003.
 [10] 인주호, 감성표현을 통한 문서 분류 시스템, 한국항공대학교, 2008
 [11] 윤훈용, 임기용, 윤우순, 스포츠화에 대한 감성만족도 분류체계에 관한 연구, 대한인공과학회 논문집. 2000