한국어 감정분석 코퍼스를 활용한 양상정보 기반의 감정분석 연구

Modality-based Sentiment Analysis through the Utilization of the Korean Sentiment Analysis Corpus

저자 (Authors)
신효필, 김문형, 박수지
Hyopil Shin, Munhyong Kim, Suzi Park

출처 (Source)
언어학 (74), 2016.4, 93-114(22 pages)
EONEOHAG : JOURNAL OF THE LINGUISTIC SOCIETY OF KOREA (74), 2016.4, 93-114(22 pages)

발행처 (Publisher)
사단법인 한국언어학회
The Linguistic Society of Korea

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06664340

APA Style
신효필, 김문형, 박수지 (2016). 한국어 감정분석 코퍼스를 활용한 양상정보 기반의 감정분석 연구. 언어학(74), 93-114

한국어 감정분석 코퍼스를 활용한 양상정보 기반의 감정분석 연구*

신효필 · 김문형 · 박수지**

- 차 례 -

1. 서론

현재 컴퓨터언어학 또는 자연언어처리에서 감정분석(sentiment analysis) 내지 의견분석(opinion analysis) 연구가 활발히 이루어지고 있다. 감정분석의 연구는 기본적으로 긍정, 부정 또는 중립으로 분류된 어휘(lexicon)들의 빈도에 따른 연산에 기초하고 있다. 이런 감정어휘들은 문서의 극성을 분류하는데 일차적으로 활용되는 자료이지만 어휘를 단순히 긍부정으로 분류하기 어려운 경우도 많으

언어학 제 74 호 (2016. 4. 30: 93-114), 사단법인 한국언어학회
* 이 논문은 2013년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2013S1A5A2A01017908).
** 서울대학교 인문대학 언어학과(hpshin@snu.ac.kr)

DOI 10.17290/jlsk.2016..74.93

며 같은 어휘라도 상황에 따라 극성이 바뀌는 경우도 존재하기 때문에 전적으로 극성어휘 연산에만 의존하는 분석은 한계가 있다.

한편 감정분석에 있어 감정어휘와 더불어 감정표현이 정교하게 주석된 코퍼스가 학습용이나 실험용으로 필요하다. 영어의 경우는 Multi-Perspective Question Answering(MPQA, 2005)이 감정분석의 대표적인 자료로 사용되고 있다. 이 코퍼스는 대략 1,000여 문장 안에 나타나는 감정표현들을 그 의미를 잘 나타낼 수 있는 주석언어를 이용하여 주석하였다. 한국어의 경우는 Shin et al. (2012)에 의해 MPQA와 같은 주석 원리와 한국어 특징을 반영한 주석 자질로 한국어 감정분석 코퍼스(KOrean Sentiment Analysis Corpus, KOSAC)가 구축되었다. 이 자료는 현재 웹으로 공개되어[1] 감정분석을 위한 기본 자료로 활용되고 있다.

본 연구는 한국어감정분석코퍼스에서 주석자질에 따른 감정표현을 추출하여 감정분석연구에 유용한 자원을 구축하는 것에서부터 시작한다. 긍정 및 부정어휘 뿐만 아니라 주관성 유형(subjectivity type)과 강도(intensity)에 따라 다양하게 나타나는 표현들을 추출하여 목록화한다. 이 코퍼스에서 추출되는 감정표현 단위는 주석 자질의 속성 중 표현-유형(expressive-type), 강도(intensity), 원천자 중첩 횟수(nested source-order), 극성(polarity), 주관성-극성(subjectivity-polarity), 그리고 주관성-유형(subjectivity-type)이다. 추출되는 감정표현의 대부분의 경우는 단어, 연결어미, 조사 등의 기능어이거나, 이것이 더 확장된 구의 형태로 되어 있다.

기존 코퍼스를 활용하여 감정표현 언어자원을 구축한 후 주관성 표현방법을 명시하는 주석 유형(type)을 중심으로 양상표현을 비롯한 화용적 자질을 감정연구에 통합하는 방법을 모색한다. 유형 속성은 direct-explicit, direct-action, direct-speech, indirect, writing-device의 하위 속성으로 이루어진다. 유형의 속성들은 서술자 및 화자의 주관성을 반영하는 표현을 주석한 것으로 간접적인(indirect) 경우를 제외하고는 주관성의 표현이 행위(action), 화행(speech act), 그리고 명시적(explicit)으로 나타나는 경우와 화행을 통해 이루어지지 않는 경우(writing-device)로 나뉜다. 이를 구성하는 표현들은 화자 지향의 부사어와 같은 내용어도 있지만 상당수가 접속부사, 접속어미와 이것이 확장된 구로 되어 있다. 따라서 이 유형에 의해 표시되는 양상정보(modality)와 화용론적 정보를 감정분석에 활용하는 것이 분석의 정확성 및 정밀성을 제고하는데 필요하다.

양상정보를 감정분석에 적용한 연구는 Shin (2014)에서 시도되었다. Shin (2014)에서는 화자뿐만 아니라 다른 사람의 개인적 상태(private state)가 어떻게 발화행위에 반영되느냐에 따라 문장의 주관성이 결정된다고 본다. Shin (2014)에서는 이를 감정분석 코퍼스의 각 문장에서 감정의 원천자가 중첩된 횟수의 분

1) http://word.snu.ac.kr/kosac

포로 파악하고 있다. 발화에서 중첩된 원천자(nested order)가 많으면 많을수록 화자 자신보다는 다른 사람의 상태가 더 많이 드러나게 되어 그 문장은 더 객관적으로 된다는 것이다. 실제로 주석된 자료에서도 중첩된 원천자와 문장의 객관성 사이에 일정한 상관관계를 보이고 있다. 그러나 이 중첩된 원천자는 언어적 표현으로 명시적으로 드러나기 보다는 원천자를 도입하는 술어(source introducing predicates)에 의해 간접적으로 파악되고, 또 코퍼스에서 정규화하기 어려운, 제한된 수의 표현으로 되어 있기 때문에 큰 규모로 감정분석 연구에 활용하기는 어렵다.

본 연구에서는 중첩된 원천자 이외에 다른 언어적 기제에 의해 드러나는 양상과 같은 화용적 정보에 초점을 맞춘다. 여기서 양상은 서술자 및 화자의 주관성을 나타내는 다양한 종류의 문법기제를 포괄하는 개념으로 사용한다. 양상은 관점에 따라 다양하게 분류되지만 크게는 명제적 양상(propositional modality)과 사건 양상(event modality)으로 대별된다(Palmer 1986). 명제적 양상은 화자의 명제에 대한 태도에 초점을 맞추고 사건 양상은 화자의 잠재적 미래 사건에 대한 태도에 초점을 맞춘다. 이 분류에 의하면 본 연구에서의 양상은 명제적, 인식론적(epistemic) 또는 증거성(evidential) 양상에 해당한다.[2] 이를 위해 본 논의에서는 한국어감정분석 코퍼스에서 양상과 같은 화용적 정보를 가장 잘 드러내는 자질로 주석되어 있는 표현들을 추출하여 주관성 분류 실험을 통해 그 중요성을 살펴볼 것이다.

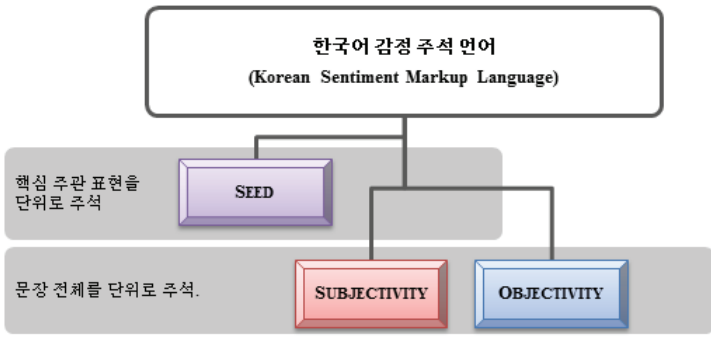
2. 한국어감정분석코퍼스

2.1 한국어 감정주석 언어

한국어감정분석코퍼스(KOSAC)는 한국연구재단의 지원을 받아 2011년부터 2013까지 2년에 걸쳐 구축되었다(Shin et al. 2012, 김문형 외 2013). 이 주석 작업을 위해 한국어 감정주석 마크업언어(the Korean Sentiment Markup Language)가 설정되었으며, Seed 태그, Subjectivity 태그, Objectivity 태그의 총 세 개의 태그로 이루어진다. 이를 도식화하면 다음과 같다.[3]

2) 양상과 감정분석의 관계에 대해서는 Shin (2014)를 참조.

3) 여기서의 기술은 상당부분 한국어 감정분석 코퍼스 구축 시 주석 가이드라인으로 사



<그림 1> 한국어 감정주석언어

SEED는 텍스트 상에 명시적으로 주관성을 드러내는 표현을 주석하기 위한 태그로 문장보다 작은 단위의 표현이 가진 주관성을 포착하는데 사용된다. SEED의 속성은 다음과 같이 정의된다.

<표 1> SEED의 태그 속성 목록

anchor: morpheme id(s)
id: tag id
nested-source: w-CDATA-CDATA
target: target id(s)
expressive-type: direct-explicit, direct-speech, direct-action, indirect, writing-device
subjectivity-type: emotion-pos, emotion-neg, emotion-neutral, emotion-complex, judgment-pos, judgment-neg, judgment-neutral, agreement-pos, agreement-neg, agreement-neutral, argument-pos, argument-neg, argument-neutral, intention-pos, intention-neg, speculation-pos, speculation-neg, others
polarity: positive, negative, neutral, complex
intensity: low, medium, high
insubstantial: TRUE, FALSE

용되었던 Korean Subjectivity Markup Language Guideline에 기초하고 있다. 이 가이드라인은 http://word.snu.ac.kr/kosac에서 살펴볼 수 있다.

2.2 표현 유형(Expressive type) 주석 속성

<표 1>에 제시되어 있는 태그 속성 중 주관표현이 주관성을 어떤 방식으로 표현하는지를 나타내는 표현 유형(expressive type) 속성에 대해 살펴보자. 이 유형은 다음과 같이 다섯 가지의 하위 속성으로 이루어져 있다.

<표 2> 표현 유형의 종류 및 예시

표현 유형	정의	예시
direct-explicit	출처가 어떤 대상에 대하여 가지는 주관성을 명시적인 감정어휘를 통하여 표현하는 경우.	좋다; 싫다; 예쁘다; 멋지다; 대단하다; 그저 그렇다; 지겹다; 최악이다; 엉망이다; 으뜸이다; 획일적이다; 획기적이다; 장점이 많다; 좋아하다; 싫어하다; 지겨워하다
		어른들은 와플을 **지겨워했지만**, 아이들은 **마음에 쏙 들어** 했다. 블라우스가 요즘 **인기다** 여기가 **포인트이다** 이 제품은 **쓸 데가 없다**
direct-speech	출처의 발화 사건(speech event)을 지시하는 술어에 의하여 출처가 취하는 태도, 즉 주관성이 드러나는 경우. 발화 동사들로 구성됨.	주장하다; 설득하다; 비난하다; 호통치다; 칭찬하다; 투덜거리다
		비평가들은 그 영화를 **극찬했다**.
direct-action	출처의 주관성이 특정 행동으로 드러나는 경우.	박수갈채를 보내다; 환호하다; 인상을 구기다; 얼굴색이 변했다
		그는 내 말에 **콧방귀를 뀌었다**; 그는 내 말에 **펄쩍 뛰었다**
indirect	서술자 및 출처의 주관성이 화행을 통하여 표현되는 direct 유형과 대조적으로, 화행을 통하지 않고 간접적으로 주관성이 드러나는 어휘. 명사류; 관형어 및 관형절; 부사어 및 부사절. direct 유형이나 writing-device 유형에 비해 출처가 상대적으로 불분명하다는 특징.	행복; 만족; 악평; 칭찬; 비정상; 강렬한; 뛰어난; **이상한** 사람 누구나 **좋아하는** 사람; 모두가 **찬성하는** 회의; **공감 가는** 댓글; 소비자들이 **가장 많이 찾고** 있는 품목; **예쁘게** 춤을 추다; **좋게** 끝난; **게걸스럽게** 먹다; 과제를 **훌륭히** 마치다; 책상을 **깔끔히** 정리하다
		프랑스의 **세계적인** 의상 디자이너 엠마누엘 웅가로가 실내 장식용 직물 디자이너로 나섰다.
writing-device	텍스트 상의 장치로서 서술자 및 화자의 주관성을 반영하는 어휘. 주관성의 표현이 화행을 통해 이루어지지 않음. 양상 표현; 특수조사; 화자 지향 부사(구); 접속부사; 접속어미	-야 한다; -ㄹ 수 있다; -ㅁ에 틀림없다; -(이)나; -(이)라도; -도; -만; -마저; -까지; 특히; 이상하게도; 신기하게도; 틀림없이; 그러나; 하지만; 그럼에도; -나; -ㅁ에도; -지만; -ㄴ데

이 중에서 indirect 유형은 간접적으로 주관성을 드러내는 명사류를 비롯한 어휘들이 주로 해당하기 때문에 접속부사나 어미에 초점을 맞추는 본 연구에서는 크게 고려하지 않지만 실험에는 사용되었다. direct-speech 유형은 발화 동사로 구성되나, 이는 발화 동사 모두가 direct-speech 유형으로 주석되어야 한다는 것은 아니다. 예를 들어, “말하다”, “밝히다”, “전하다”와 같은 술어는 발화 사건을 지시하지만 전혀 주관성이 나타나지 않는 경우가 있다. 이때, 술어 자체가 주관성을 가지지 않아 Seed 태그의 주석 대상 자체가 될 수 없기 때문에, direct-speech와 같은 속성값이 할당되어 있지 않다.

writing-device 유형은 주관성을 가진다는 점 외에는 다른 네 가지 유형들과 공유하는 특성이 없다. 주관 표현이기에 Seed 태그로 주석되지만, Seed 태그가 가진 상세 속성으로 규정되지 않는다. 이 주관 표현들은 서술자의 주관성을 드러내는 요소로서, 이후 Subjectivity 태그와 Objectivity 태그를 결정하는 데 중요한 요소로 작동한다. Writing-device의 종류를 실제 주석 예로 살펴보면 다음과 같다.

(1) a. 양상: 이번 관광전은 한국 시장을 외국에 널리 알린다는 개최 의도와는 거리가 먼 것으로 보인다. / 더 높은 것은 발목과 무릎에 무리를 줄 수 있으므로 주의해야 한다.

<SEED> **anchor**=“-으로 보이-”, **id**=“u1”, **nested-source**=“w”, **type**=“writing-device” </SEED>

<SEED> **anchor**=“-수 있-”, **id**=“u2”, **nested-source**=“w”, **type**=“writing-device” </SEED>

<SEED> **anchor**=“-야 하”, **id**=“u3”, **nested-source**=“w”, **type**=“writing-device” </SEED>

b. 접속부사: 이번 관광전은 그러나 개최 의도와는 거리가 먼 것으로 보인다.

<SEED> **anchor**=“그러나”, **id**=“u1”, **nested-source**=“w”, **type**=“writing-device” </SEED>

c. 화자지향부사: 특히 5천원 미만의 제품은 대부분 고무 밴드가 쉽게 찢어지는 게 흠이었다. 요즘은 이상하게도 국산 제품이 인기가 많다.

<SEED> **anchor**=“특히”, **id**=“u1”, **nested-source**=“w”, **type**=“writing-device” </SEED>

<SEED> **anchor**=“이상하게도”, **id**=“u2”, **nested-source**=“w”, **type**=“writing-device” </SEED>

d. 특수조사: 제대로 보이지 않는 제품이라도 조사 대상의 20%나 됐었다.

<SEED> **anchor**="-라도", **id**="u1", **nested-source**="w", **type**="writing-device" </SEED>

<SEED> **anchor**="-나", **id**="u2", **nested-source**="w", **type**="writing-device" </SEED>

e. 철수는 꼴에 집에 갔다.

<SEED> **anchor**="꼴에", **id**="u1", **nested-source**="w", **type**="writing-device" … </SEED>

f. 영수는 철수가 꼴에 집에 갔다고 말했다.

<SEED> **anchor**="꼴에", **id**="u2", **nested-source**="w-영수", **type**="writing-device" … </SEED>

위의 예시에서 나타나는 양상 표현, 접속 부사, 화자 지향 부사, 특수조사는 그 통사, 의미적 작용이 문장 전체를 범위로 일어나며, 문장들 사이에 화용적 효과를 가져 온다. 이 때문에 문장 내 보다 작은 범위에서 대상을 찾을 수 없고, 문장의 것과는 구분되는 어휘 고유의 극성이 존재하지 않는 경우가 대부분이다.

3. 감정표현 추출

감정분석 연구에서 가장 일반적으로 활용할 수 있는 것은 주관성을 띠는 어휘 단위의 목록이다. 한국어감정분석코퍼스에서 문장보다 작은 단위의 핵심 주관 표현으로 주석한 SEED는 길이의 분포가 일정하지 않고 단일 형태소에서부터 인용구를 포함한 절까지 편차가 크므로 어휘 단위 표현에 직접 대응하기 어렵다. 특히 SEED가 긴 경우 새로운 코퍼스에서 이와 일치하는 표현이 출현할 가능성이 낮으므로 주석된 정보를 활용하는 데 한계가 있다. 따라서 본 연구에서는 SEED에 주석된 의미 정보를 어휘 단위로 살펴보기 위해 개별 어휘의 감정 특성이 그 어휘를 포함하는 SEED의 감정 특성에서 도출될 수 있다고 가정하고, SEED의 anchor에서 형태소의 유니그램, 바이그램, 트라이그램을 추출하여 감정 어휘 목록을 구축하였다. 단, 한 SEED가 다른 SEED를 포함하는 경우 상위 SEED가 하위 SEED를 인용・부정・강조하면서 감정 특성값을 전환할 수 있으므로, 일관된 감정 특성값을 얻기 위해 다른 SEED와 중첩되지 않는 최하위

SEED에 포함된 형태소만을 사용하였다. 형태소 N-그램은 가능한 모든 것을 뽑되 한글 이외의 문자나 문장 부호가 포함된 것은 제외하였다. 이러한 방식으로 총 16,362가지(유니그램 3,476가지, 바이그램 6,579가지, 트라이그램 6,307가지)의 감정표현을 추출하고, 각 표현의 의미 속성 자질값은 해당 표현을 포함하는 SEED 중에서 가장 높은 빈도로 출현한 값과 비율로 계산하였다. 표현 유형 속성에 따라 추출된 목록에서 각 감정표현이 지니는 속성을 정리하면 (2)와 같다.

(2) 어휘 단위 감정표현의 속성
- ngram (N-그램): 표제어 N-그램을 이루는 형태소
- freq (빈도): 해당 N-그램을 포함하는 SEED의 개수
- dir-action: 해당 N-그램을 포함하는 SEED 중 dir-action의 비율
- dir-explicit: 해당 N-그램을 포함하는 SEED 중 dir-explicit의 비율
- dir-speech: 해당 N-그램을 포함하는 SEED 중 dir-speech의 비율
- indirect: 해당 N-그램을 포함하는 SEED 중 indirect의 비율
- writing-device: 해당 N-그램을 포함하는 SEED 중 writing-device의 비율
- max.value (최대 자질): 가장 높은 비율을 차지하는 자질
- max.prop (최대 비율): 가장 높은 비율의 수치

구체적인 예로 <표 3>에서 유니그램 “그러나/MAJ”를 포함하는 SEED는 총 47개이며, 47개의 SEED 중 표현 유형값이 dir-speech인 것의 비율이 0.0851, indirect인 것의 비율이 0.0212, writing-device인 것의 비율이 0.8936임을 알 수 있다. 마찬가지로 트라이그램 “ㄹ/ETM;수/NNB;밖에/JX”를 포함한 7개 SEED 중 85.71%가 writing-device 값을 갖는다.

<표 3> 감정표현 추출 예시

ngram	freq	dir-action	dir-explicit	dir-speech	indirect	writing-device	max. value	max. prop
그러나/MAJ	47	0	0	0.0851	0.0212	0.8936	writing-device	0.8936
ㄹ/ETM;수/NNB;밖에/JX	7	0	0	0.1428	0	0.8571	writing-device	0.8571

감정표현 중에서는 가장 높은 비율을 차지하는 자질값이 다수인 것도 있다. <표 4>에서 바이그램 “하/VV;지/EC”를 포함하는 SEED는 dir-explicit와 indirect의 비율이 0.3333으로 일치한다. 이 경우 최대 비율이 0.5를 넘을 수 없으므로 감정

표현으로서 정확도가 높지 않다고 보고 배제하였다.

<표 4> 감정표현 추출 예시

ngram	freq	dir-action	dir-explicit	dir-speech	indirect	writing-device	max. value	max. prop
하/VV;지/EC	6	0.1666	0.3333	0.1666	0.3333	0.0000	dir-explicit	0.3333

추출된 감정표현을 최대 비율의 범위에 따라 분류하면 <표 5>에서 볼 수 있듯이 0.5 미만의 값을 가지는 것은 16,362가지 중 1,105가지에 불과하다. 따라서 대부분의 감정표현이 다섯 가지 표현 유형 중 하나의 자질을 우세하게 가짐을 알 수 있다.

<표 5> 최대 유형 비율 범위에 따른 감정표현 분류

최대 비율 범위	0.5 이하	0.5 초과 1.0 미만	1.0	총합
개수	1,105	687	14,570	16,362

최대 자질 비율이 0.5를 초과하는 15,257가지 감정표현을 최대 표현 유형 자질에 따라 분류한 결과는 <표 6>에서처럼 dir-speech > indirect > dir-explicit > writing-device > dir-action 순으로 많이 나타났다.

<표 6> 최대 표현 유형 자질에 따른 감정표현 분류

최대 자질	dir-action	dir-explicit	dir-speech	indirect	writing-device	총합
개수	60	4,035	6,084	4,545	533	15,257

추출된 감정표현의 정확도를 일정 이상 확보하기 위해서는 최대 자질 비율의 값뿐만 아니라 코퍼스에서 출현한 빈도도 고려해야 한다. freq 값이 1을 초과하는, 즉 두 개 이상의 SEED에 나타나는 감정표현의 표현 유형을 최대 자질과 비율에 따라 정리한 결과는 <표 7>과 같다.

<표 7> 최대 표현 유형 자질 및 비율에 따른 감정표현 분류

최대 비율	dir-action	dir-explicit	dir-speech	indirect	writing-device
0.5 이하	20	654	332	74	25
0.5 초과 1.0 미만	1	202	204	182	98
1.0	0	225	370	275	67

최대 표현 유형 자질이 writing-device인 감정표현 중 두 개 이상의 SEED에 출현하면서 최대 비율이 100%인 67가지의 목록을 살펴보면, <표 8>에서처럼 유니그램이 주로 접속부사(그래서/MAJ, 그런데/MAJ, 따라서/MAJ 등), 접속어미(ㄴ데/EC, 려/EC 등), 화자지향부사(겨우/MAG, 무려/MAG, 물론/NNG, 심지어/MAG 등)을 포착하는 한편 바이그램에 특수조사(까지/JX;도/JX, 보다/JKB;는/JX, 뿐/JX;아니/VCN 등)가 포함되고 트라이그램에 양상(수/NNB;있/VV;을/ETM, 아야/EC;하/VX;겠/EP, 있/VV;을/ETM;것/NNB)이 해당하는 것을 확인할 수 있다.

<표 8> writing-device 속성에 해당하는 감정표현 분류

유니그램	겨우/MAG, 그래서/MAJ, 그런데/MAJ, 따라서/MAG, ㄴ데/EC, 는가/EF, 다/JX, 다만/MAG, 따라서/MAJ, 또한/MAG, ㄹ/JKO, 려/EC, 무려/MAG, 물론/NNG, 바람/NNB, 불구/XR, 뿐/JX, 사실/MAG, 심지어/MAG, 야/JX, 우선/MAG, 워낙/MAG, 자/EF, 한편/MAG
바이그램	고/EC;말/VX, 그러/VV;니/EC, 그런데/MAJ;도/JX, 기/ETN;ㄴ/JX, 기/ETN;만/JX, 까지/JX;도/JX, ㄴ/ETM;만큼/NNB, ㄴ/JX;하/VX, 있/VV;을/ETM, 는/ETM;바람/NNB, 기/ETN;때문/NNB, 던/ETM;만큼/NNB, 데/NNB;다/JX, 도/JX;불구/XR, 만/JX;하/VX, 물론/NNG;이/VCP, 바람/NNB;에/JKB, 보다/JKB;는/JX, 뿐/NNB;아니/VCN, 아야/EC;하/VV, 불구/XR;하/XSV, 뿐/JX;아니/VCN, 에/JKB;야/JX, 으면/EC;하/VX, 을/ETM;정도/NNG, 하/VX;겠/EP
트라이그램	기/ETN;ㄴ/JX;하/VX, 기/ETN;때문/NNB;이/VCP, 기/ETN;만/JX;하/VX, ㄴ/ETM;데/NNB;다/JX, 는/ETM;바람/NNB;에/JKB, 다고/EC;하/VV;ㄹ/ETM, 도/JX;불구/XR;하/XSV, 불구/XR;하/XSV;고/EC, 뿐/JX;아니/VCN;라/EC, 뿐/NNB;아니/VCN;라/EC, 수/NNB;있/VV;을/ETM, 아야/EC;하/VX;겠/EP, 에/JKB;도/JX;불구/XR, 은/ETM;것/NNB;이/VCP, 을/ETM;수/NNB;있/VV, 을/ETM;정도/NNG;이/VCP, 있/VV;을/ETM;것/NNB

<표 8>에 추출된 감정표현은 writing-device로서 가장 높은 정확도를 가지는 것에 해당한다. 여기에서 출현 횟수와 최대 비율의 기준을 완화하면 더 많은 표현을 추출하여 적용 범위를 넓힐 수 있다. 다음 절에서는 이러한 방식으로 추출한 감정어휘가 문장의 주관성을 예측하는 데 어떻게 활용될 수 있을지를 평가하는 실험을 수행하고자 한다.

4. 양상정보를 이용한 감정 분석

본 절에서는 한국어감정분석코퍼스에서 추출한 자질을 사용하여 주관적인 문장과 객관적인 문장을 분류하는 실험을 진행한다.

4.1 데이터와 실험개요

실험에 사용된 코퍼스는 한국어감정분석코퍼스에서 Subjectivity 태그와 Objectivity 태그에 해당하는 문장이다. 주석자 세 명이 교차 검증한 결과 총 7,713개 문장 중 2,658개가 주관적인 것으로, 5,055개가 객관적인 것으로 판정되었다. 이 중에서 직접 인용표지(“ ”)로 시작하여 이 표지로 끝나는 직접 인용 문장은 인용자의 의도에 따라 문장 전체가 객관적인 것으로 주석되었으나, 이 실험에서는 문장의 인용자가 아닌 인용문의 화자의 의도에 따라 문장의 주관성을 Subjective 혹은 Objective로 맞게 재 주석하여 사용했다.

전체 문장을 10겹 교차검증법(10-fold cross-validation)에 따라 훈련집합 90%와 실험집합 10%로 분할한 뒤, 총 10회의 분할에 대하여 선형 SVM 모형(Cortes & Vapnik 1995, Fan et al. 2008)을 사용하여 개별 문장의 주관성을 결정하는 자질의 효과를 훈련집합에서 학습하고 이 학습결과에 따라 실험집합에 속하는 문장의 주관성을 예측하는 과정을 Scikit-learn(Pedregosa et al. 2011)으로 수행하였다.

4.2 자질 소개

이 실험에서는 동일한 코퍼스에 대한 기존 연구 Kim & Shin (2014)에서 사용한 자질에 본 연구에서 도입한 양상 정보 자질을 추가하였을 때의 효과를 비교하고자 하였다.

우선 Kim & Shin (2014)에서는 SEED 표현에서 감정 표현 빈도, 확률, expressive-type, semantic-type, intensity 다섯 가지 범주의 자질을 자동으로 추출하였다.

표현 빈도 자질 (SF, N-SF, T-SF). 가장 단순한 가정은 문장 주관성이 문장 내의 주관 표현의 개수로 결정된다는 것이다. 이 그룹에 속하는 자질로 빈도(SF), 문장 길이로 정규화된 빈도(N-SF), 0 이상의 정수를 세 개 범위로(0, 1 or > 2)로

분류한 빈도(T-SF) 세 가지를 추출하였다. 이 그룹의 자질은 문장의 주관성만 반영하고 객관성을 반영하기는 어렵다는 한계가 있다.

확률 자질 (PES, PPOSS, PSS, PSS-POS). 이 그룹에는 우선 표현 단위의 주관성 확률(PES), 품사 시퀀스 단위의 주관성 확률(PPOSS), 각 표현이 속하는 문장의 주관성 확률 (PSS), 각 품사 시퀀스가 속하는 문장의 주관성 확률(PSS-POS) 네 가지 자질 집합이 있다. (1) PES 자질은 한 표현이 주관적일 확률로, 표현의 전체 빈도를 그 표현이 주관적으로 주석된 빈도로 나눈 값이다. PPOSS 자질은 품사 시퀀스에 대해 같은 방식으로 측정한 값이다. (2) PSS 자질은 훈련집합 내에서 한 표현이 속한 문장이 주관적일 확률이며, PSS-POS는 표현 대신 품사 시퀀스에 대해서 같은 방식으로 정의된 값이다. 이 네 척도가 개별 자질이 아닌 자질들의 집합으로 간주되는 이유는 한국어 감정분석코퍼스에서 SEED로 태그된 모든 표현을 포함하기 때문이다. 각 표현에 대한 PES와 PSS 자질값을 수식으로 표현하면 (3)과 같다.

(3) PES 자질과 PSS 자질의 정의(Kim & Shin 2014)

$$PES = \frac{S_{eo}}{N_{eo}} \times \frac{N_{eo}}{N_{eo} + K} \quad 4)$$

$$PSS = \frac{S_{ss}}{N_s} \times \frac{N_s}{N_s + K} \quad 5)$$

이 그룹에 속하는 나머지 자질 네 개(AVG-PES, AVG-PPOSS, AVG-PSS, AVG-PSS-POS)는 앞서 설명한 PES, PPOSS, PSS, PSS-POS 자질값들을 각각 문장 내에 출현한 단어들에 대하여 평균을 낸 값이다. 이 자질은 문장 내에 주관적 표현이 최소한 하나는 있어야 구할 수 있다. 문장 내에 주관적 표현이 없다면 자질값을 0으로 설정하여 문장의 객관적 측면을 반영하였다. 따라서 평균값을 사용한 자질은 문장의 주관성과 객관성을 양 방향으로 포착할 수 있다는 이점이 있다.

강도 자질 (CAT-INT, CONT-INT, AVG-CONT-INT). 이 그룹은 두 개의 자질 집합으로 각 표현의 강도를 범주로 간주한 CAT-INT와 연속적인 값으로 간주한

4) N_{eo}: the total number of the expression occurrences; S_{eo}: the number of the expression occurrences as subjective; $N_{eo}/(N_{eo}+K)$: smoothing term for compensating the effect of low frequent words; K: a parameter that determines the degree of smoothing. K is set to 1.

5) N_s: number of sentences that the seed belongs to; S_{ss}: number of subjective sentences that the subjective expression belongs to

CONT-INT와, 하나의 자질로 문장 내 주관 표현들의 연속적 강도 값의 평균인 AVG-CONT-INT로 이루어졌다. 확률 자질과 마찬가지로 주관 표현들의 출현에만 의존하는 두 자질 집합은 주관성만 반영하며, 주관 표현들의 부재도 고려한 평균값 자질은 문장의 객관성까지 반영할 수 있다.

각 표현의 강도를 범주로 나타낼 때는 가장 자주 나타난 값 (high-3, medium-2, low-1)을 선택하였고, 연속적인 값으로 나타낼 때는 (4)와 같은 식으로 계산하였다.

(4) 주관 표현 강도의 연속적인 값 측정 방법(Kim & Shin 2014)

$$CONT-INT = \frac{((3*freq_{high})+(2*freq_{midium})+(1*freq_{low}))}{total\ frequency\ of\ the\ word}$$

표현 유형 자질 (EXP-TYPE-FREQ, EXP-TYPE-NORM, EXP- TYPE-THREE). 한국어 감정분석코퍼스 SEED 태그에서 표현 유형은 direct-speech, direct-action, direct-explicit, indirect, writing-device 여섯 가지이다. 이 중 direct로 시작하는 유형은 문장의 술어에 해당하는 주관적 표현이고 indirect 유형은 주로 명사, 부사, 형용사에 해당한다. writing-device 유형에 속하는 문장부사, 양상, 접속부사 등이 주로 문장의 주관성을 결정하는 데 중요한 역할을 한다. 이 그룹의 자질은 표현 빈도 자질과 마찬가지로 표현 유형의 빈도(EXP-TYPE-FREQ), 문장 길이로 정규화된 빈도(EXP-TYPE-NORM), 세 개 값으로 분류된 빈도(EXP-TYPE-THREE) 세 가지로 정의되었다. 각 유형에 해당하는 표현이 없는 경우 빈도에 0을 할당하여 문장의 객관성을 반영하였다.

의미 유형 자질 (SEM-TYPE-FREQ, SEM-TYPE-NORM, SEM- TYPE-THREE). SEED 표현의 의미 유형 agreement, argument, emotion, intention, judgment, speculation, others에 대해서도 표현 유형과 같은 방식으로 SEM-TYPE-FREQ, SEM-TYPE-NORM, SEM-TYPE-THREE 세 가지 자질을 정의하였다.

단어 주머니 자질 (BOW). 이 자질에서는 각 문장을 유니그램의 집합과 이들의 빈도로 표현하였다. 이 자질은 단순하고 자명하지만, 앞서 사용된 자질에서 포착하기 어려웠던 객관성을 시사하는 자질을 포함할 수 있다는 데 의의가 있다.

정제된 표현 유형 자질 (REF-TYPE-ALL, REF-TYPE-EXPLICIT, REF-TYPE-ACTION, REF-TYPE-SPEECH, REF-TYPE-INDIRECT, REF-TYPE- DEVICE). 이 자질 그룹은 본 연구를 위해 정제된 어휘 리스트 중에서 표현 유형(expressive type)에 따라 어휘를 이용하여 문장의 주관성을 판단하려는 자질이다. 어휘 리스트는 dir-explicit, dir-action, dir-speech, indirect, writing-device에 따라 그 종류가 구분되며 각각의

어휘 리스트는 어휘마다 갖고 있는 코퍼스에서의 출현 빈도값(freq)과 최대 비율값(max.prop)을 사용하여 어휘의 수는 더 늘이거나 줄일 수 있다. 어떤 출현 빈도와 최대 비율값을 이용하여 어휘의 수를 조절해야 하는지는 실험을 통해서 결정하였다. 이 어휘 목록은 유니그램, 바이그램, 트라이그램으로 구성되어 있기 때문에 자질을 추출하려는 문장도 유니그램, 바이그램, 트라이그램으로 각각 변환하여 어휘 목록과 매칭을 통해서 자질을 추출하였다. 여섯 가지 표현 유형마다 각각 어휘의 빈도와 각 어휘 토큰을 자질로 삼아 빈도수를 자질화 하였다. 이 자질 그룹은 각 표현 유형만을 포함하는 자질 조건과 (REF-TYPE-EXPLICIT, REF-TYPE-ACTION, REF-TYPE-SPEECH, REF-TYPE-INDIRECT, REF-TYPE-DEVICE) 모든 표현 유형을 포함하는 자질 조건(REF-TYPE-ALL)을 사용하여 총 6가지 자질 조건으로 실험되었다.

4.3 실험 결과

본 장에서는 두 가지 실험을 소개한다. 첫 번째 실험에서는 어휘의 빈도와 최대 비율값의 기준을 조정할 때 문장의 주관성 판단 성능에서 나타나는 차이를 관찰하고, 이를 바탕으로 최적의 최대 비율값과 빈도값을 찾는다. 두 번째 실험은 Kim & Shin (2014)의 문장 주관성 판단 결과와 본 연구에서 표현 유형에 따라 구성한 어휘 사전을 사용할 때의 결과를 비교하는 연구이다.

4.3.1 사전 정제 파라미터 실험

최대 비율값을 조절할 때는 최소 빈도를 2 이상으로 고정시켰고, 반대로 빈도값을 조절할 때는 최대 비율값이 0.7 이상인 어휘로 제한하였다.

동일한 조건의 실험을 위해 기존 연구에서 효과적인 자질의 조합으로 밝혀진 N-FS, AVG-PES, AVG-PPOSS의 자질에 본 연구의 정제된 어휘 자질인 REF-TYPE-ALL자질을 조합하여 이 자질의 조합이 파라미터의 변화에 따른 문장 주관성 판단 실험을 위해 사용하도록 했다.

<표 9>의 실험 결과를 살펴보면 최대 비율값이 0.5 이상인 어휘를 추출했을 때 가장 높은 정확도를 보였다. 그리고 최대 비율값이 1에 가까워지면 포함되는 어휘의 수가 줄어들면서 성능의 저하가 관측되었다.

어휘의 빈도에 따라 사전에 포함되는 어휘의 수를 조정하는 실험 결과를 보여주는 <표 10>에서는 빈도수의 조절에 따라 큰 성능에 차이는 보이지 않지만 빈도 2 또는 3 이상의 어휘만을 포함시켰을 때 최상의 성능을 보이는 것으로

관찰되었다. 따라서 본 연구에서 표현 유형에 따른 어휘 사전을 구성할 때 0.5 이상의 최대 비율값과 2 이상의 빈도를 갖는 어휘만을 포함하도록 하여 차후 실험을 진행하도록 하였다.

<표 9> 어휘의 최대 비율값 파라미터 조정에 따른 주관성 판단 성능 변화표
(목적 클래스: 주관적 문장/객관적 문장)

	0.5	0.6	0.7	0.8	0.9
precision	70.03/80.08%	69.34/80.08%	69.43/80.24%	68.44/78.64%	68.15/78.75%
recall	83.07/65.67%	83.44/64.30%	83.62/64.38%	82.29/63.25%	82.55/62.71%
f1	75.96/72.11%	75.71/71.29%	75.84/71.40%	74.71/70.08%	74.63/69.78%
accuracy	74.19/74.23%	73.72/73.72%	73.84/73.84%	72.63/72.63%	72.43/72.46%

<표 10> 어휘의 빈도값 파라미터 조정에 따른 주관성 판단 성능 변화표
(목적 클래스: 주관적 문장/객관적 문장)

	1	2	3	4	5
precision	69.41/80.24%	69.43/80.24%	69.42/80.24%	69.39/80.24%	69.41/80.24%
recall	83.62/64.35%	83.62/64.38%	83.66/64.38%	83.62/64.35%	83.57/64.35%
f1	75.83/71.38%	75.84/71.40%	75.85/71.40%	75.82/71.38%	75.81/71.38%
accuracy	73.83/73.83%	73.84/73.84%	73.84/73.84%	73.81/73.83%	73.81/73.83%

4.3.2 양상정보를 활용한 문장 주관성 판단 실험

KOSAC 코퍼스를 사용하여 문장의 주관성을 판단한 기존 연구(Kim & Shin 2014)에 따르면 <표 11>에서 보는 것처럼 문장 내의 SEED의 개수(N-SF), 문장 내의 SEED의 주관성 확률 값의 평균(AVG-PES), 문장 내의 SEED의 형태소 연쇄가 주관적일 확률의 평균(AVG_PPOSS)의 자질 조합에 EXP-TYPE-FREQ 또는 SEM_TYPE_FREQ 자질을 조합하는 것이 가장 정확한 주관성 판단 결과를 보였다.

<표 11> KOSAC코퍼스를 활용한 자질 조합 실험(Kim & Shin 2014)

N.	Feature Combination	precision	recall	f1	accuracy
(1)	N-SF	53.44/84.87	96.57/18.71	68.77/30.61	56.94
(2)	N-SF + AVG-PES* + AVG-PPOSS*	57.90/78.26	89.33/37.24	70.23/50.42	62.83
(3)	N-SF + AVG-PSS* + AVG-PSS-POS*	52.72/83.40	96.63/16.28	68.18/27.18	55.73
(4)	(2) + CAT-INT	65.10/69.15	70.39/63.55	67.53/66.08	66.89
(5)	(4) + CONT-INT	64.23/68.34	69.99/62.22	66.85/64.96	66.00
(6)	(4) + PSS	64.13/70.00	73.10/60.44	68.23/64.75	66.63
(7)	(4) + EXP-TYPE-FREQ*+ SEM-TYPE-FREQ*	61.10/76.55	84.69/47.92	70.93/58.84	65.96
(8)	(2) + EXP-TYPE-FREQ*	64.39/71.93	75.99/59.37	69.66/64.99	67.52
(9)	(2) + SEM-TYPE-FREQ*	66.36/70.13	71.38/64.96	68.71/67.37	68.11
(10)	ALL-FEATURES	57.33/69.03	80.01/69.03	66.67/52.20	60.85

본 연구에서는 정제된 사전으로부터 얻은 어휘 목록 중 양상정보와 관련이 있는 writing-device 또는 문장을 주관적으로 만드는 문장의 술어와 관련된 dir-action, dir-explicit, dir-speech와 같은 표현 유형의 어휘 목록을 이용하여 기존의 연구 결과와 비교하여 문장의 주관성 판단에 있어서 어떤 차이를 보이는지 실험했다.

기존 Kim & Shin (2014)에 보고된 최적의 자질 조합 중 이 연구에서 사용한 어휘와 그 효과가 중첩되는 EXP-TYPE-FREQ과 SEM_TYPE-FREQ를 제거한 조합에 본 연구의 어휘목록을 적용한 실험을 진행한 결과 <표 12>에서 보는 바와 같이 개별적인 표현 유형의 어휘목록을 이용하여 문장에서 주관성에 영향을 주는 자질을 추출하는 것만으로도 기존 연구 결과의 가장 높은 정확률 68.11%를 넘어서는 정확도(accuracy)를 관찰할 수 있었다. 표현 유형 다섯 가지 중 한 가지씩만 사용한 것 중에서는 화자의 직접적인 화행을 나타내는 REF_TYPE_SPEECH 자질 그룹이 기존 연구 대비 가장 큰 정확률의 증가를 보였으며 (71.84 / 71.73%) 다양한 양상 정보를 포함하는 REF_TYPE_DEVICE도 65.39 / 65.36%의 정확률로 기존의 연구 자질 그룹과 조합되어 문장 주관성 분석 연구에 그 기여하는 바를 확인 할 수 있었다. 또한, 모든 표현 유형을 조합하여 자질로 삼아 문장의 주관성 판단에 사용할 때는 기존결과 대비 6% 포인트 이상의 accuracy의 증가를 확인할 수 있었다. 이러한 실험 결과를 통해 다양한 화용론적인 양상정보를 포함하는 어휘 목록을 KOSAC에서 추출하는 것이 문장의 주관성 판단과 같은 실험에 적극적으로 활용될 수 있음을 볼 수 있다.

<표 12> 정제된 양상정보 어휘를 활용한 문장의 주관성 판단 실험

	precision	recall	f1	accuracy
(2) + REF-TYPE-ACTION	57.64/90.97%	96.81/31.28%	72.24/46.50%	63.48/63.48%
(2) + REF-TYPE-EXPLICIT	60.91/83.79%	91.31/43.40%	73.05/57.15%	66.95/66.95%
(2) + REF-TYPE-SPEECH	65.88/82.94%	88.45/56.29%	75.45/66.97%	71.84/71.73%
(2) + REF_TYPE_INDIRECT	61.20/80.33%	88.43/45.86%	72.32/58.35%	66.76/66.79%
(2) + REF_TYPE_DEVICE	59.22/88.16%	94.90/36.89%	72.90/51.96%	65.39/65.36%
(2) + REF_TYPE_ALL	70.02/80.19%	83.07/65.67%	75.95/72.16%	74.29/74.17%

5. 결론

본 연구는 기구축된 한국어감정분석코퍼스에서 여러 자질로 주석된 정보들을 추출하고 다차원적으로 결합하여 감정표현 자원을 구축하고 그 자원을 실제 감정분석 연구에 활용해 보는 것을 목표로 하였다. 본 연구에서는 기존의 감정분석 연구에 있어 단순한 극성어휘들의 분포에 의존한 연산 방법에서 벗어나 더 정교한 감정분석을 행하기 위해 양상정보를 포함하는 화용적 정보의 도입을 시도하였다.

이를 위해 한국어감정분석코퍼스에서 화자의 태도와 관련된 화용적 정보를 포함하고 있는 표현 유형 속성을 그 빈도와 확률에 따라 목록화 하였고 이를 실제 감정분석연구에 적용하였다. 이 자질들을 사용하였을 때 선행 연구에서의 정확성보다 6% 포인트나 향상되는 결과를 보였다. 이는 극성어휘 외에 다양한 양상 및 화용적 정보를 포함시키는 것이 필요하다는 것을 잘 보여준다. 또한 기존의 감정표현 주석만 되어 있던 코퍼스로부터 그 속성자체의 특질과 다른 속성과의 결합관계에서 나타나는 특질을 고려하여 표현을 추출하는 것은 코퍼스의 활용도를 높이는 것뿐만 아니라 감정분석 연구에 있어 새로운 자료로 활용될 수 있다는 점에서도 의의가 있다. 본 연구에서 사용된 자료는 현재 웹상에서 공개되어 있어 앞으로 이 분야 연구에 크게 활용될 수 있으리라 기대한다.

참고문헌

김문형 · 장하연 · 조유미 · 신효필(2013), “KOSAC(Korean Sentiment Analysis Corpus): 한국어 감정 및 의견 분석 코퍼스”, 「2013년 한국컴퓨터종합학술대회 논문집」.

Banfield, Ann (1982), *Unspeakable Sentences*, Boston: Routledge and Kegan Paul.

Bergler, Sabine (1992), *Evidential Analysis or Reported Speech*, Ph.D Dissertation, Brandeis University.

Chafe, Wallace & Joanna Nicholas (eds.) (1986), *Evidentiality: The Linguistic Coding of Epistemology*, Norwood: Ablex.

Choi, Yejin, Claire Cardie, Ellen Riloff & Siddharth Patwardhan (2005), Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns, *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP 2005).

Cortes, Corinna & Vladimir Vapnik (1995), Support-Vector Networks, *Machine Learning* 20, 273-297.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsiech, Xiang-Rui Wang & Chih-Jen Lin (2008), LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research* 9, 1871-1874.

Finegan, Edward (1995), Subjectivity and subjectification: an introduction, in D. Stein & S. Wright (eds), *Some observations on factivity, Papers in Linguistics* 47, 340-358.

Karttunen, Lauri (1973), Presuppositions of compound sentences, *Linguistic Inquiry* 4(2), 169-193.

Kim, Munhyong & Hyopil Shin (2014), Pinpointing Sentence-Level Subjectivity through Balanced Subjective and Objective Features, *Lecture Notes in Artificial Intelligence* 8686.

KOSAC (2012), Korean Sentiment Analysis Corpus, http://word.snu.ac.kr/kosac.

Langacker, Ronald (1985), Observations and speculations in subjectivity, in J. Haiman (ed), *Iconicity in Syntax. Typological Studies in Language* 6, Amsterdam/ Philadelphia: John Benjamins.

Langacker, Ronald (1991), *Foundations of Cognitive Grammar, vol. 2, Descriptive*

Application, Stanford University Press.

Langacker, Ronald (1997), The contextual basis of cognitive semantics. in J. Nyuts & E. Pederson (eds), *Language and Conceptualization*, Cambridge: Cambridge University Press, 229-252.

MPQA (2005), Multi-Perspective Question Answering. University of Pittsburgh. http://www.cs.pitt.edu/mpqa/.

Mushin, Llana (2001), *Evidentiality and Epistemological Stance*, Amsterdam/ Philadelphia: John Benjamins.

Palmer, Frank (1986), *Mood and Modality*, Cambridge: Cambridge University Press.

Pustejovsky, James, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizaukas, Andrea Setzer, Beth Sundheim, Lisa Ferro, Marcia Lazo, Inderjeet Mani, & Dragomir (2003), The TimeBank corpus, in *Proceedings of Corpus Linguistics* 2003, 647-656.

Riloff, Ellen & Janyce Wiebe (2003), Learning Extraction Patterns for Subjective Expressions, *Conference on Empirical Methods in Natural Language Processing* (EMNLP-03), ACL-SIGDAT.

Riloff, Ellen, Janyce Wiebe, & William Phillips (2005), Exploiting subjectivity classification to improve information extraction, in *Proceedings of the 20th National Conference on Artificial Intelligence* (AAAI-2005), 1106-1111.

Rubin, Victoria (2010), Epistemic modality: from uncertainty to certainty in the context of information seeking as interactions with texts, *Information Processing and Management* 46, 533-540.

Sauri, Roger (2008), *A Factuality Profiler for Eventualities in Text*, Ph.D Dissertation, Brandeis University.

Sauri, Roger & James Pustejovsky (2007), Determining modality and factuality for text entailment, in *Proceedings of the First IEEE International Conference on Semantics Computing*, 509-516.

Shin, Hyopil, Munhyong Kim, Yu-Mi Jo, Hayeon Jang & Andrew Cattle (2012), Annotation Scheme for Constructing Sentiment Corpus in Korean, in *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, 181-190.

Shin, Hyopil (2014), A Corpus Study of Nested Sources for Subjective Analysis, *Eoneohag* 69.

Wiebe, Janyce (2002), Instructions for annotating opinions in newspaper articles,

Department of Computer Science Technical Report TR-02-101, University of Pittsburgh.

Wiebe, Janyce, Theresa Wilson, & Claire Cardie (2005), Annotating expressions of opinions and emotions in Language, *Language Resources and Evaluations* 39(2), 165-210.

<Abstract>

Modality-based Sentiment Analysis through the Utilization of the Korean Sentiment Analysis Corpus

Hyopil Shin, Munhyong Kim, and Suzi Park

This study develops a practical application of language resources from the Korean Sentiment Analysis Corpus (KOSAC) for sentiment analysis research. With this in mind, based on their sentiment properties and the probabilistic factors of annotated expressions from KOSAC, we extracted annotated expressions and refined them to be a sentiment analysis research resource. This study attempted to break away from simple calculation methods dependant on the distribution of lexical polarity items seen in previous research. Additionally, in order to perform more sophisticated sentiment analysis, we attempted to introduce pragmatic information which includes modality. In order to achieve this, we cataloged expressions that include pragmatic information related to the speaker's attitude, based on their relative probability in KOSAC. After doing so, this study shows a practical application of this new language resource to subjectivity analysis research. When using this new resource, this research demonstrates an accuracy improvement of around 6%. This demonstrates very clearly that, in addition to polarity items, there exists a need to include a variety of aspects and lexical information when doing this type of research. Moreover, this extraction of sentiment expressions, depending on their semantic and pragmatic properties, not only shows an additional use of KOSAC, but also establishes a new resource in the field of sentiment analysis.

Keywords: modality, sentiment analysis, pragmatics, KOSAC

논문 접수: 2016.04.06
논문 수정: 2016.04.12
게재 결정: 2016.04.15